ABSTRACT

Title of Document: **PRAGMATIC ENRICHMENT IN LANGUAGE PROCESSING AND DEVELOPMENT**

Shevaun N. Lewis, Doctor of Philosophy, 2013

Directed By: Professor Colin Phillips
Department of Linguistics

The goal of language comprehension for humans is not just to decode the semantic content of sentences, but rather to grasp what speakers intend to communicate. To infer speaker meaning, listeners must at minimum assess whether and how the literal meaning of an utterance addresses a question under discussion in the conversation. In cases of implicature, where the speaker intends to communicate more than just the literal meaning, listeners must access additional relevant information in order to understand the intended contribution of the utterance. I argue that the primary challenge for inferring speaker meaning is in identifying and accessing this relevant contextual information.

In this dissertation, I integrate evidence from several different types of implicature to argue that both adults and children are able to execute complex pragmatic inferences relatively efficiently, but encounter some difficulty finding what is relevant in context. I argue that the variability observed in processing costs associated with adults' computation of scalar implicatures can be better understood

by examining how the critical contextual information is presented in the discourse context. I show that children's oft-cited hyper-literal interpretation style is limited to scalar quantifiers. Even 3-year-olds are adept at understanding indirect requests and "parenthetical" readings of belief reports. Their ability to infer speaker meanings is limited only by their relative inexperience in conversation and lack of world knowledge.

PRAGMATIC ENRICHMENG IN LANGUAGE PROCESSING AND DEVELOPMENT

By

Shevaun N. Lewis

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2013

Advisory Committee:
Professor Colin Phillips, Chair
Professor Valentine Hacquard
Professor Yi Ting Huang
Professor Jeffrey Lidz
Professor Rochelle Newman

Acknowledgements

I'm overwhelmed at the prospect of trying to summarize in just a few paragraphs how much my academic family at Maryland has meant to me. It's impossible to cordon off a subset of them who specifically contributed to the writing of this dissertation. They all did, by teaching me almost everything I know about linguistics and helping me through the last five years one way or another. It was sheer luck that I even applied to Maryland, but I loved it from the beginning and now it's hard to image a future of "doing science" anywhere else.

Colin was everything I could have hoped for in an advisor—both a stalwart supporter and a tough critic. Although I may have fooled him in my interview into believing I had some idea what I was talking about, I don't think I've gotten much past him since then. I'm grateful that even as my research began to wander away from my original plans and his areas of interest, he continued to contribute his ideas and feedback. And of course I have Colin to thank for the shiny bronze baton on my bookshelf and the fact that I can now run 10 miles without collapsing. Who else would have entertained the twisted notion that 12 linguists can and should run 200 miles together, not just once but annually?

I'm much obliged to Valentine and Jeff for getting me started on my first experiment that actually worked. The attitudes group has been a very fertile source of inspiration for me over the last four years, as Chapter 5 of this dissertation demonstrates. Valentine made me want to learn something about semantics, and the more I do, the more it pays off. Jeff is an invaluable resource on pretty much anything related to kids or linguistics. I have long since forgiven him for endorsing the idea for my first experiment that didn't work (and didn't work and didn't work and didn't work).

Almost every faculty member in the linguistics department has affected my intellectual development in one way or another. I'm especially grateful to Andrea Zukowski for all our discussions over the years and her patience with my failure to ever implement our plans for experiments. Alexander Williams, who always makes himself heard in classes and lab meetings, has had a huge impact on the ideas developed in this dissertation. I'm still in awe of Paul Pietroski, who makes old-timey philosophers make sense, and moreover mixes a mean cocktail. Norbert Hornstein has always provided encouragement with both words and cookies. Even more importantly, he constantly reminds us that enjoying linguistics doesn't mean you have to give up enjoying other things and living the good life.

Many thanks to Yi Ting Huang and Rochelle Newman for rounding out my dissertation committee. Although my work never got as far into the Hearing & Speech side of things as I originally hoped, I learned a lot from Rochelle in reading and discussion groups over the years. I was impressed by Yi Ting from the first time I met her at a CUNY poster session, and I'm glad that we overlapped at Maryland for a couple years. I'm eager and a little bit scared to get her feedback on this dissertation.

The ideas in this dissertation owe a lot to discussions in the attitudes group—Valentine Hacquard, Jeff Lidz, Erin Eaker, Kate Harrigan, Aaron White, Rachel Dudley, Naho Orita, and Morgan Moyer—as well as the various people who have been kind enough to visit with us: Jill de Villiers, John Grinstead, Chris Kennedy, Keir Moulton, Paul Portner, Aynat Rubinstein, Meredith Rowe, and Florian Schwarz.

The students in the linguistics department are an amazing bunch, both personally and professionally. I learn something from one of them almost every day, whether in class, a lab meeting, a windowless office, the lunch room, or a bar. I'm not sure why they put up with my constant contrariness, but I'm sincerely grateful they do (though I don't often show it). I depended on them more than ever to survive this last year.

My cohort—Wing Yee Chow, Dave Kush, Ewan Dunbar, Michaël Gagnon, Terje Lohndal, Chris Laterza, and Brad Larson—have been a continuing source of inspiration and support. When Wing Yee and I shared the floor of Annie's apartment as prospective students five years ago, I had no idea how much I would come to depend on her as a collaborator, officemate, and friend. Any time I got stuck—on writing, statistics, materials, anything—I could just spin my chair around and there she'd be with some helpful advice. Dave took a long time to come around (although he'll always insist it was the other way round), but now that he has he's a frequent companion in drinking, Settlers, complaining, and occasionally science. If I ever need to be reminded that I'm difficult, I know who to call. Alexis Wellwood, the unofficial 9th member of our cohort, is a terrific linguist and loyal friend. She was there with whiskey and Thai food when I needed it.

Sol Lago can always be depended on to tell it like it is (or at least how she sees it). Shannon Barrios and Naho Orita along with Sol and Wing Yee formed a sort of unofficial girls' club this past year, providing much needed support with sometimes silly and sometimes tearful companionship. Annie Gagliardi got me out of bed for many an early morning run last summer, and also co-created the best (and most underappreciated) Peep diorama ever. Aaron White was the first Drog back when laminated scenes with velcro seemed like a good idea, but now is my resource on beer, jazz, and anything else the cool kids are doing. Kate Harrigan is not only a great collaborator (see Experiments 3-5) and friend, but also a patient role model on how to be fabulous and nail Fancy Friday. Morgan Moyer's sunny attitude is so infectious that she can make a Saturday in the preschool lab seem like fun. I've appreciated the laughs and the whiskey shots this year.

I am indebted to the many undergraduate research assistants who have contributed to the "Puffin" studies: Faina Kostyukovsky, Jessica Lee, Leah Whitehill, Sam Blitzstein, Laura Sherry, Sara McVeigh, Amber Frazier, and Amritha Mallikarjun. At 262 children and counting, this project would never have come so far without them. I'm especially grateful to Faina, who worked on the project for three semesters and two summers. I'm not sure what kept her coming back—it certainly was not my total lack of enthusiasm for Harry Potter movies—but I'm glad she was there to make sure no one took themselves too seriously, me included. Tara Mease is of course completely essential to maintaining the sanity of the complex institution that is the infant lab: I greatly admire her ability to keep her cool under pressure. I dread the prospect of having to run acquisition experiments without her in the background making everything go smoothly.

I wouldn't have gotten to UMD at all if not for the great advising I got at Yale. From my first class with Maria Piñango, I knew that psycholinguistics was for me. I admired her confrontational teaching style immediately. People who know me now are shocked to hear that I rarely spoke up in class before. Discussions with Maria prepared me more than anything else for the intense debates I enjoyed in grad school.

Everything I know about working with kids I owe to my mentors at the University of Washington Autism Center, as well as the children who frustrated and delighted me by turns. I feel so honored to have had the chance to have a positive impact on their lives.

Of course, it's my family who made me, and not just by providing me with a fancy education and a ridiculously comfortable life. I'm a bookish, sarcastic, thick-skinned Lewis through and through. From my dad I learned that making up well-reasoned scientific arguments needn't depend on my knowing any actual facts. Since I'm turning out to be exactly like my mom whether I like it or not, you can take your pick of my most endearing or frustrating traits to blame on her. Jared and Meghan show me the kind of life I could live if I had any social skills, or the guts to occasionally do something dangerous. Keith is an endless source of humor and debate. He makes sure I will never forget that not all minds are made alike.

As for Brad, what can I say? You make me a better person. You insist that I appreciate the small things. I'm pretty sure we've seen the worst of each other this year, but I couldn't have done it without you.

Table of Contents

List of Tables

List of Figures

1 Introduction

1.1 Overview

Language comprehension is an act of communication. The goal of comprehension is not merely to decode a linguistic representation, but rather to grasp the "thought" that the speaker intended to convey. Thus, a full account of the development and real-time deployment of comprehension processes must go beyond "parsing": we must characterize how linguistic representations are integrated with knowledge of the physical and social context to derive the speaker's intended message. Moreover, it may be impossible to fully understand parsing itself without understanding the communicative system it is embedded in.

Although theories differ slightly on how much meaning should be encoded in semantic representations, all agree that the gap between semantically-encoded "meanings" and the speaker's intended message is vast. The aspects of meaning that are not semantically encoded are supplied by "pragmatic enrichment"—a large and diverse grab-bag of pragmatic operations. The exchange in (1) demonstrates the diversity of pragmatic enrichments necessary even for very commonplace conversation. (2) gives a rough translation of some of the enrichments. The goal for pragmatic theories is to explain how the complex messages in (2) can be derived from the rather simpler utterances in (1).

(1) Host: I'm making some coffee, if you're interested.

Guest: Actually I need to get to bed early tonight.

(2) Host: I (*the speaker*) am making some coffee (*concurrently with this conversation, or at least very soon*). *I'm telling you in case* you're interested *in drinking some coffee now, and if you are, I hereby offer you some*.
Guest: *In slight contradiction to what you're suggesting*, I (*the speaker*) need to get to bed early tonight *compared to a conventional bedtime. Since going to bed early would be difficult if I drank some coffee, I hereby refuse your offer*.

I am concerned with one broad type of pragmatic enrichment—conversational implicature. People often mean more than they say: they *implicate* meanings that are not directly expressed by the literal content of their utterances. These indirect speech acts can be diagnosed through the mismatch between the literal meaning and the goals of the conversation. In (1), the host literally says that she is making coffee, and implicates that she is offering some to the guest. The guest literally says that he needs to go to bed early, and implicates that he is refusing the offer of coffee.

This dissertation focuses on implicature in language comprehension. The basic structure of the problem for comprehension is as follows. The literal meaning of an utterance is assumed to be relatively accessible for comprehenders: they have some efficient way of recovering it from the meanings of the words and the syntactic structure. However, comprehenders (unless they are logicians or lawyers) are not interested in the raw literal meaning of an utterance: they want to understand what the speaker intended to communicate. To infer what the speaker intended, they must combine the literal meaning of the utterance with some relevant contextual information. This contextual information can involve the immediate linguistic context, the purpose of the conversation, or the hearer's knowledge about the speaker or the world. In fact, almost any information,

linguistic or non-linguistic, is in principle relevant to understanding a speaker's intention. This is a problem because comprehenders are finite systems, and they have a finite (in fact very short) amount of time to access all of the relevant information in the course of determining the speaker's meaning.

In (1), the relevant non-linguistic information might be characterized as a scheme for a conventional social interaction: A offers B a drink, B accepts or refuses. In other situations, other information could have been relevant to interpreting the host's statement. For example, she could have delivered the same utterance to her husband, with a sarcastic tone and a raised eyebrow, implicating that *once again, you didn't make the coffee, and I had to do it myself like everything else.* Comprehenders must be prepared to complete an unbounded array of inferences as well as a linguistic analysis of the utterance in order to arrive at the intended interpretation. Nevertheless, although most utterances are indirect in one way or another, adults experience little difficulty determining what people mean: normal conversations generally do not feel like puzzles.

1.2 Methodological goals

Experimental pragmatics has become a rich and vibrant field, and we have made significant advances in our understanding of how speaker meaning is computed in utterance comprehension. In this dissertation, I contribute some novel empirical findings to this literature. More importantly, however, I propose some adjustments to how we approach experimental investigation of implicature. I emphasize the utility of integrating insights from different domains—processing and development, and different kinds of implicature. I also highlight the necessity of stating psycholinguistic hypotheses at the level of algorithm, rather than computational theory.

1.2.1 Integrating insights from different domains

Pragmatics is an exceedingly broad and diverse field, and yet many pragmatic phenomena in the psycholinguistic and developmental literatures are studied in near isolation. A theme in this dissertation is that integrating insights from multiple domains will be useful for understanding specific areas of competence as well as the big picture of pragmatic competence.

I will pull together evidence about adults' (Chapter 3) and children's (Chapters 4-6) comprehension of implicature. Psycholinguistic evidence suggests that adults readily compute implicatures, but that these pragmatically enriched interpretations incur higher processing costs than literal interpretations. Children are reported to rarely compute implicature-enriched interpretations. Although it is often proposed that children's difficulty is related to the processing costs in adults, the observation usually ends there. To the extent that the processing cost has been attributed to any specific aspect of implicature, it is generally to the supposed inferences required to compute the implicated meaning. I will argue that neither children nor adults are unduly burdened by the inference processes themselves. Rather, both groups have difficulty identifying and accessing relevant contextual information.

I will also integrate insights from a range of different implicature phenomena. Much of the experimental literature has become single-mindedly focused on scalar implicature, although it is not necessarily representative of implicature as a whole. While most conversational implicatures can only be explained with reference to the entire utterance in context, scalar implicatures arise predictably for certain types of expressions. I will discuss scalar implicature at length, but I also look at some relevance implicatures:

indirect requests (Chapter 5) and parenthetical uses of belief reports (Chapter 6). Considering these other kinds of implicatures changes our understanding of the generalizations in the literature on scalar implicatures in children. We will see that children are not always "hyper-literal": with other kinds of implicatures, their limited understanding of context leads children to over-generate enriched interpretations.

1.2.2 Distinguishing computational and algorithmic theories

Throughout the dissertation I will emphasize the difference between pragmatic theories and models of pragmatic interpretation in comprehension. The cognitive system(s) responsible for pragmatic enrichment can be described at multiple levels, and it is important to keep them distinct. Following Marr (1982), I distinguish theories at the *computational* level from those at the *algorithmic* level. At the computational level, we seek to describe the purpose of a system, its inputs and outputs, and the general logic for the transformation from input to output. Theories of pragmatic enrichment from the semantic and philosophical literature are generally stated at the computational level. They describe the logic of the relation between literal meaning and speaker meaning. At the algorithmic level, we characterize how the inputs and outputs of the system are represented, and the specific processes that implement the transformation between them. Models of pragmatic enrichment in production or comprehension should be stated at the algorithmic level. They describe the steps that a speaker would go through to express his meaning through an utterance with a particular literal meaning, or the steps that a listener would go through to grasp the intended meaning of an utterance with a particular literal meaning.

To build an algorithm for implicature in comprehension, it is essential to have a good understanding of the structure of the problem at the computational level. In Chapter 2 I give a brief overview of three major theories of implicature, which have been influential in the psycholinguistic and developmental as well as theoretical literatures. I also introduce scalar implicature, which has been the primary focus of psycholinguistic and developmental work on implicature.

My general approach is to try to outline the components of an algorithm for comprehension, based on the constraints of the computational-level theory and what we know about comprehension in general in adults and children. Then I systematically review the evidence from the previous literature, with an eye to understanding how the results are informative about specific components of the algorithm. Much of the psycholinguistic and developmental literatures on implicature have attempted to test competing computational-level pragmatic theories, rather than models of implicature computation in comprehension. This approach is problematic, and has led to confusion about the implication of some of the findings. I attempt to reinterpret previous findings with more specific algorithm-level hypotheses in mind. I also report several experiments which help clarify previous results as well as contribute novel insights.

1.3 Overview of findings

Although I will consider multiple components of pragmatic interpretation over the course of this dissertation, my primary theme will be the question of how comprehenders determine what information in the context is relevant for understanding a speaker's meaning, and how they access that information efficiently to license and compute implicatures. I will argue that the variability in the accessibility of contextual information

in studies of adults' real-time comprehension of implicatures can explain a lot of the observed variability in processing costs. I will also claim that accessing relevant information is a source of vulnerability in children's ability to derive speaker meanings.

1.3.1 Implicature in real-time comprehension

The primary challenge for real-time pragmatic interpretation during comprehension is doing it incrementally. Most theories of implicature are stated at the level of the proposition (although I will discuss some exceptions). How can such a theory be implemented in a comprehension system that proceeds word-by-word? It is helpful to take cues from the literature on real-time syntactic parsing. Traditional theories of syntactic competence are also not stated in a way that lends itself well to incremental processing. Although the parser can only "see" one word of the input at a time, sophisticated predictive mechanisms and constant retrieval of previously processed input allow the parser to build sentence-level syntactic representations that correspond to those described by the computational-level theories. Although the path to uncovering such mechanisms for pragmatic interpretation is more opaque, increasingly sophisticated methods for observing real-time interpretation are beginning to make the problem more tractable.

In Chapter 3 I will review some of the growing literature on scalar implicature in adults' real-time comprehension. Most of this work has been directly inspired by pragmatic theory with, in my opinion, too little influence from mainstream psycholinguistics. I think the most promising directions for research on real-time pragmatic interpretation are those that incorporate concepts from work on syntactic processing, specifically prediction and retrieval. When and how do listeners predict what

the speaker's intended meaning will be? When and how do they retrieve previously encoded information to feed pragmatic inference? What extra-linguistic information do they track and encode for subsequent retrieval in the service of pragmatic inference?

With these goals in mind, I will re-interpret some of the most-cited results from the literature on pragmatic processing in an effort to better distinguish theoretical questions from questions about the algorithm, and establish a framework for future research. I contribute two experiments on adults' real-time comprehension of scalar implicature. Both attempt to clarify the source of processing costs that have been observed in previous studies on the comprehension of scalar implicature. The first distinguishes semantic and verification-related processes from pragmatic inference. The second investigates how relevant alternatives in context play their role in real-time implicature computation.

A common thread throughout the previous findings and my own work is the difficulty of identifying and accessing relevant information needed to infer the speaker's intended meaning. When considering implicature from the point of view of comprehension, this is a very pressing problem, but most theories stated at the computational level do not have much to say about how to solve it

1.3.2 Implicature in first language acquisition

In first-language acquisition, there is constant feedback between the information that children glean from context and from their representations of the linguistic signal. Children must guess at what speakers are trying to communicate based on the part of the linguistic signal that they understand and the part of the context that they understand, and then map that richer interpretation onto a richer linguistic representation. The sound-

meaning mapping is difficult to acquire because of the layers of hidden structure. Most research on first language acquisition focuses on the hidden layers that are closer to the sound side—phonology, words, and syntax—under the assumption, perhaps, that multiple levels of meaning are simply beyond the capacity of 0-3 year-olds. However, ignoring the problem of pragmatic-level meaning won't make it go away—it is an unavoidable property of children's early language experience.

Luckily, children are by no means as pragmatically obtuse as they are often painted in the literature. I will argue that children have the necessary pragmatic principles and inferential capacities for computing implicatures as early as they have been tested (around 3 years). In Chapter 4 I review the previous literature on children's understanding of implicature, focusing primarily on scalar implicature. I conclude that the low rate of generating pragmatically-enriched interpretations that has been observed in younger children is due to confusion about the experimenter's goals and difficulty accessing relevant information from the context. In Chapter 5, I turn to the somewhat dusty literature on children's understanding of indirect requests. I find that children are quite sophisticated in their interpretation of a variety of forms of indirect request—most likely because of their rich experience with them. I report a series of experiments investigating children's understanding of the limits of what can be considered relevant for an indirect request. In Chapter 6, I tackle an acquisition problem which has not previously been considered from a pragmatic perspective: young children's interpretations of belief reports. I argue that difficulty determining the relevance of beliefs in context leads children to non-adult-like speaker meanings.

The comparison of children's understanding of scalar implicature with other types of implicature turns out to be quite informative. When taken together, the findings on suggest that children do not have a general problem with enriched interpretation, or suffer from excessive "literalness". Children's surface behavior is by no means the same across scalar implicature, indirect requests, and belief reports. I argue that the same fundamental problem underlies children's non-adult-like behavior for all three implicature types: a difficulty identifying and accessing relevant contextual information. The differences can be understood by considering the different demands of accessing relevant information and the role of conventionality in each case.

2 Theories of implicature

Conversational implicature bridges the divide between direct speech acts (generally referred to as "literal meaning" in the experimental literature) and indirect speech acts ("speaker meaning"). The challenge for computational-level theories of implicature is to explain the logic of the relation between these two levels of meaning.

2.1 Three computational-level theories of implicature

Although the theoretical literature on implicature is enormous, accounts can be divided into three camps. The first is just Grice by himself: his ideas on implicature as a part of rational communication set the stage for most subsequent research. The second camp, the "neo-Griceans", includes those who consider their accounts to be developments of Grice's ideas, but who depart from him in various ways. The third camp is the proponents of Relevance Theory, an attempt at a cognitive account of utterance understanding.

The accounts promoted by these three camps are not directly comparable. Grice and the neo-Griceans are most concerned with an account of speaker meaning at the computational level, while relevance theorists pursue an account of utterance interpretation that is nearer to the algorithmic level (Saul, 2002). I focus on three points that are important at both levels. First, what is the relationship between the semantic representation and the literal meaning of the utterance (*what is said*)? Second, what are the respective roles of the speaker and the hearer in achieving successful communication? Third, how are implicatures derived in utterance interpretation?

2.1.1 Grice

Grice (1975) introduced the notion of implicature to capture the aspects of speaker meaning that are not part of "what is said". He wanted to distinguish saying from implicating in order to preserve a role for truth-conditional semantics in determining speaker meaning (Grice, 1957; 1975; Neale, 1992). He proposed to account for speaker meanings—and divergences between saying and meaning—by situating conversation in a general theory of rational interaction. Conversations, he claimed, are cooperative efforts in which the participants recognize a common goal and make rational contributions to move toward that goal. That is, participants in a conversation observe the Cooperative Principle (CP): "Make your conversational contribution such as is required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which you are engaged" (Grice, 1975, p. 45).

The Cooperative Principle only sets the goals for a conversational contribution; it does not specify how a contribution might satisfy or fail to satisfy the goals. Grice explains how the CP constrains contributions by describing "maxims" of different types, derived from the CP, which could act as guidelines for an appropriate contribution. Maxims of Quality dictate that the contribution should be true, at least to the speaker's knowledge. Maxims of Quantity state that the contribution should be informative enough, but not too informative. The maxim of Relation enjoins speakers to provide relevant contributions. Finally, maxims of Manner determine how something should be said, addressing prolixity and ambiguity.

Since hearers always assume that the speaker intends to be cooperative, they infer that the speaker intends an implicature when what the speaker says (the literal meaning)

fails to satisfy one of the maxims. The implicature is the missing link that makes apparently uncooperative utterances into appropriate conversational contributions. For example, let's consider an exchange like (1), slightly modified to make the conversational goals more explicit.

(3) Host: Would you like some coffee?

Guest: Actually I need to get to bed early tonight.

The literal meaning of the guest's response in (3) could be seen as violating multiple maxims. A relevant, informative response to the host's question should state whether or not the guest would like some coffee. Since what the guest says does not provide this information, the literal meaning of the utterance fails to satisfy maxims of Quantity and Relation. It may also violate a maxim of Manner, by being unnecessarily obscure.

Confronted with an irrelevant, underinformative, and obscure response, the hearer (the host) does not conclude that the speaker (the guest) has opted out of the normal rules of conversation. Rather, she assumes that the guest is being cooperative, and his intended meaning is actually fully relevant, informative, and expressed in an appropriate manner. She infers that the guest assumed that she would know that coffee would keep him awake, and thus if he wants to go to bed early he does not want coffee. The implicated meaning, then, is *No, I would not like some coffee*, which satisfies the maxims of Relation and Quantity. To explain the obscurity, the host can infer that the guest prioritized politeness over directness, as is culturally conventional. Providing an excuse is considered less rude than an outright refusal.

2.1.1.1 Semantic representation and literal meaning

It was important for Grice that the literal meaning is very closely related to the semantic representation. As Neale (1992) puts it, "what is said is to be found in the area where sentence meaning and [speaker's] meaning overlap" (p. 554).

If *what is said* is an assertion, it should be truth-conditional. Meanings that are conventionally associated with particular words, but not truth-conditional, should not be considered part of the literal meaning. For example, (4) and (5) are truth-conditionally equivalent—both would be considered true just in those worlds where Bill is both rich and honest. The difference in meaning between them—the fact that 'but' suggests a potential contrast between 'rich' and 'honest', while 'and' does not—is due to a "conventional" implicature. While conversational implicature is taken to arise from the rational nature of conversation (as I explain below), conventional implicature is not.

(4) Bill is rich but honest.

(5) Bill is rich and honest.

The close alignment between *what is said* and truth-conditional sentence meaning breaks down in cases where the speaker does not actually *mean* the proposition literally conveyed by the sentence. When a speaker utters a sentence sarcastically (6) or metaphorically (7), the meaning of the sentence is not what the speaker said, according to Grice. In such cases, the speaker is only "making as if to say" the literal meaning of the sentence.

(6) Keep crying—that'll probably help.

(7) You're the cream in my coffee.

2.1.1.2 Roles for speaker and hearer

For Grice, conversation is a game of coordination: the speaker and the hearer each take active roles, and each assumes that the other is able to access mutually known information. The speaker says something, and means something by saying it, with the assumption that the hearer can work out the intended meaning based on what was said: "the speaker thinks (and would expect the hearer to think that the speaker thinks) that it is within the competence of the hearer to work out [the implicated meaning]" (Grice, 1975, p. 50). The hearer's job is to infer the speaker's meaning, including any implicatures, based on the conventional meaning of the words in the utterance, the Cooperative Principle and associated maxims, linguistic and non-linguistic context, and background knowledge. The hearer assumes that the speaker intended him to compute this inference. Grice describes this aspect of the hearer's reasoning: "he knows (and knows that I know that he knows) that I can see that [the implicated supposition] *q* is required; he has done nothing to stop me thinking that *q*; he intends me to think, or is at least willing to allow me to think, that *q*; and so he has implicated that *q*" (p. 50).

It is important to note that although Grice (and other Gricean theorists) often describe implicatures by explaining the reasoning that the hearer would have to do to derive them, the speakers are the ones who create implicatures, by intending them. If a hearer derives a possible meaning of an utterance through some reasonable inference, but it is not the meaning that the speaker intended, then in fact the utterance does not carry that implicature. Put simply, hearers can be wrong about the implicatures of an utterance, but speakers by definition cannot be.

2.1.1.3 Implicature in utterance interpretation

Although the inferences that Grice describes sound like processes that the hearer could actually undertake, he meant them to be only definitional of a conversational implicature, not an algorithm for computing one. A conversational implicature "must be capable of being worked out" (p. 50), but there is no claim that such an explicit "working out" has to happen every time the speaker meaning includes an implicature. Nevertheless, it is useful to point out certain features of the account which could be (and in some cases, have been) applied to an algorithm for implicature comprehension.

What is said—the speaker's literal meaning—plays a critical role in the hearer's inference process. In general, the hearer assumes (and the speaker knows that he assumes) that the speaker said what he said with the intention of giving the hearer a particular notion about his intended meaning. Because of this assumption of cooperation, the literal meaning has a privileged role in the reasoning process compared to other sources of information. In an analysis of the meaning of an utterance, the semantic representation is "logically prior" to the literal meaning, which in turn is logically prior to conversational implicature (Neale, 1992, p. 543). It is the mismatch between the literal meaning and what would be expected given the conversational maxims that triggers the hearer to infer that an implicature is intended. The apparent priority of the literal meaning has led some psycholinguists to hypothesize that computing the literal meaning must happen before computing implicatures in real-time comprehension and over developmental time (e.g. Noveck, 2001; Bott & Noveck, 2004).

Although literal meaning plays an important role, the words of an utterance cannot directly trigger a conversational implicature. A given sentence type can be

associated with different implicatures when uttered in different contexts. The example in (3) demonstrates that there is not a deterministic route to a given implicature: different maxims can be included in the reasoning process, leading to slightly different interpretations. One of the main controversies in psycholinguistic work on implicature is the question of whether their derivation can be automatized. Although Grice's theory does not directly bear on this question (since, as I stated above, it is not meant to apply at the level of algorithm), his ideas are often cited in these debates.

2.1.2 Neo-Griceans

Neo-Griceans all consider themselves to be developing Grice's theories of meaning and conversation, but they are a diverse group (and there are occasional shouting matches in the literature about whose theory is most Gricean). They preserve the emphasis on cooperation and coordination between the speaker and the hearer. There have been multiple efforts to systematize the conversational maxims associated with the Cooperative Principle. Another important line of work focuses on the relationship between semantic representations (sentence meaning) and literal meaning (what is said), which Grice left somewhat underspecified. I discuss each of these developments, and their consequences for theories of utterance interpretation.

2.1.2.1 New conversational maxims

Horn (1984) proposed to simplify the conversational maxims by boiling them down to two principles. The Q Principle favors the hearer by pushing for more explicit information: "Say enough." Q-based implicatures tend to set an upper bound on the speaker's intended meaning: the hearer reasons that if the speaker knew enough to say more—by making a stronger assertion, or contributing additional assertions—he would

have. The R Principle favors the speaker by pushing for economy: “Don’t say too much.” R-based implicatures add extra suppositions to the speaker’s intended meaning: the hearer reasons that the speaker did not express his full meaning directly for some important reason—economy or politeness, for example.

One potentially positive effect of this simplification is that a given implicature can only be attributed to one maxim or the other. Under Grice’s theory, the implicature associated with the guest’s utterance in (3) could be derived from Quantity, Relation, or Manner maxims. In Horn’s system, we can simply say that the implicature is R-based. The two principles push against each other; the intended meaning of any given utterance relies on a point of balance between them, which will be more on either the Q side or the R side. This potentially simplifies the work a comprehender has to do: rather than considering multiple different types of explanations for a speaker’s utterance, the comprehender must infer where the speaker found a balance between Q and R.

Levinson (2000) reformulates Grice’s maxims into three “heuristics”. The I-Heuristic (“What is expressed simply is stereotypically exemplified”), the Q-Heuristic (“What isn’t said, isn’t”), and the M-Heuristic (“What’s said in an abnormal way isn’t normal”). The details of how these heuristics work and how they compare to Horn’s principles aren’t important for my purposes here (but see Traugott (2004) for a very helpful discussion). One aspect that is relevant is the overt emphasis on “normal”, conventional usage, which is more implicit in the work of Grice and other neo-Griceans. In building an algorithm for pragmatic enrichment in comprehension, it will be necessary to determine what exactly is meant by “normal” or “conventional”.

2.1.2.2 Semantic representations and literal meaning

A large body of work has demonstrated that the gap between semantic representations and *what is said* (literal meaning) is wider than Grice would have hoped. Effects of context and speaker intention on the literal meaning of utterances are pervasive. The semantic representation seems to dramatically underdetermine the literal meaning.

For example, the utterance in (8) could be used to express the thought that the speaker has never made coffee before, or the thought that the speaker hasn't made coffee recently, for some contextually-relevant purpose. The literal meaning of (9) depends on whether 'my arm' is meant to refer to the speaker's own arm, or perhaps a robotic arm that the speaker, an engineer, is currently working on. (10) is most likely to mean that the addressee isn't going to die from whatever injury is currently under discussion, but the same utterance could also be used by a scientist informing a participant in a successful drug trial that immortality had been achieved.

(8) I haven't made coffee.

(9) I broke my arm.

(10) You're not going to die.

The pragmatic enrichments that give (8)-(10) different literal meanings in different contexts have been termed "implicitures" (Bach, 1994; 2000; 2006) or, in Relevance Theory, "explicatures" (Sperber & Wilson, 1986; Carston, 1988; 2000). Although computing these aspects of literal meaning is an important part of pragmatic competence, I largely ignore them here. I only note that we should be careful about

assuming that comprehenders have easy access to literal meanings, since they do not arise directly from the bottom-up information in the linguistic signal.

2.1.2.3 Implicature in utterance interpretation

For the most part, Neo-Gricean theories do not have much more to say about utterance interpretation than Grice himself. However, as I mentioned above, conventionality plays a more explicit role, especially in Levinson (2000). The idea is that implications which are not part of the literal meaning can nevertheless become directly associated with certain words or expressions. They become part of the "normal" meaning, which the hearer should infer that the speaker intended unless something else about the utterance or context suggests otherwise.

This notion of conventionality does have potential consequences for algorithms for utterance interpretation. Levinson's ideas have been quite influential in the psycholinguistic literature on implicature, most likely in part because psycholinguists are quite comfortable with the concept of conventionality. Conventionality can be straightforwardly operationalized as frequency, and there are plenty of models available for how frequency affects real-time comprehension. Another reason that Levinson's local, conventional implicatures are appealing is that lexically-triggered operations are easier to implement in an incremental interpretation system. In a strict Gricean theory, in which implicatures are licensed only by the full utterance in context, it is more difficult to determine when and how implicatures would be computed in the course of real-time interpretation.

2.1.3 Relevance Theory

Relevance Theory (Sperber & Wilson, 1986; Wilson & Sperber, 1986) is a model of utterance interpretation, primarily from the point of view of the comprehender. It is therefore closer to the algorithmic level than Grice's theory or the different neo-Gricean theories. Relevance Theory takes as a starting point the Gricean idea that communication—and particularly comprehension—is fundamentally inferential. However, Relevance theorists reject the idea that the hearer's inferences require an acknowledgment of the speaker's intent, or any kind of "mutual" knowledge (the "speaker assumes that the hearer assumes that the speaker intends…" aspect of Gricean theories).

In order to formalize the reasoning process for computing implicatures, Relevance Theory reduces all the Gricean maxims to a single principle of *relevance*. Hearers interpret utterances by attempting to maximize their relevance with respect to the context.

A *context* is taken to be a set of propositions that are entertained by the hearer, each of which is associated with a "confirmation value" representing its truth value and the hearer's confidence in it. The starting context for any given utterance is the set of propositions that were most recently processed by the hearer. Propositions may be added to the context only if they are linked to a proposition that is already present in the context. A proposition that is connected to the current context through a relatively short chain of links is less costly to access than one that requires a longer chain.

A new proposition is relevant to the context only if it directly or indirectly affects the confirmation value of one of the propositions already in the context, or adds a new implication to the context. An implication can be added by deductive inference from the

combination of the new proposition with the propositions already present in the context. A proposition is more relevant if it changes the confirmation value of more propositions, or makes a larger change in a confirmation value, or adds more implications.

When a speaker utters something expressing a certain proposition, the hearer immediately combines that proposition with the current context by modifying confirmation values and deriving new implications through deductive inference. If the speaker means his utterance literally, the proposition that is most easily extracted from the utterance will be highly relevant to the context, in that it will change confirmation values or add new implications. If we've been talking about our plans for the fourth of July, the context you're entertaining might contain a few propositions relating to what I might be planning to do, like *Shevaun will see fireworks*. If you have no previous evidence about my plans, the confirmation value of these propositions would be relatively low, since your certainty would be low. If I then say, "I'm going to see fireworks," you can raise the confirmation value of *Shevaun will see fireworks* while lowering the confirmation value of incompatible propositions.

If the speaker utters something with an indirect intended meaning, the hearer has to do some work to figure it out. Suppose we've been talking about our plans for the fourth of July, and I say, "I don't want to do anything illegal." If there are no propositions in your context that are related to illegal activity, the proposition most immediately accessible from my utterance is not relevant. To make it relevant, you will have to add new propositions to your context by pursuing links in your personal encyclopedia of world knowledge. You will add new propositions until one of them can combine with the proposition expressed by my utterance to yield a new implication. You could start with

the new proposition from my utterance, and start listing all the illegal things I don't want to do: *Shevaun doesn't want to kill someone*, *Shevaun doesn't want to rob a bank*; etc. It might take you a while to get to the right one. Starting with one of the propositions in your current context might work better: *People often set off fireworks on the fourth of July*. From there you could relatively easily access *It's illegal to set off some kinds of fireworks in the city*. That proposition can be combined with the new one from my utterance to yield a new implication, *Shevaun doesn't want to set off any fireworks that would be illegal in the city*. However, there would be no need to stop there. Depending on what other propositions are easily accessible from your current context, you could derive all sorts of additional implications. If the proposition *Shevaun has a car* is easily accessible, perhaps because we were talking about it before, you might eventually get to the implication *Shevaun wants to set off some fireworks in the country*. Or if the previous conversation was about how some other people wanted to set off some fireworks in the city, you might get the implication *Shevaun doesn't want to do what those people are planning*.

The limit on the implications a hearer will derive from my single utterance is imposed by processing limitations. The hearer's goal is to get as many implications as possible with as little work as possible. Since the hearer's knowledge about the speaker's knowledge and potential goals is part of the context, the derived implications should be fairly close to the speaker's intentions. However, it is ultimately the hearer that determines the implications of the utterance, and it is possible that the utterance may end up having implications that the speaker didn't intend, or lacking implications that the speaker did intend.

2.1.3.1 Semantic representation and literal meaning

In Relevance Theory, the output of the syntax/semantics system is a bare, context-invariant structure; it is not propositional or truth-evaluable. The pragmatic system uses any tools available to derive a proposition from this linguistic representation, which then launches the deductive interpretation process described above (Carston, 2004). Thus, the major difference between Relevance Theory and Gricean/Neo-Gricean views is that there is not a distinct level of literal meaning, or "what is said". This can be seen as an advantage, since the debate over how to distinguish impliciture (or "explicature" in the terms of Relevance Theory) and implicature is moot: both are accomplished by the same kinds of pragmatic mechanisms. On the other hand, many theorists and naïve speakers alike have an intuition that although it may be difficult to determine where exactly the line should be drawn, there is a real and fundamental distinction between what is explicitly communicated and what is implicit. In Relevance Theory, explicitness is a continuum determined by the amount of processing it takes to get from the logical form to the meaning.

2.1.3.2 Roles for speaker and hearer

In Relevance Theory is a model of utterance interpretation from the hearer's point of view; the hearer is responsible for the work of computing implicatures. There is much less reliance on coordination between the speaker and the hearer. The speaker does have an intended meaning in mind, and she would like the hearer to be able to grasp that meaning. The hearer has an expectation that the speaker's utterance will have the property of *optimal relevance*: it will be relevant enough to justify the hearer's processing effort, and also as relevant as possible given the constraints of the speaker's current state

of knowledge, preferences, and goals. Thus, the speaker must keep the hearer's processing limitations in mind when choosing an utterance form: if the meaning is too obscure, the hearer will not be able to infer it without expending processing resources to an unreasonable degree.

In Relevance theory, the meaning of an utterance is only loosely tied to the semantic representation or to the speaker's intentions. Although both play a role, the interpretation that an utterance ends up having is largely determined by the state of the hearer. This is quite a dramatic departure from more traditional Gricean and Neo-Gricean views.

2.1.3.3 Implicature in utterance interpretation

Since Relevance Theory is a theory of utterance interpretation, in principle it should relate fairly directly to a model at the level of algorithm. However, it departs so radically from how we are used to conceiving of linguistic interpretation that it does not lend itself well to psycholinguistic models.

To build an algorithm of comprehension, you need to know what the relevant representations are like and what rules or operations allow you to go from one representation to the next. In Relevance Theory, the representations are straightforward enough—non-propositional logical forms on the input side, and propositions on the output side. The rules and operations, however, are completely unconstrained. When high-level reasoning processes are involved, it is hard to implement them in a way that seems plausibly efficient for the purpose of real-time comprehension. Comprehenders seem to understand speakers' intended meanings more quickly than they would be able to work out an explicit logical deduction.

The characterization of context as a set of propositions recently processed by the hearer is appealing in its simplicity of representation. However, it does not make the problem of identifying what is in a hearer's context any more tractable. Similarly, attributing processing costs to the complexity of the path from the current context to the implicated meaning has the advantage of being explicit at the computational level, but it does not help us guess what an actual knowledge structure in an actual human looks like. In order to have a substantive theory at the level of algorithm that makes any kind of useful predictions, we'll need a plausible model of the actual knowledge structures in question.

2.1.4 Summary

I have reviewed three major theories of implicature: Grice's account of rational, cooperative conversation, Neo-Gricean theories inspired by Grice's ideas, Relevance Theory, which moves all interpretation to the pragmatic level in an effort to build a cognitively plausible model of utterance interpretation.

Although Relevance Theory would seem to be most closely related to the algorithm-level models that psycholinguists are interested in, most of the psycholinguistic literature frames its questions and hypotheses in Gricean terms. I will do likewise for the sake of convenience.

2.2 An introduction to scalar implicature

The vast majority of psycholinguistic studies on implicature have focused on scalar quantity implicature, particularly that based on lexicalized scales like <'all', 'some'>, exemplified in (11). In this section, I give an overview of how the prominent

theories of implicature have treated scalar implicature, and discuss how the different theories might suggest different algorithms for scalar implicature computation.

(11) Some students graduated on time.

→ It is not the case that all students graduated on time.

2.2.1 Scales

Quantity implicatures in general arise when the literal meaning of an utterance is underinformative for the purposes of the conversation. Since the speaker is a cooperative participant in the conversation, the intended meaning is in fact more informative than the literal meaning, and the listener is able to figure that out.

Horn (1972) argued that scalar expressions give rise to quantity implicatures of a predictable form. A scale is a set of expressions of the same type which can be arranged in order of their informativeness. When a speaker uses a less informative expression on the scale, the intended meaning often includes the negation of a more informative expression on the scale. Thus, the scalar implicature sets an upper bound on the interpretation of the scalar expression. For example, if the speaker uttering (11) knew that in fact all students had graduated on time, he would have said so to make his contribution maximally informative. Since he didn't say so, it must be the case that he either knows for a fact that not all students graduated on time, or that he doesn't have enough information to assert that all did.

Although it turns out to be somewhat difficult to define what counts as "the same type" or to constrain possible "informativeness" orderings (Hirschberg, 1985; Matsumoto, 1995), there are nevertheless numerous scales that seem to consistently license quantity implicatures. These scales include (at least) quantifiers (12), logical

operators (13), adjectives (14), adverbs (15), and modals (16) (Horn, 1972; Gazdar, 1979; Levinson, 1983).

(12) <all, most, many, some, few>

(13) <and, or>

(14) <hot, warm>

(15) <always, often, sometimes>

(16) <must, may>

Scales like (12)-(16) seem to exist necessarily because of the meanings of the lexical items involved. Horn initially argued that the stronger members of the scale must entail the weaker ones. However, other scales have been argued to arise "ad-hoc" in context, leading to inferences of a similar form without the need for a pre-existing entailment relation. For example, an ordered set of stages in a process can be treated as a scale, as in (17) (Hirschberg, 1985).

(17) [Context: Mary is baking a cake.]

Mary has made the batter.

→ Mary has not put the cake in the oven.

2.2.2 Local vs. default vs. incremental

Scalar implicature is an attractive area for both theoretical and psycholinguistic investigation because it is relatively predictable and constrained, at least compared to implicature generally. As I discussed in the introduction, the flexibility of pragmatic inference allows any utterance to implicate almost anything, under the right circumstances. By contrast, scalar implicatures are associated with a particular part of the

utterance, rather than the entire utterance, and the intended meaning can be computed predictably. For psycholinguists, these features are crucial. A specific trigger for the implicature allows us to time-lock our observations of processing to the beginning of the computation. Consistent, predictable computations give rise to regular patterns of processing cost in different contexts.

However, these appealing features of scalar implicature also potentially limit the generalizability of the findings. If scalar implicatures are computed through specialized automatic processes triggered by certain lexical items, those processes would be irrelevant to the computation of less constrained conversational implicatures. In fact, multiple theories that propose a special status for certain conventional scalar implicatures have become quite prominent in both the theoretical and psycholinguistic literatures. Both Levinson (2000) and Chierchia (Chierchia, 2004; Chierchia, Fox, & Spector, 2008) argue that conventional scalar implicatures are computed locally—rather than at the utterance level, as in Gricean accounts—and arise in the "normal" case. Chierchia goes so far as to claim that conventional scalar implicatures are computed at the syntactic rather than the pragmatic level. Since these theories only apply to a select group of scalar implicatures, they have nothing to say about pragmatic inference generally. Thus, the worst case scenario (for us scientists) is that one of these "localist" theories is correct, and the phenomena that have been the subject of such intense scrutiny in the psycholinguistic and developmental literatures are actually irrelevant for understanding pragmatic interpretation in general. I take the more optimistic view—that these accounts are wrong. The phenomena that these theories were designed to explain can be accounted for with

Gricean utterance-level principles (Sauerland, 2004; Russell, 2006; Geurts, 2009), even if they are empirically accurate, which is also doubtful (Geurts & Pouscoulous, 2009).

The influence of these localist theories has led to some confusing discussion of several distinct but overlapping concepts: local implicature, default implicature, and incremental interpretation. A *local implicature* is one that is derived for some subpart of an utterance (e.g. a scalar expression like 'some students') and feeds into subsequent syntactic/semantic computations. There has been relatively little experimental work on local implicature (but see Geurts & Pouscoulous, 2009; Chemla & Spector, 2011).

A *default implicature* is one that arises in the normal case, all else being equal. The psycholinguistics literature has generally assumed that a default implicature must also be automatic and insensitive to context. This is not quite right, as Horn (2006) points out: you might shave every morning by default, but that doesn't mean that it will happen automatically. However, it is fair to say that without the assumption that it leads to automaticity, being a default implicature doesn't seem to mean that much. If the context must still be checked to ensure that the conditions for cancelling the default implicature do not hold, then default-ness doesn't save any work. It's also not clear what context would count as a default context, since most natural conversational situations provide a wealth of relevant information sufficient to infer the intended interpretation, implicature or no.

Finally, *incremental interpretation* is the idea that sentences are parsed and interpreted in real time "from left to right", building up as much structure and meaning as possible based on each word as it arrives, plus any top-down expectations about the likely intended structure or meaning. Whether an implicature is local or default are questions

that can be stated at the computational level of description. Incrementality, by contrast, is a feature of the algorithm. It should be apparent why these concepts have been related in the literature: an implicature that is both local and default will be relatively easy to implement in incremental interpretation. A particular scalar expression would trigger the implicature without the need for too much evaluation of the context. However, it is important to keep in mind that non-local, non-default implicatures may also be implementable in incremental interpretation. The psycholinguistic literature is full of examples of the parser making structural or interpretive leaps based on incomplete information in the bottom-up input. Implicature need not be any different.

2.2.3 Summary

Scalar implicatures are relatively constrained and predictable compared to implicature in general. They seem to be triggered by certain types of expressions, and the implicated meaning can be computed using the same set of operations every time. Although these features are beneficial for psycholinguistic and developmental research, it is also possible that they are to be explained by mechanisms that are fundamentally different from those involved in other types of implicature. I am optimistic that this is not the case, but for the most part this worry will not be relevant to the rest of my discussion.

3 Scalar implicature processing in real-time comprehension

How do comprehenders grasp the intended meaning of a speaker's utterance in real time? As an utterance unfolds over time, the combination of information from the utterance with other sources of information makes several levels of meanings available. Lexical meanings become available when an uttered word is accessed in long-term memory. The semantic content is the product of the lexical meanings combined with syntactic structure, as well as the valuing of any contextual indices. The direct speech act—"what is said"—may require some additional relativizing to context. Finally, the intended message may include indirect speech acts which can only be accessed by considering the speaker's possible intentions in context, which may be constrained by knowledge of the world or the speaker in particular. The level of the intended message exists regardless of whether it contains any indirect speech acts: the goal for comprehension is always to understand the speaker's meaning, not just decode a semantic representation. We focus on indirect speech acts and the implicatures that generate them not because they are qualitatively different from the "normal" case, but because the gap between the literal and intended meaning makes the two levels more distinctly observable.

Numerous psycholinguistic studies have investigated the real-time processing of implicature, especially scalar implicature (see section 2.2 for an introduction to scalar implicature). Researchers have often attempted to use psycholinguistic results to arbitrate between different computational-level theories—Relevance Theory vs. Neo-Gricean theories, for example. This is a mistake, however, as computational-level theories are each compatible with many different algorithms for real-time comprehension, and vice

versa. Here I generally eschew hypotheses stated at the computational level, in favor of specific questions about the algorithm. Where relevant, I note how a given feature of the algorithm may relate to the computational-level theories.

In sections 3.1-3.2, I review the evidence that comprehenders do compute scalar implicatures on a relatively short time scale and in a way that is sensitive to context, mostly consistent with the generalizations that have been reported in the theoretical literature. In section 3.3 I review evidence that computing implicature-enriched interpretations of scalar expressions is more costly than computing literal meanings. In the rest of the chapter, I delve into the possible sources of this processing cost. The evidence from the previous literature is far from conclusive, and I report two novel experiments that help to clarify some of the previous results.

3.1 Scalar implicatures in judgments

Although context-free judgments may seem far removed from the task of comprehending speaker meaning generally, most of the theoretical pragmatic literature is based on low- or no-context judgments generated by individual theorists. It is useful to begin by confirming that naïve speakers arrive at the same interpretations and judgments, particularly for expressions that are considered to trigger implicatures by default in some theories. It may be that the perceived regularity of scalar implicature is due to a bias inherent in the procedure of collecting introspective judgments: by considering whether an utterance implicates a proposition, the introspector implicitly introduces that proposition as a relevant possibility, thereby licensing the implicature (Geurts, 2009).

Noveck (2001) used truth-value judgment tasks to investigate whether adults (and children) readily generate scalar implicatures for statements with underinformative

modals (Experiment 1) or quantifiers (Experiment 3). In Experiment 1, participants judged modal statements like (18) with respect to a visual display and certain rules about the possible contents of a closed box. In Experiment 3, participants judged statements with quantifiers (19) with respect to their own world knowledge. This latter task was called the "statement evaluation task", to distinguish it from truth-value judgment tasks which usually provide more context for evaluation.

(18) There might be a parrot in the box. [*Context: There* must *be a parrot in the box.*]

(19) Some giraffes have long necks.

In both experiments, the critical trials involved a statement which was true under a "logical" (literal) interpretation and false under a "pragmatic" (implicature-enriched) interpretation. For example, (19) is logically true, since it is the case that a non-zero number of giraffes have long necks. However, since the expression "some giraffes" is logically consistent with any number of giraffes—whether it be two giraffes or all the giraffes in the world—it is underinformative. With an upper-bounding implicature, the interpretation of (19) would be something like (20). Under this pragmatic interpretation, (19) is false.

(20) Some but not all giraffes have long necks.

Noveck found that adults often—but not always—rejected underinformative statements: 65% for modal statements like (18) and 59% for statements with quantifiers like (19). These results suggest that while upper-bounded interpretations of scalar expressions are by no means universal (as I discuss more in the next section), they are clearly available to naïve listeners.

3.2 Context-sensitivity

The debate over whether conventional scalar implicatures arise by "default" led to some research on whether people are sensitive to context when they generate (or fail to generate) these implicatures. Although the "defaultist" views advocated by Levinson and Chierchia do not require conventional scalar implicatures to be entirely insensitive to context—such a view would be untenable—they certainly operate under the assumption that scalar expressions are overwhelmingly interpreted with upper-bounded meanings. Upper-bounded interpretations are assumed to be the "normal" and most frequent case, while literal lower-bounded interpretations are the exception.

I will now review some of the abundant evidence that this assumption is incorrect. Comprehenders are very sensitive to context when interpreting scalar expressions. More importantly, when little or no context is provided, they choose an interpretation arbitrarily, preferring upper-bounded interpretations no more often than lower-bounded interpretations on the whole. Although there is relatively little evidence available about the frequency of upper-bounded interpretations in the "real world", recent corpus-based work suggests that at least for 'some', upper-bounded interpretations are in fact much less frequent than lower-bounded interpretations overall (Degen, 2013, submitted).

Adult judgments of underinformative statements in a variety of contexts have been collected as a control in studies on children's pragmatic competence. Although there is substantial variation in adults' judgments across different studies, the differences form an interpretable pattern: richer contexts lead to a dramatically higher rate of implicatures. For example, Guasti and colleagues (2005) compared judgments in Noveck's statement evaluation task to a more naturalistic truth-value judgment task involving an illustrated

story. In the statement evaluation task (their Experiment 1), adults rejected underinformative utterances 50% of the time, replicating the middling rate of pragmatic interpretations observed by Noveck (2001). In the story-based truth-value judgment task (Experiment 4), adults rejected underinformative utterances 83% of the time. The authors attribute this much higher rate of pragmatic interpretations to the shared conversational context provided in the story-based task. In the statement evaluation task, participants must guess at whether an upper-bounded interpretation is relevant. In the story-based task, the upper bound is clearly relevant, thus licensing the implicature. Papafragou and Musolino (2003) also observed a very high rate of pragmatic interpretations (92.5%) in a similar story-based truth-value judgment task.

Pouscoulous and colleagues (2007) observed a substantial increase in the rate of pragmatic interpretations in an action-based task compared to a picture-based truth-value judgment task. Adults rejected underinformative utterances like (21) only 47% of the time in a truth value judgment task, but responded to instructions like (22) by enforcing an upper-bounded interpretation of 'some' 86% of the time. Placing the underinformative expression inside a request in an action-based task makes the goal of the utterance clear. A descriptive statement like (21) presented out of context has no conversational purpose; in fact, it simply would not occur outside an experimental setting. Participants, therefore, are in the situation of having to guess what the experimenter is trying to test them on. By contrast, a request like (22) has an obvious goal: the speaker wants to enact a certain change on the physical environment. In that context, alternative actions are relevant: *should I put tokens in all the boxes, or only some of them?* These relevant alternatives are what license the scalar implicature.

(21) Some turtles are in the boxes.

(22) I would like some boxes to contain a token.

Although these task differences are suggestive, better evidence for context sensitivity comes from controlled experiments where context was manipulated as a factor.

Chierchia and colleagues (2001) investigated adults' interpretations of statements involving disjunction, such as (23). 'Or' is logically inclusive, allowing the interpretation in (24), but is often pragmatically enriched to an exclusive interpretation via the implicature in (25). They found that adults derived a pragmatic (exclusive) interpretation of disjunction 100% of the time in upward entailing contexts like (23), but almost never (5%) in downward entailing contexts like (26). Although Chierchia argues that the lack of scalar implicatures in downward-entailing contexts supports his grammatical account of scalar implicatures, the important point for us here is that adults' interpretation of scalar expressions is highly sensitive to linguistic context. However, since this kind of context would be easily accommodated by otherwise "default" implicature mechanisms, I now turn to manipulations of discourse context.

(23) Every boy chose a bike or a skateboard.

(24) Every boy chose a bike or a skateboard, possibly both.

(25) No boy chose both a bike *and* a skateboard.

(26) Every dwarf who chose a banana or a strawberry received a jewel.

Zondervan (2009) investigated adults' interpretations of disjunction in different discourse contexts. He found that adults rejected underinformative utterances involving

disjunction more often when the disjunction was part of a constituent that provided an answer to the Question Under Discussion (QUD). In Experiment 1, for example, adults rejected underinformative statements like (27) 73% of the time when the explicit QUD focused the disjunctive expression (28), compared to only 55% of the time when it focused the subject (29). This pattern held even when the QUD was left implicit (Experiment 3). Zondervan (2011) replicated these effects with underinformative utterances with the scalar quantifier 'most', as in (30).

(27) Harry brought bread or chips.

(28) *Focus:* What did Harry bring?

(29) *Non-focus:* Who brought bread or chips?

(30) a. *Focus*: How many of the students drank beer?

b. *Non-focus*: What did most students drink?

c. *Target sentence*: Most students drank beer.

Several reading experiments have found evidence that adults' interpretations of scalar expressions are context sensitive (Breheny, Katsos, & Williams, 2006; Bergen & Grodner, 2012; Hartshorne & Snedeker, submitted). These studies all employ the same measure to gauge interpretation. In each case, the triggering scalar expression is 'some of the X', which can be interpreted logically as 'at least some of the X' or pragmatically as 'some but not all of the X'. The mention of 'some of the X' is followed by a mention of 'the rest', as in the example in (31) from Breheny and colleagues' (2006) Experiment 2. The anaphoric 'the rest' requires a bridging inference to be interpreted as 'the rest of the X'. This bridging inference should be easier if 'not all of the X' has already been computed. Thus, higher processing costs at 'the rest' should be associated with logical

interpretations of ‘some of the X’, while lower processing costs are associated with pragmatic interpretations.

(31) Some of the consultants had a meeting with the director. The rest did not manage to attend.

In their Experiment 2 Breheny and colleagues found, similar to Zondervan (2009; 2011), that pragmatic interpretations of ‘some of the X’ were more likely when the ‘some’ expression was focused. In Greek, which has flexible word order, the entity in subject position is very likely to be interpreted as focused. Thus, when the ‘some’ expression appeared in subject position as in (32) it was likely to be considered the answer to an implicit Question Under Discussion (QUD) like, “What happened with the consultants?” This accommodated QUD would license the upper-bounding implicature. By contrast, when the ‘some’ expression was in object position as in (33), it would not be considered focused. No such QUD would be accommodated into the discourse, and the scalar implicature would not be licensed.

(32) (Only) some of the consultants had a meeting with the director. The rest did not manage to attend.

(33) The director had a meeting with (only) some of the consultants. The rest did not manage to attend.

Reading times at ‘the rest’ were compared for sentences with and without the disambiguating ‘only’. When the ‘some’ expression was in subject position, there was no difference based on the presence of ‘only’, suggesting that the upper-bounding implicature was readily computed. When the ‘some’ expression was in object position,

reading times at ‘the rest’ were significantly longer when only was absent, suggesting that the upper-bounding implicature was not initially computed.

In their Experiment 3, Breheny and colleagues found that pragmatic interpretations of ‘some of the X’ were more likely (as evidenced by lower processing cost at ‘the rest’) when the preceding context explicitly mentioned the upper bound ‘all of the X’, as in (34), compared to when the context did not mention the set of X at all, as in (35).

(34) Mary asked John whether he intended to host all his relatives in his tiny apartment. John replied that he intended to host some of his relatives. The rest would stay in a nearby hotel.

(35) Mary was surprised to see John cleaning his apartment and she asked the reason why. John told her that he intended to host some of his relatives. The rest would stay in a nearby hotel.

Bergen and Grodner (2012) found that pragmatic interpretations of ‘some of the X’ were more likely when the context suggested that the speaker had “full knowledge” (36), as opposed to “partial knowledge” (37), which was relevant to the statement. Reading times at *the rest* were longer in the partial knowledge condition, suggesting that the upper-bounding implicature had not initially been computed for ‘some of the real estate investments’.

(36) At my client’s request, I meticulously compiled the investment report. Some of the real estate investments lost money. The rest were successful despite the recent economic downturn.

(37) At my client's request, I skimmed the investment report. Some of the real estate investments lost money. The rest were successful despite the recent economic downturn.

Finally, Hartshorne and Snedeker (submitted) found that pragmatic interpretations of 'some of the X' were less likely when the expression was embedded in a conditional—a downward-entailing context—as in (38), than when it was not (39), as evidenced by longer reading times at 'the rest'.

(38) If Mary did some of her homework this morning before breakfast, then the rest must be done later today.

(39) Mary did some of her homework this morning before breakfast, and the rest must be done later today.

To summarize the experimental findings, adults are more likely to compute a scalar implicature when the context specifically licenses it, than when a neutral context or no context is provided. When the context makes the purpose of an utterance clear, adults are more likely to interpret it pragmatically rather than logically (Guasti, et al., 2005; Papafragou & Musolino, 2003; Pouscoulous, Noveck, Politzer, & Bastide, 2007). Adults are more likely to derive an upper-bounded interpretation of a scalar expression (1) when the scalar expression is in focus, providing an answer to the Question Under Discussion (Zondervan, 2009; 2011; Breheny, Katsos, & Williams, 2006), (2) when the relevant alternative is mentioned and relevant in the context (Breheny, Katsos, & Williams, 2006), (3) when the speaker has knowledge of the relevant alternative (Bergen & Grodner, 2012), and (4) when the scalar expression is not in a downward-entailing environment

(Chierchia, Crain, Guasti, Gualmini, & Meroni, 2001; Hartshorne & Snedeker, submitted). If scalar implicatures were computed by default, we would not expect such sizeable context effects, because implicatures would also be computed in the low-context conditions.

Recent corpus-based work by Degen (2013, submitted) supports the conclusions of the psycholinguistic studies. She found that upper-bounded interpretations of the quantifier 'some' are in fact relatively infrequent overall—an estimated 28% of all uses of 'some'. The probability of an upper-bounded interpretation was higher in the presence of certain grammatical and prosodic cues, and in certain discourse contexts.

Taken together, all of these studies provide compelling evidence that adults do not consistently compute scalar implicatures in the absence of a licensing context. Upper-bounded interpretations do not seem to be the "default" option, although again, it's not entirely clear what that would look like. In any case, I now set aside the issue of default-ness, and turn to questions about how and when scalar implicatures are computed during real time comprehension.

3.3 Processing costs associated with scalar implicature

Processing cost is the primary currency of psycholinguistics, so to get at the "how" and "when" of real-time implicature comprehension, we must ask whether implicatures are costly to compute compared to literal meanings. Implicature-related processing costs—like any other psycholinguistic observation—can be attributed to numerous different sources. Although most researchers offer a specific interpretation of their results, much of the evidence is in fact compatible with multiple models of real-time interpretation. In this section I simply review the evidence that scalar implicature is often

associated with increased processing cost. In the next section I outline a few different interpretation algorithms that could lead to the observed costs, and discuss the somewhat limited evidence that might narrow down the options.

Bott & Noveck (2004) used the statement evaluation task developed in Noveck (2001), with the enhancement that participants' judgments were also timed. In their Experiment 3, they found that "false" (pragmatic) judgments of sentences like (40) were slower than "true" (logical) judgments. In Experiment 4, they showed that when response time was controlled by forcing participants to give their responses after an auditory signal, judgments elicited after a shorter interval were less likely to reflect pragmatic (upper-bounded) interpretations (28%) than those elicited after a longer interval (44%).

(40) Some giraffes have long necks.

De Neys and Schaeken (2007) directly tested the hypothesis that pragmatic judgments require more processing resources by eliciting truth value judgments under conditions of high and low cognitive load induced by a spatial memory task. They found that participants were slightly less likely to interpret sentences like (40) pragmatically when under a higher cognitive load (73%) compared to a lower cognitive load (79%). Response latencies for pragmatic interpretations were about 700ms longer on high-load trials than low-load trials, but latencies for logical interpretations were not affected by cognitive load. These results suggest that pragmatic interpretations were more costly than logical interpretations.

Marty, Chemla and Spector (2013) also investigated the rate of pragmatic judgments under cognitive load. They improved on De Neys and Schaeken's work by eliciting judgments for descriptions of simple scenes, rather than the Noveck-type

statements of fact, and including many more trials per condition and participant. In each trial, participants had to memorize either 2 (low load) or 4 (high load) letter sequences. After judging a sentence like (41) with respect to a display of colored dots, they had to reproduce the letter sequence in reverse order. Participants rarely rejected the underinformative utterances, but they were slightly more likely to do so in the low-load (89%) than in the high-load (81%) condition.

(41) Some dots are red.

Numerous additional studies have also found that pragmatic interpretations are more costly than literal interpretations (Katsos, Breheny, & Williams, 2005; Breheny, Ferguson, & Katsos, 2013; Huang & Snedeker, 2009; Huang & Snedeker, 2011; Bergen & Grodner, 2012; Bott, Bailey, & Grodner, 2012). Since each of these studies also provides more detailed information about the time-course of interpretive processing, I defer an in-depth discussion of their results to the following sections.

It is worth noting that not all studies have observed higher processing costs for pragmatic interpretations: sometimes there is no difference in cost for the two types of interpretation. For example, Feeney, Scrafton, Duckworth & Handley (2004) found no difference in reaction times between logical and pragmatic responses in a verification task very similar to that of Bott & Noveck (2004). They argue that since the logical and pragmatic responses in Bott and Noveck's study came from mostly non-overlapping groups of participants, it is unwise to draw conclusions based on the reaction time differences. In their Experiment 3, they used a version of the verification task designed to prevent participants from developing a pragmatic or logical response strategy. Most of their participants produced both logical and pragmatic responses to underinformative

sentences, and there was no difference in reaction times between the two response types. However, they also found that 'true' (logical) responses to infelicitous 'some' sentences were slower than 'true' responses to felicitous 'some' sentences. They argue that this apparent processing cost is due to some participants initially computing a pragmatic interpretation and then suppressing it to respond logically.

Several reading studies have also failed to find differences in processing cost for literal and pragmatic interpretations. (Grodner, Klein, Carbary, & Tanenhaus, 2010; Breheny, Ferguson, & Katsos, 2013; Politzer-Ahles & Fiorentino, 2013; Hartshorne & Snedeker, submitted). I will review these studies in the following sections as I discuss possible sources of the processing cost.

To summarize, computing upper-bounded meanings for scalar expressions is often, but not always more costly than computing literal lower-bounded meanings. Truth-value judgments reflecting upper-bounded interpretations are slower (Bott & Noveck, 2004), and are less likely to occur under time pressure (Bott & Noveck, 2004) or high cognitive load (De Neys & Schaeken, 2007; Marty, Chemla, & Spector, 2013). However, this cost is not always observed for judgments (Feeney, Scrafton, Duckworth, & Handley, 2004). Studies of processing without a judgment component (reading and visual-world eye-tracking studies) vary in whether costs for upper-bounded scalar implicatures are observed. Since this variation may be informative about the source of the cost, I discuss it in more detail in sections 3.5-3.7.

3.4 Understanding processing costs: possible algorithms

The results reviewed so far demonstrate that adults readily compute scalar implicatures in appropriate contexts, and that this computation is sometimes costly. What

can this cost tell us about the algorithm for computing scalar implicatures in the course of comprehension? To begin, we need an idea of the space of possibilities in characterizing an algorithm for scalar implicatures. One of my goals in this discussion is to make clear that there are many more possible algorithms than are generally acknowledged in the psycholinguistic literature. Researchers tend to choose two possible algorithms to compare (often aligning them with neo-Gricean theories vs. Relevance Theory), without regard to the many alternatives which could predict the same results.

3.4.1 Algorithms for computing scalar implicatures

We are interested in characterizing an algorithm whose purpose is to compute upper-bounded interpretations of scalar expressions in cases where that is the meaning that the speaker intended. The potential inputs to the algorithm include any information available in the linguistic content of the utterance and any information available about the conversational context, the speaker, or the world at large. For concreteness, let's consider algorithms for computing an upper-bounded interpretation of the utterance in (42).

(42) John ate some of the cookies.

I will discuss two sources of variation between different potential algorithms: (1) how is the implicated meaning accessed? and (2) how does context play its role?

3.4.1.1 How is the implicated meaning accessed?

There are three logically possible methods for accessing the implicated meaning of an utterance. The first is to access the meaning *directly*, based on information present in the linguistic content. For example, the lexical item 'some' could be directly linked to the meaning *some but not all*. This type of access is most consistent with the default,

local models of Levinson (2000) and Chierchia (2004). It also might be consistent with a constraint-based account (Degen & Tanenhaus, 2011), in which all the potential meanings of an expression could be available in parallel, but more or less preferred depending on how the context satisfies different constraints.

The second method is to compute the implicated meaning using a set of procedures specific to scalar implicature, triggered by certain scalar expressions. These might be as in (43).

(43) a. Compute the literal meaning of the scalar expression. → *(at least) some*

b. Retrieve the lexical scale associated with the expression. → <*all, some*>

c. Find the next most informative member of the scale. → *all*

d. Create a copy of the original utterance, replacing the scalar with the more informative version retrieved in step (c). → *John ate all of the cookies.*

e. Negate the modified copy of the original utterance. → ¬ *(John ate all of the cookies.)*

e. Add the modified copy to literal meaning computed in (a). → *John ate at least some of the cookies, and it is not the case that he ate all of them.*

The third method for computing the implicated meaning is to use completely general inference procedures, like those invoked by Grice. This method would not be restricted to scalar implicatures, but could be used for any type of conversational implicature.

It should be noted that most Gricean, neo-Gricean, and Relevance theoretic accounts would be consistent with all three of these methods of accessing the implicated meaning.

3.4.1.2 How does context play its role?

There are likewise three logical possibilities as to how context could play its role in interpretation: before, during, and after. Depending on the timing, we assume different mechanisms. By "context" here, I mean more specifically the pieces of contextual information that are relevant for interpreting a given utterance. Conversational participants must have some way of encoding contextual information into discrete units. I'm interested in mechanisms for identifying and accessing the contextual units that are relevant for a particular interpretive task.

If conversational participants continually maintain a model of the Question Under Discussion, they could use it to predict potential contributions to the conversation and constrain the interpretation of utterances. For example, in a conversation about how many cookies are left in the jar, the listener may carry forward an expectation that the speaker will provide information about the whole set of cookies. This prediction may be represented at a more conceptual level—an expectation about the content of the speaker's future utterances—which may in some cases lead to more specific lexical predictions. Either way, upon encountering 'some' in (42), the listener anticipates that the entire set of cookies is relevant, and thus immediately interprets the quantifier as upper-bounded. On the other hand, in a conversation about why John is feeling sick, the listener has no such expectation, and will interpret 'some' as lower-bounded.

Alternatively, context could have its effect during the interpretation of the critical expression. In the course of interpreting a given utterance or part of an utterance, the listener may choose to access potentially relevant contextual information as input to the computation. Instead of carrying forward an expectation about what the speaker will

contribute, the listener is ready to access relevant information on the fly. Upon hearing 'some of the cookies' in (42), the listener searches the previous discourse for the set of cookies under discussion to determine the role of the cookies in the discourse. If the whole set of cookies is relevant, this information will feed into a decision to interpret 'some' as upper-bounded.

Finally, context could have its effect only after initial interpretation has taken place. For example, it could be that listeners always interpret utterances literally at first, and try to fit the literal meaning into the context. If there is a mismatch between the literal contribution of the utterance and the needs of the current discourse, the listener launches additional pragmatic processes to change the interpretation.

Contrary to common assumptions in the psycholinguistic literature, the role of context in the algorithm is orthogonal to the question of how the implicated meaning is accessed. Any of the access methods described in the previous sub-section could be combined with any of the context mechanisms described here. For example, it is often assumed that if the implicated meaning is accessed directly, context could only play a role after the fact. However, it would be entirely possible for a context-sensitive predictive mechanism to determine which meaning of a scalar expression should be accessed.

3.4.2 Sources of observed processing costs

Given this wide array of possible algorithms for computing pragmatically-enriched interpretations, what can we glean from the existing evidence? I see three possible explanations for the costs observed in the literature for "pragmatic" upper-bounded interpretations.

First, generating or evaluating an upper-bounded interpretation of a scalar expression may be more costly than a lower-bounded interpretation, regardless of any pragmatic processes that may be involved. For example, if computing the necessarily upper-bounded interpretation of (44) is more costly than the necessarily lower-bounded interpretation of (45), then this same cost difference should also be observed in the computation of pragmatic compared to logical interpretations of underinformative sentences like (46). In section 3.5 I review previous studies that indicate that upper-bounding is indeed costly, at least in verification tasks, regardless of the need for pragmatic inference. I also report an experiment that addresses this issue.

(44) Only some elephants are Indian.

(45) At least some elephants are Indian.

(46) Some elephants are Indian.

Although a difference in meaning can explain some of the observed processing costs, it probably cannot explain all of it. The hypotheses that are most commonly offered in the literature have to do with how the implicated meaning is computed. If the upper-bounded meaning of 'some' can be accessed directly, it should not be more costly than computing a lower-bounded meaning. If some more elaborate procedure is involved in computing the meaning—whether it be a domain-specific process or a general inference mechanism—then upper-bounded meanings should be more costly. In section 3.6 I review evidence that literal meanings are available earlier than implicated meanings.

A final type of explanation for the observed processing cost has to do with how context has its effect on interpretation. The earlier that contextual information can be integrated into ongoing interpretation, the less delay we should observe for scalar

implicatures relative to literal meanings. In section 3.7 I review the evidence that the availability of relevant contextual information may affect the cost of scalar implicatures. I also report an experiment designed to investigate the role of the accessibility of contextual information in previously observed processing costs.

3.5 The cost of upper-bounded interpretation

3.5.1 Previous evidence

One way to observe the cost of upper-bounding without the confounding effects of inference is to specifically instruct participants on how to interpret the underinformative scalar expression of interest. If the comprehender knows that in the context of the experiment, 'some' is always intended to mean 'some but not all', there is no need to perform an inference to arrive at the intended interpretation.

Bott & Noveck's (2004) instructed participants to interpret 'some' either as "some and possibly all" ("certains et peut être tous") or "some but not all" ("certains mais pas tous"). In their Experiment 1, they gave participants the two types of instruction in two separate experimental sessions. They found that judgments of sentences like (47) were slower and less accurate when participants were instructed to use an upper-bounded interpretation (and thus reject the sentence), compared to when they were instructed to use a lower-bounded interpretation (and thus accept the sentence).

(47) Some giraffes have long necks.

Experiment 2 used a slightly different task to ensure that the difference in reaction times was not due to the need to provide a "false" response for an upper-bounded interpretation and a "true" response for a lower-bounded interpretation. Each target

sentence was preceded by the statement, “Mary says that the following sentence is true/false.” Participants were asked to indicate whether they agreed with Mary. This paradigm makes it possible to compare accuracy and reaction times for the same response (“agree”) for upper-bounded and lower-bounded interpretations. In this experiment, separate groups of participants were given each type of instruction. The results were the same as in Experiment 1: judgments were slower and less accurate for participants instructed to use an upper-bounded interpretation of ‘some’. However, if there is a special cost for rejecting a statement, it presumably arises not from the mere need to say “false”, but rather from the verification process that generated the rejection. Thus it seems unlikely that this version of the task really controls for the potential difference between true and false responses.

Bott, Bailey & Grodner (2012) also compared judgments provided by groups of participants given different instructions about the intended interpretation. They used a speed-accuracy tradeoff (SAT) task to gain a finer-grained measure of the processing costs associated with different interpretations of 'some'. In their Experiment 1, they replicated Bott & Noveck’s (2004) finding that participants instructed to use an upper-bounded interpretation of ‘some’ provided slower and less accurate judgments.

The findings so far suggest that judgments reflecting upper-bounded interpretations seem to be more costly, even when participants are instructed to use such an interpretation, so that no inference is required. A second strategy for isolating upper-bounding from inference is to compare processing costs for upper-bounded interpretations that are pragmatically-derived and semantically-derived. As mentioned

above, ‘only’ semantically enforces an upper-bounded interpretation of some. Several studies have investigated the processing of ‘some’ compared to ‘only some’.

Bott, Bailey & Grodner’s (2012) Experiment 2 compared judgments in an SAT task to sentences like (48) and (49). All participants were instructed to use an upper-bounded (pragmatic) interpretation of ‘some’. They found that although asymptotic accuracy was the same for both sentence types, reaction times were significantly faster for ‘only some’. Specifically, the model of accuracy conditioned by reaction time required an earlier intercept in the ‘only some’ condition.

(48) Some elephants are mammals.

(49) Only some elephants are mammals.

It is difficult to determine why ‘only some’ was easier to process than ‘some’ in Bott and colleagues’ study. Since participants were instructed which interpretation to use (and completed a substantial amount of practice with feedback on how to judge the critical sentences), it is unlikely that any inference was necessary in the ‘some’ condition. One possibility is that even though only one interpretation was allowed, participants obligatorily computed the logical, lower-bounded interpretation of ‘some’ before or simultaneously with the upper-bounded interpretation. Another possibility is that ‘only’ triggers the upper-bounding process earlier, so that it finishes sooner even though it requires the same amount of total processing time.

Marty and Chemla (2013) used a dual-task paradigm like that of De Neys & Schaeken (2007) to investigate the cost of upper-bounded interpretations of ‘some’ compared to ‘only some’. Participants judged sentences like (48)-(49) while holding in memory a simple (low-load) or complex (high-load) pattern of dots. Participants were not

instructed on how to interpret ‘some’ in the sentences where it was ambiguous. Participants rejected false ‘only some’ sentences at a relatively high rate in both high load (72%) and low load (79%) trials. By contrast, participants rejected infelicitous ‘some’ sentences at a lower rate overall, and they were significantly less likely to do so under high load (34%) than under low load (54%). Although the authors do not emphasize this finding, it is worth noting that there were no differences in accuracy based on sentence type or memory load for true/felicitous sentences like (50)-(51). This could be because upper-bounding is not costly with ‘only’, or because participants automatically computed an upper-bounded interpretation of ‘some’.

(50) Some reptiles are snakes.

(51) Only some reptiles are snakes.

These results suggest somewhat more convincingly that computing an upper-bounded interpretation of ‘some’ is more cognitively demanding than computing the same interpretation for ‘only some’. Unfortunately, the authors do not report whether individual subjects were consistent in their response strategy, as has been observed in other studies. In order to attribute the higher cost for upper-bounded interpretations to an inference, we need to be certain that participants were actually performing an inference, rather than deciding ahead of time how they would interpret ‘some’. Another concern is that the observed differences might be attributable to the fact that ‘only’ triggers the upper-bounding process at the beginning of the sentence, while the bare ‘some’ sentences may need to be interpreted further before deciding on the intended interpretation of the quantifier.

Although several reading studies have also compared ‘some’ and ‘only some’, reading times for the quantifier cannot be compared between the two conditions because of the potential confounding effects of different wording (Breheny, Katsos, & Williams, 2006; Bergen & Grodner, 2012; Politzer-Ahles & Fiorentino, 2013; Hartshorne & Snedeker, submitted). Thus, it still remains to be determined whether semantically-encoded upper-bounded meanings are more costly to process when no verification is necessary.

In summary, results from previous studies are mixed. The critical comparison between ‘some’ and ‘only some’ is problematic because the cue for upper-bounding comes earlier for ‘only some’. The goal of Experiment 1 is to compare the cost of evaluating upper-bounded and lower-bounded interpretations of ‘some’, removing this potential confound.

3.5.2 Experiment 1

The goal of Experiment 1 was to determine the cost of upper-bounding independent of the need for pragmatic inference by encoding the intended interpretation of the quantifier in the literal meaning of the sentence. I enforced upper-bounded interpretations using the focus operator ‘only’, and lower-bounded interpretations using ‘at least’. I adopted a truth-value judgment paradigm similar to that of Marty, Chemla, & Spector (2013), using simple visual contexts rather than world knowledge for verification.

In addition to comparing upper-bounded and lower-bounded interpretations, I also tested two different quantifier scales: <‘all’, ‘some’> and numerals. Some previous studies have found that processing upper-bounded (exact) interpretations of numerals is

less costly than for ‘some’ and other quantifiers (Huang & Snedeker, 2009; 2011; Degen & Tanenhaus, 2011; Marty, Chemla, & Spector, 2013), and furthermore that children compute upper-bounded interpretations of numerals more readily (Papafragou & Musolino, 2003; Huang & Snedeker, 2009), some argue that the exact interpretation of numerals is encoded in their semantics (Geurts, 2006; Breheny, 2008), while others propose that the difference is due to a high level of familiarity with the number scale compared to other scales (Barner, Brooks, & Bale, 2010). There are also concerns about how the two types of quantifiers are tested, given that numeral expressions representing subitizable quantities may be evaluated using different visual mechanisms for estimating quantity (Feigenson, Dehaene, & Spelke, 2004). The manipulation in my study will not substantially contribute to this debate, since I am not testing pragmatically-derived upper-bounded interpretations. The main purpose for testing numerals was methodological: including sentences with numerals made the test sentences less predictable. The results demonstrate that upper-bounded interpretations are no less costly to evaluate for numerals than for ‘some’ when (a) the quantities are not subitizable and (b) the upper or lower bound is part of the semantic meaning, and need not be inferred. This result should come as a surprise to no one.

Again, my primary goal in this experiment was to investigate processing costs for evaluating semantically-encoded upper-bounded and lower-bounded interpretations of ‘some’.

3.5.2.1 Methods

Participants

20 students at the University of Maryland (18-22 years, 14 female) participated in return for course credit.

Design

Participants judged sentences with scalar quantifiers with respect to an image. The scalar quantifier was either 'some' or a numeral. All quantifiers were either explicitly lower-bounded (preceded by 'at least') or explicitly upper-bounded (preceded by 'only'), as shown in (52)-(53) and (54)-(55), respectively.

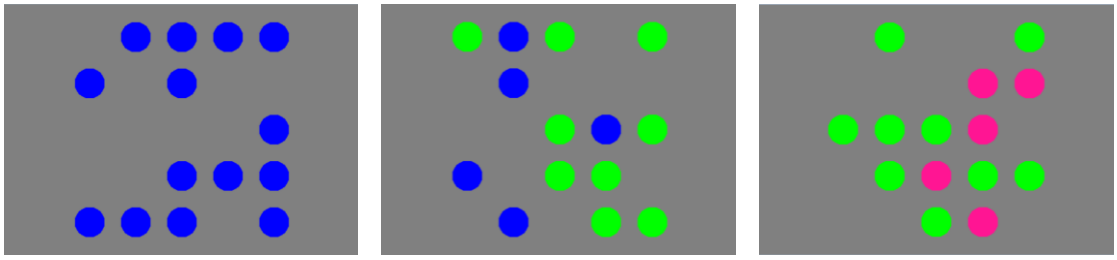
(52) At least five dots are blue.

(53) At least some dots are blue.

(54) Only five dots are blue.

(55) Only some dots are blue.

The sentences were accompanied by an image of an array of dots. I manipulated the number of dots in the array matching the target color mentioned in the sentence (e.g. blue): *all*, *none*, or a *subset* (see Figure 3-1). The *none* scenes were constructed by replacing the target color with a second non-target ("distractor") color in a *subset* scene.



(a) (b) (c)

Figure 3-1 Experiment 1: Scene types.
The target color is blue. **(a)** *All* dots are the target color. **(b)** A *subset* of dots are the target color. **(c)** *None* of the dots are the target color.

Quantifier Type	Explicit Bound	Sample sentence	Scene Type/Truth Value		
			All	*Subset*	*None*
some	*at least*	At least some dots are blue.	T	T	F
	only	Only some dots are blue.	F	T	F
numeral	*at least*	At least five dots are blue.	T	T	F
	only	Only five dots are blue.	F	T	F

Table 3-1 Experiment 1: Conditions.

If verifying upper-bounded interpretations is costly even when no pragmatic inference is necessary, accuracy should be lower and response times higher in the *only* condition compared to the *at least* condition, even for *subset* scenes where both are true. If the processing advantage for numerals that has been observed in previous studies is related to a pragmatic process or to subitizable numbers, the higher cost for *only* should hold for both Quantifier Types.

Materials

The two sentence manipulations (QUANTIFIER TYPE and EXPLICIT BOUND) and the scene manipulation (SCENE TYPE) were combined factorially for 12 total conditions, shown in Table 3-1. There were 20 trials for each condition, for 240 trials total. I created 20 different scene schemas by systematically varying non-experimental properties of the scene, including the target color, the non-target ("contrast") color, the total number of dots, and the ratio of target dots to total dots (see Table 3-2). The color pairings were chosen for maximum discriminability. The colored dots appeared on a dark gray background. The specific positions of the dots on an invisible 5 x 5 grid was determined randomly on each trial.

Target Color	Distract Color	Contrast Color	Total Dots	Subset Dots	Scene Types: All	Subset	None
blue	pink	green	14	5	14 blue	5 blue, 9 green	5 pink, 9 green
		brown	14	9	14 blue	9 blue, 5 brown	9 pink, 5 brown
		yellow	16	5	16 blue	5 blue, 11 yellow	5 pink, 11 yellow
		orange	16	11	16 blue	11 blue, 5 orange	11 pink, 5 orange
red	white	blue	18	6	18 blue	6 red, 12 blue	6 white, 12 blue
		purple	18	12	18 blue	12 red, 6 purple	12 white, 6 purple
		yellow	20	7	20 blue	7 red, 13 yellow	7 white, 13 yellow
		black	20	13	20 blue	13 red, 7 black	13 white, 7 black
pink	green	white	22	7	22 blue	7 pink, 15 white	7 green, 15 white
		black	22	15	22 blue	15 pink, 7 black	15 green, 7 black
		yellow	14	5	14 blue	5 pink, 9 yellow	5 green, 9 yellow
		black	14	9	14 blue	9 pink, 5 black	9 green, 5 black
green	blue	red	16	5	16 blue	5 green, 11 red	5 blue, 11 red
		purple	16	11	16 blue	11 green, 5 purple	11 blue, 5 purple
		brown	18	6	18 blue	6 green, 12 brown	6 blue, 12 brown
		orange	18	12	18 blue	12 green, 6 orange	12 blue, 6 orange
white	red	pink	20	7	20 blue	7 white, 13 pink	7 red, 13 pink
		purple	20	13	20 blue	13 white, 7 purple	13 red, 7 purple
		orange	22	7	22 blue	7 white, 15 orange	7 red, 15 orange
		brown	22	15	22 blue	15 white, 7 brown	15 red, 7 brown

Table 3-2 Experiment 1: Trial types.

Procedure

Participants sat in front of a CRT monitor and wore headphones. The experimental materials were presented using the Python with the package PsychoPy2. The experiment began with instructions and five practice trials. The first two practice trials were presented more slowly to allow the participant to adjust to the format of the experiment. Feedback was provided after all practice trials. After the practice trials, the 240 experimental trials were presented in random order. There were two automatic breaks, after 80 trials and 160 trials.

Each trial began with a black square presented in the middle of the screen. After a 500ms pause, the test sentence played over the headphones. The sentences were 1700-1900ms long. The sound files were trimmed at the beginning and end with the aid of visual inspection of the waveform in Praat (Boersma, 2001).

At the offset of the sentence, the array of dots appeared on the screen for about 250ms. The actual duration of the presentation was dependent on the monitor refresh rate, and thus varied slightly from trial to trial (250-297ms, mean 258ms). Participants' response times were recorded from the beginning of the visual presentation. Since participants heard the sentence before seeing the display, their verification procedures should not be much affected by a few milliseconds of extra display time. Thus, time-locking their responses to the beginning of the visual presentation should prevent slight variation in the duration of the presentation having an effect on response times. In any case, since the significant differences in response times we observed were as large as 150-300ms, and the duration of the presentation fell within a 16ms range (250-266ms) for 99% of trials, the presentation differences were unlikely to have affected our results.

After the dots disappeared, a response cue appeared on the screen, prompting participants to press 'F' for "true" or 'J' for "false". Participants were instructed to respond as quickly as possible. The instructions also explained that since the visual presentation was so brief, they would not always be fully confident in their answers. The next trial began immediately after the participant responded.

Data analysis

Accuracy rates were analyzed using logistic mixed effects models. Accuracy for each SCENE TYPE was modeled separately, with fixed effects for QUANTIFER TYPE and EXPLICIT BOUND, and the maximal random effects structure (a random by-subject intercept and random by-subject slopes for QUANTIFIER TYPE, EXPLICIT BOUND, and their interaction). All fixed effects were coded orthogonally.

Response times were analyzed using a linear mixed effects model with the same structure as the model used for accuracy rates. Given the large number of observations (400 per condition), the significance of fixed effects can be estimated using the t-statistic (Baayen, 2008). Fixed effects are considered significant at the 5% level if the absolute value of their t-statistic is greater than 2, and marginally significant (at the 10% level) if it is greater than 1.67.

3.5.2.2 Results

Accuracy

Accuracy rates for each condition are shown in Table 3-3 and Figure 3-2.

In the *all* scenes, there were significant main effects for QUANTIFIER TYPE and EXPLICIT BOUND: accuracy was higher for *some* than for *numerals* ($p = 0.0020$), and for *at least* than for *only* ($p = 0.0016$). The interaction was not significant. Since the interaction looks substantial in the overall means, I fitted a model without a random by-

Quantifier Type	**Explicit Bound**	**Scene Type** *All*	*Subset*	*None*
Some	*At least*	0.855 (T)	0.968 (T)	0.990 (F)
	Only	0.803 (F)	0.943 (T)	0.993 (F)
Numeral	*At least*	0.895 (T)	0.858 (T)	0.990 (F)
	Only	0.710 (F)	0.888 (T)	0.988 (F)

Table 3-3 **Experiment 1: Accuracy (grand means).**
Target responses in parentheses.

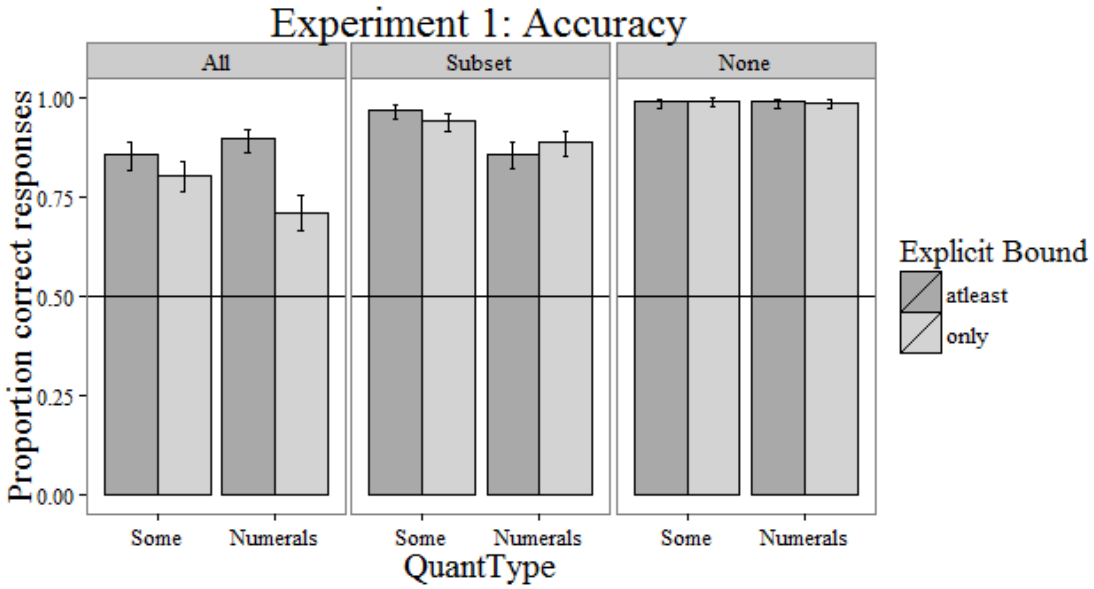


Figure 3-2 **Experiment 1: Accuracy**
By-subject means. Error bars represent 95% confidence intervals based on the binomial distribution.

subject slope for the interaction term, and compared it to the original model with a likelihood ration test. The model with the random by-subject interaction was a significantly better fit to the data ($p << 0.0001$), suggesting that there was significant variation across subjects in the pattern of accuracy.

In the *subset* scenes, there was only a significant main effect of QUANTIFIER TYPE: accuracy was higher for *some* than for *numerals* ($p << 0.0001$).

In the *none* scenes, accuracy was near ceiling, with no significant differences between conditions.

Response time

Mean response times for each condition are shown in Table 3-4 and Figure 3-3.

In the *all* scenes, there was a highly significant main effect of EXPLICIT BOUND: response times for *only* were significantly longer than for *at least* ($t = 3.84$). There was also a significant interaction between EXPLICIT BOUND and QUANTIFIER TYPE ($t = -2.17$): the effect of EXPLICIT BOUND was larger for *numerals* than for *some*.

In the *subset* scenes, there was a highly significant main effect of QUANTIFIER TYPE: response times for *numerals* were significantly longer than for *some* ($t = -4.98$). There was also a significant main effect of EXPLICIT BOUND: response times for *only* were longer than for *at least* ($t = 2.78$).

In the *none* scenes, there was again a highly significant main effect of QUANTIFIER TYPE: response times for *numerals* were significantly longer than for *some* ($t = -4.35$).

Quantifier Type	Explicit Bound	Scene Type		
		All	*Subset*	*None*
Some	*At least*	836 (42)	701 (36)	624 (32)
	Only	901 (49)	771 (33)	632 (36)
Numeral	*At least*	831 (32)	993 (82)	685 (36)
	Only	980 (49)	1029 (79)	687 (29)

Table 3-4 **Experiment 1: Response times.**
By-subject means. Standard error in parentheses.

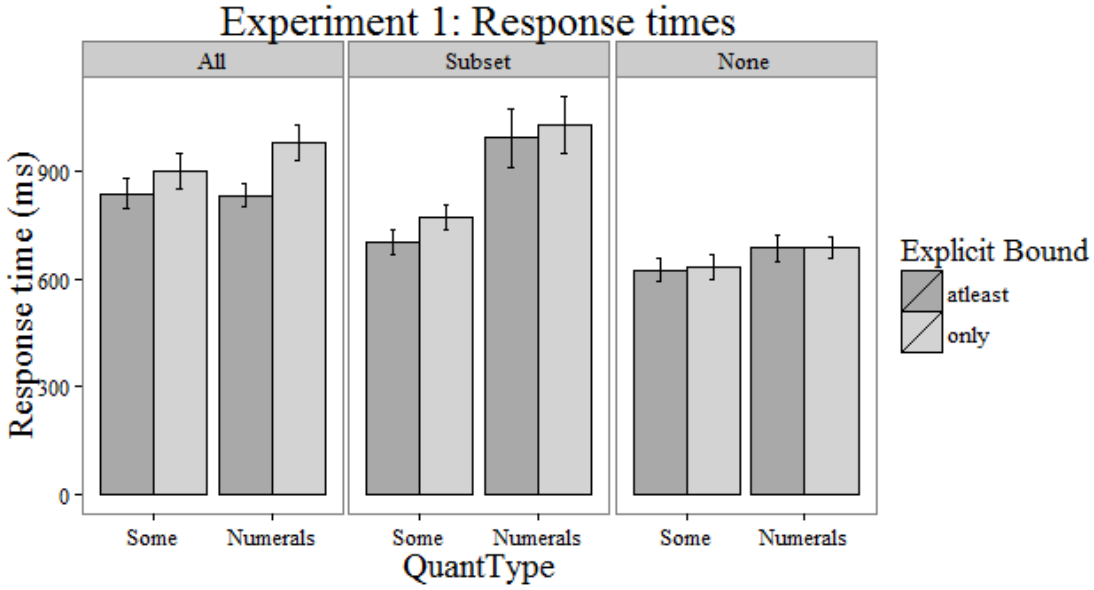


Figure 3-3 **Experiment 1: Response times.**
By-subject means. Error bars represent standard error.

3.5.2.3 Discussion

The goal of Experiment 1 was to test the generalization that evaluating upper-bounded interpretations of quantifiers is costly, even when the upper bound is required by the literal meaning of the sentence.

The *all* scenes are comparable to the critical conditions of the judgment studies cited above (Bott & Noveck, 2004; Bott, Bailey, & Grodner, 2012; Marty & Chemla, 2013). In these scenes, the sentence is true with the lower bound (‘at least’) and false with

the upper bound (‘only’). For both ‘some’ and numerals, accuracy was lower and response times were longer for the ‘only’ sentences, suggesting that the upper-bounded interpretation was more costly to evaluate for this scene type. Thus, upper-bounding—at least in judgment studies—is costly even when it is required by the literal meaning of the sentence. This cost may be due to differences in the verification procedure for lower-bounded and upper-bounded interpretations. To verify that ‘at least some’ dots are blue, it is only necessary to find more than one blue dot in the scene. To verify that ‘only some’ dots are blue, it is necessary to find more than one blue dot in the scene, and determine that there is a dot of another color in the scene. The extra step for upper-bounding plausibly takes extra time.

The cost for rejecting false sentences with upper-bounded quantifiers was greater for numerals than for ‘some’ (at least in response times, and also in accuracy for some subjects). The greater difficulty of rejecting the false sentences with exact numerals must be attributable to the verification mechanisms used in each case. To verify that ‘only seven’ dots are blue, it is necessary to estimate the number of blue dots in the scene (using the Approximate Number System) and determine whether it matches the number in the sentence within some acceptable margin of error. The estimation of the number of blue dots plausibly takes longer (or gives rise to greater uncertainty) than merely determining that multiple blue dots and dots of another color are present in the scene.

The difference in response times to upper- and lower-bounded quantifiers in the *all* scenes could be attributed to the different target responses: upper-bounded quantifiers elicited a “false” response, while lower-bounded quantifiers elicited “true”. For this reason, the *subset* scenes are also of interest, since the sentence is true regardless of

which explicit bound is set. Response times for the *subset* scenes were also slower for the upper-bounded 'only' sentences, demonstrating that evaluating upper-bounded interpretations is more costly regardless of whether it results in a rejection of the sentence.

In summary, the results of Experiment 1 support the hypothesis that the cost for upper-bounded interpretations of scalar expressions that has been observed in some judgment studies is attributable at least in part to the upper-bounded meaning itself. I suggest that cost likely arises from multi-step verification procedures for upper-bounded interpretations; this hypothesis could be tested using visual-world eye-tracking to observe participants' verification procedures. Since verification procedures may not be deployed in studies that do not involve truth-value judgments, there must be some other source(s) of cost for computing upper-bounded interpretations. I turn to these in the following sections.

3.6 The cost of computing implicated meanings

Implicature-enriched interpretations may be more costly than literal interpretations because of the method used for accessing implicated meanings. Given that implicature-enriched interpretations are often more costly, it seems unlikely that the implicated meaning is accessed directly. On the other hand, if accessing the implicated meaning requires a multi-step process (whether it be domain specific or completely general), we would expect literal interpretations to become available earlier than implicature-enriched interpretations. To find evidence for this temporal ordering of potential interpretations, we need methods that allow us to observe the interpretation changing over time, rather than just the final product.

Bott, Bailey & Grodner's (2012) SAT task allows us to observe the output of verification processes after specific periods of time (27, 100, 200, 300, 400, 600, 800, and 2500 ms). If upper-bounded interpretations of 'some' are obligatorily preceded by logical, lower-bounded interpretations, the rate of incorrect responses should actually increase (i.e., exceed chance) before it begins to fall towards an asymptotic rate of near zero. In their Experiment 1, Bott and colleagues analyzed pseudo-*d'* scores—incorrect response rates in critical conditions compared to control conditions—as a function of response time. Pseudo-*d'* should be high at times when participants are choosing the incorrect response more often in the critical conditions than the control conditions, and close to zero when they are making mistakes at the same rate in critical and control conditions. They found that pseudo-*d'* was non-monotonic for participants who were instructed to use an upper-bounded interpretation: it increased at first before falling towards zero. By contrast, pseudo-*d'* was monotonic for participants who were instructed to use a lower-bounded interpretation: it moved steadily towards zero. This pattern is consistent with the hypothesis that pragmatically-derived upper-bounded interpretations are preceded by logical lower-bounded interpretations. Unfortunately, pseudo-*d'* was not reported for Experiment 2, which compared 'some' and 'only some', so it is not possible to determine whether the pattern of responses reflected an implicature-specific process. If pseudo-*d'* does not differ for 'some' and 'only some', it may be that access to the lower-bounded interpretation of 'some' is simply a necessary part of deriving the upper-bounded interpretation, regardless of whether an implicature is involved.

Tomlinson, Bailey & Bott (2013) used mouse-tracking to observe how judgments for underinformative sentences like (48) unfold over time. Participants used a mouse to

click on “true” or “false” to indicate their judgments. The mouse pointer started in the center of the bottom edge of the screen, and the response options were located in each of the top corners. Deviations from a direct path from the starting point to the target are the critical evidence for processing difficulty or consideration of an alternative response. In Experiment 1, participants were divided into two groups and trained to provide either logical (lower-bounded) or pragmatic (upper-bounded) responses, as in the studies described above. Mouse trajectories for each group for judgments of underinformative sentences like (48) were compared to true/felicitous (56) and false (57) sentences.

(56) Some mammals are elephants.

(57) Some elephants are insects.

For participants in the logical (lower-bounded) group, mouse trajectories for answering “true” to underinformative sentences like (48) were no different from those for answering “true” to fully felicitous sentences like (56). For participants in the pragmatic (upper-bounded) group, mouse trajectories for answering “false” to underinformative sentences deviated substantially from those for answering “false” to sentences like (57). The trajectory veered significantly toward the alternative (logical/“true”) response. This pattern was replicated in Experiment 2, where participants were not instructed which interpretation to use.

These results suggest that an interpretation based on the logical (lower-bounded) interpretation of ‘some’ is available earlier than an upper-bounded interpretation. However, this was true regardless of whether the upper-bounded interpretation arose due to a pragmatic inference (Experiment 2) or specific instruction about the intended interpretation (Experiment 1). Thus, the computation of the lower bound seems to be a

necessary part of computing an upper-bounded interpretation, not as evidence for a specific type of pragmatic inference.

A final type of relevant evidence comes from studies of reference resolution based on scalar implicatures measured in visual world eye-tracking studies. Huang & Snedeker (2009) were the first to use this method to observe the interpretation of a scalar expression unfold in real time. In their study, participants heard instructions like (58) with underinformative 'some'. Their eye movements were recorded for the duration of the instruction as they looked at a scene with four quadrants containing girls and boys with socks and soccer balls.

(58) Point to the girl that has some of the socks.

The target quadrant contained a girl who has a subset (two out of four) of the socks. The critical distractor quadrant contained a girl who has all (three) of the soccer balls in the scene. The instruction would be ambiguous between *socks* and *soccer balls* until the final *–s* of *socks*. Thus, participants' preference to look at the target quadrant before disambiguation could be taken as evidence that they preferred the upper-bounded interpretation of 'some' at that point. However, Huang and Snedeker found that participants did not converge on the target quadrant until after the disambiguating information. By contrast, participants were very fast to find the target quadrant with instructions like (59), even though the literal, lower-bounded interpretation of 'two' would be consistent with the three soccer balls in the distractor quadrant.

(59) Point to the girl that has two of the socks.

The authors argue that participants have access to the literal, lower-bounded interpretation of ‘some’ before the upper-bounded interpretation, since there is a period of time when they consider both of the girls as potential targets. To confirm that participants looked at both the target and distractor because they considered both consistent with the instruction, and not just because they had not yet computed any meaning for ‘some’, Huang and Snedeker’s Experiment 3 compared the critical condition to a new control condition. In the “1-referent” condition, socks are the only set under discussion (there are no soccer balls). The distractor quadrant contains a girl with no socks, which is inconsistent with both the logical and pragmatic interpretations of ‘some’. In that condition, participants looked toward the target quadrant immediately upon hearing ‘some’, demonstrating that they had computed the logical meaning of the expression (*not none*).

In a subsequent study, Huang & Snedeker (2011) extended the ambiguous period in the instruction as in (60) to determine exactly how long it would take for participants to show a preference for the upper-bounded interpretation of ‘some’. They found that participants did not reliably look to the target quadrant until 1100 ms after the quantifier. This delay sharply contrasted with the near-immediate preference for an upper-bounded interpretation of ‘two’.

(60) Point to the girl that has some of the ice cream {sandwiches/cones}.

In summary, results from a variety of methods suggest that the literal interpretation of a scalar expression becomes available to guide responses earlier than the implicature-enriched interpretation. This generalization would be consistent with an algorithm in which literal meanings are obligatorily computed first to feed into a

pragmatic interpretation process. It would also be consistent with an algorithm where multiple potential speaker meanings (including the literal meaning) are computed in parallel, but literal meanings require fewer steps and are finished faster.

3.7 The cost of using contextual information

So far I have largely ignored the role of context in computing upper-bounded interpretations, although the experiments I have reviewed vary widely in their use of context. As I discussed in section 3.4.1.2, context could play its role before, during, or after the interpretation of the critical expression. It seems likely that in fact context plays a role in all of these ways. Sometimes contextual information may be available early, and sometimes not. Some conversations are less predictable than others. My goal in this section is not to attempt a general theory of how context affects real-time interpretation, but rather to investigate more thoroughly the role context might have played in the processing costs observed in the literature. I argue that the perplexing pattern of cost in some studies and lack of cost in other studies can be better understood by taking contextual differences into account.

I will consider two kinds of evidence. First, some studies suggest that if the properties of the discourse of the experiment are modified such that implicature-enriched interpretations are more predictable, the cost of implicatures is reduced. Second, the variation in processing costs observed across different reading studies suggests that the cost of the implicature is affected by the accessibility of relevant alternatives in the context.

3.7.1 Making upper-bounded interpretations more predictable

Grodner, Klein, Carbary & Tanenhaus (2010) used a paradigm similar to that of Huang & Snedeker (2009; 2011) to look at how quickly participants could use a scalar implicature to constrain the reference of a quantified expression in instructions like (61).

(61) Click on the girl who has summa the balls.

While Huang & Snedeker's experiments included instructions with numerals, 62 out of the 72 instructions in Grodner and colleagues' experiment contained *some*, *all*, or *none*. (The other ten used *the*, as in *the girl who has the balls*). Thus, nearly all of the sentences in the experiment were relevant alternatives for each other: the only relevant scale was <*all*, *some*>. They found that looks began to converge on the target well before disambiguation, in contrast to Huang & Snedeker's results. There was no delay in using an implicature-enriched meaning for *some of the balls* compared to a literal meaning for *all of the balls* or *none of the balls*. The authors conclude that the processing costs observed by Huang & Snedeker most likely reflect the relative difficulty of accessing relevant alternatives in an experiment which made use of a wider variety of descriptions for the target sets. By rehearsing the relevant alternatives in nearly every trial, Grodner and colleagues' experiment made it easy for participants to access 'all' for the purpose of upper-bounding 'some'.

Degen & Tanenhaus (2013, under review) arrive at a similar conclusion using a somewhat different paradigm. Participants saw an image of a gumball machine with orange and blue gumballs in an upper chamber. Then some gumballs dropped to a lower chamber, creating a contrast between a partitioned set of gumballs of one color and an unpartitioned set of the other color. Participants evaluated sentences like (62) against

these scenes. In Experiment 1, participants rated the naturalness of the sentences as descriptions for different set types. The description in (62) was rated as less natural for partitioned sets of certain sizes when the experiment included descriptions with numerals like (63). Response times and eye-movements for truth-value judgments reflected greater difficulty evaluating the sentences with 'some' when they were competing with number alternatives in the experiment.

(62) You got some of the blue gumballs.

(63) You got two of the blue gumballs.

Degen and Tanenhaus' results suggest that the discrepancy between Grodner and colleagues' (2010) results and Huang and Snedeker's (2009; 2011) may well have been due to the different range of sentence types used. Since Huang and Snedeker included number terms which would compete with 'some' as descriptions for the target images, evaluating sentences with 'some' may have been more difficult.

It is not clear which experiment is more representative of the difficulty experienced by listeners in real conversation. On the one hand, it is certainly not the case that people are constantly and exclusively talking about 'all', 'some', or 'none' of a set, so that listeners are ready to efficiently compute comparisons between them. On the other hand, the rich information available in real conversations may make the relevant alternatives salient in other ways. Huang & Snedeker made an effort to embed their critical sentences in natural contexts, but those contexts did not include any mention of the full set of socks or soccer balls. The full set that was relevant for interpreting the critical sentence was the one present in the visual scene, not one from the discourse context, which may have affected participants' ability to use it.

Degen and Tanenhaus' findings speak to a broader point as well: listeners' expectations about the utterances that speakers will use to express their meanings are informed not only by immediately relevant alternatives, but also the distribution of the utterances they have encountered in other contexts. This suggests an early, proactive role for at least some kind of contexts.

3.7.2 Accessing relevant alternatives: reading studies

In reading studies, sometimes increased processing cost is observed at the point of the triggering scalar expression, and sometimes not. By comparing the contexts that lead to increased cost to those that do not, we can begin to infer the source of the cost. I will review results of each type in turn.

3.7.2.1 Exclusive 'or' and ad-hoc scales

Breheny, Katsos & Williams (2006) report two experiments on the processing costs of scalar implicature. In their Experiment 1, they examined the cost of computing upper-bounded "exclusive" readings of *or* compared to lower-bounded "inclusive" readings. The different readings were induced through brief contexts, as in (64)-(65) (translated from Greek). They found that reading times on the critical region containing *or* (underlined in the examples) were longer in the Upper-bounding condition than in the Lower-bounding condition.

(64) Upper-bounding context: John was taking a university course and working at the same time. For the exams he had to study from short and comprehensive sources. Depending on the course, he decided to read the class notes or the summary.

(65) Lower-bounding context: John heard that the textbook for Geophysics was very advanced. Nobody understood it properly. He heard that if he wanted to pass the course he should read the class notes or the summary.

Katsos, Breheny & Williams (2005) also looked at exclusive and inclusive readings of *or* in a self-paced reading study in British English (their Experiment 2). Their upper-bounding and lower-bounding contexts were quite similar, as shown in (65)-(66).

(66) *Upper-bounding context*: The manager asked: Who has the report on last year's profits? Her secretary replied: Jones or Barnes from the department of Finance has. Would you like to see the report?

(67) *Lower-bounding context*: The manager asked: Who has a report on last year's profits to show me? Her secretary replied: Jones or Barnes from the department of Finance has. Would you like to see the report?

They found that reading times on the critical segment (e.g. *Jones or Barnes*) were longer in the upper-bounding condition than in the lower-bounding condition.

Katsos, Breheny & Williams (2005) examined the cost of scalar implicatures with "ad-hoc" scales in their Experiment 3. The upper-bounding context included a question that explicitly mentioned the stronger element on the scale (68), while the lower-bounding context did not mention any member of the scale (69).

(68) George went to pick up Mary from the station. He was covered in paint. Mary asked him: Were you painting the house? George replied: I was painting the roof with an insulating paint.

(69) George went to pick up Mary from the station. He was covered in paint. Mary asked him: What were you painting? George replied: I was painting the roof with an insulating paint.

Since there have not been any studies reporting a lack of cost for upper-bounding implicatures for 'or' or ad-hoc scales, it is difficult to say what the source of the cost was in these studies. It's important to note that none of these studies includes a measure of the participants' final interpretation (in contrast to the studies I review in the next section). Thus, we cannot be sure that an upper-bounded interpretation was successfully computed. I will return to this issue in the discussion of Experiment 2.

3.7.2.2 Quantifier scales

Breheny, Katsos & Williams (2006) looked at the cost of computing upper-bounded and lower-bounded readings of *some* in their Experiment 3. As mentioned above in section 3.2, reading times at *the rest* indicated that an upper-bounded interpretation of *some* had been computed in the upper-bounding context (70), but not the lower-bounding context (71). Furthermore, reading times for the region containing *some* were longer in the upper-bounding condition than the lower-bounding condition, suggesting a cost for the implicature.

(70) Upper-bounding context: Mary asked John whether he intended to host all his relatives in his tiny apartment. John replied that he intended to host some of his relatives. The rest would stay in a nearby hotel.

(71) Lower-bounding context: Mary was surprised to see John cleaning his apartment and she asked the reason why. John told her that he intended to host some of his relatives. The rest would stay in a nearby hotel.

Bergen & Grodner (2012) also looked at the cost of implicature-enriched readings of *some*. As mentioned above in section 3.2, they also found a difference in reading times at *the rest* indicating that an upper-bounded interpretation of *some* had been computed in the "full knowledge" context, but not in the "partial knowledge" context. Like Breheny and colleagues, they found that reading times for the region containing the scalar quantifier (*some of*) and the spillover region (*the real estate*) were longer in the full knowledge context (72) than the partial knowledge context (73), suggesting a cost for the implicature.

(72) Full knowledge: At my client's request, I meticulously compiled the investment report. Some of the real estate investments lost money. The rest were successful despite the recent economic downturn.

(73) Partial knowledge: At my client's request, I skimmed the investment report. Some of the real estate investments lost money. The rest were successful despite the recent economic downturn.

I turn now to the studies which did not find any evidence of cost for implicature-enriched interpretations. Both studies employed the design used by previous studies: they manipulated whether an upper-bounded reading of *some* was licensed, and measured reading times for *some* as well as for a region containing a reference to the complement set (*the rest*).

Hartshorne & Snedeker (submitted) manipulated the intended reading of *some* by embedding it in an upward-entailing (UE) or downward-entailing (DE) context, as in (74)-(75). They also manipulated whether *some* was preceded by *only*, which would semantically force an upper-bounded interpretation.

(74) *UE*: Addison ate (only) some of the cookies before breakfast this morning, and the rest are on the counter.

(75) *DE*: If Addison ate (only) some of the cookies before breakfast this morning, then the rest are on the counter.

In Experiment 1, they found that when *only* was present, reading times at *the rest* were the same in each condition, since the interpretation of *some* is upper-bounded in both cases. When *only* was absent, reading times for the spillover region following *the rest* were shorter in the UE condition, suggesting that an upper-bounded interpretation of *some* had initially been computed in that condition, but not the DE condition. However, there were no differences in reading times at *some* or its spillover region between the two conditions, suggesting that there was no cost for computing the upper-bounded interpretation at that point, in contrast to the findings of Breheny, Katsos, & Williams (2006) and Bergen & Grodner (2012).

In Experiment 2, they reduced the distance between *some* and *the rest* as in (76)-(77). This reduced the time between the scalar trigger and the target from about 2.5s to about 900ms. With these modified materials, they found no significant differences in reading times at or following *the rest* across the different conditions, suggesting that there was not sufficient time to compute distinct interpretations of *some* in each context.

(76) *UE*: Addison ate (only) some of the cookies, and the rest are on the counter.

(77) *DE*: If Addison ate (only) some of the cookies, then the rest are on the counter.

In multiple replications of Experiments 1 and 2, Hartshorne & Snedeker consistently failed to find any cost for upper-bounded interpretations of *some*. The authors argue that scalar implicatures must take more than 900ms to get started (Experiment 2), but that they do not give rise to additional processing cost such that readers slow down (Experiment 1). They propose that the costs observed in Breheny et al. (2006) may have been due to confounding differences between the contexts.

Before discussing those differences, I will review the other study that failed to find any cost for scalar implicature. Politzer-Ahles & Fiorentino (2013) designed a self-paced reading study very similar to Breheny et al.'s Experiment 3, but with more tightly controlled contexts, as in (78)-(79). Although they replicated the longer reading times at *the rest* in the lower bound condition, there were no differences in reading times at the scalar quantifier.

(78) *Upper bound*: Mary was preparing to throw a party for John's relatives. She asked John whether all of them were staying in his apartment. John said that some of them were. He added that the rest would be staying in a hotel.

(79) *Lower bound*: Mary was preparing to throw a party for John's relatives. She asked John whether any of them were staying in his apartment. John said that some of them were. He added that the rest would be staying in a hotel.

One important point of variation across these studies is whether the set to be quantified with *some* in the critical region is mentioned earlier in the context. In Breheny

et al.'s (2006) Experiment 3, the upper bound condition (which licensed an implicature) always mentioned the set (e.g. *all of his relatives*) while the lower bound condition did not. In Politzer-Ahles & Fiorentino's (2013) otherwise very similar materials, the set was mentioned in both conditions, and in the critical region it was referred to with *them* instead of a repetition of the noun. Since Politzer-Ahles and Fiorentino did not observe any cost for the implicature, the cost reported by Breheny et al. in the upper bound condition may have been an instance of the "Repeated Name Penalty" (Gordon, Grosz, & Gilliom, 1993; Gordon & Chan, 1995). Consider again the upper bound condition of Breheny et al's Experiment 3, repeated as (80). Although it is certainly possible to avoid the infelicity of the repeated 'his relatives' by using appropriate prosody (e.g. "SOME of his relatives"), this is only possible if you are expecting the repetition. It is therefore unlikely that participants would find the repetition felicitous upon first encountering it.

(80) Mary asked John whether he intended to host <u>all his relatives</u> in his tiny apartment. John replied that he intended to host <u>some of his relatives</u>. The rest…

While the materials in Bergen & Grodner's (2012) experiment varied, in 18 of the 24 items the set was not mentioned in either condition: it was merely implied. A few representative examples are given in (81)-(83).

(81) This morning, I took attendance at an important meeting with the manager. <u>Some of the company's accountants</u> were there.

(82) In the school parking lot, I carefully inspected an old bus. <u>Some of its tires</u> were flat.

(83) At a friend's suggestion, I completely worked through an entire math textbook. Some of its problems were difficult.

Despite the lack of repeated names, they still found a cost for the upper-bounding implicature. This suggests that the cost arose due to the need in the upper-bounding condition to infer how the critical set should be related to the context. In Hartshorne & Snedeker's (submitted) experiments, there was no context, so there was no need to relate the set to anything.

We can identify two potential sources for the slow-downs observed at the scalar quantifier in some studies. (1) A Repeated Name Penalty may be observed in the implicature condition if it includes an infelicitous repetition of the quantified NP. (2) An upper-bounded interpretation of 'some' requires that the full set be relevant to the discourse. If the full set has not yet been explicitly mentioned, it must be inferred, and this inference could be costly.

In Experiment 2, I designed the materials so that there was no infelicitous repetition. If the cost observed in Breheny et al.'s (2006) study was due to the Repeated Name Penalty, it should be not be replicated here. I also compared lower-bounded interpretations in contexts that explicitly mention the set to those in contexts with no previous mention of the quantified set, in order to investigate the cost of inferring the relevant set. Finally, I tested ad-hoc scales that were closely matched to the 'some'/'all' scale, as a first step in determining how generalizable the results from 'some' will be to other kinds of implicature.

3.7.3 Experiment 2

3.7.3.1 Methods

Participants

24 native English speakers from the University of Maryland community (18-33 years, mean 21.1 years, 18 females) participated for payment or course credit. 10 additional people participated but were not able to finish the experiment due to difficulty calibrating the eye-tracker and various tracking problems. Data from 3 others were excluded due to poor tracking and excessive artifacts. One participant withdrew from the study due to discomfort in the eye-tracking apparatus.

Design/Materials

The experimental items were three-sentence passages like (84). The first two sentences served to establish a context, and the third sentence contained the critical scalar expression.

(84) Julia had decided to tour all of the historic sites during her visit to Prague.
She spent a week walking around with her guidebook.
On her travel blog, she described how she had toured some of the historic sites in her whirlwind visit, but the rest were too difficult to find.

The first context sentence determined whether an upper-bounding scalar implicature would be licensed for the scalar expression (e.g. *some of the historic sites*) in the third sentence. In the Upper Bound condition, the upper bound of the scale (e.g. *all of the historic sites*) was mentioned (85). In the Lower Bound condition, the lower bound

(e.g. *some of the historic sites*) was mentioned (86). In the Neutral condition, the scale was not mentioned at all (87).

(85) Julia had decided to tour all of the historic sites during her visit to Prague.

(86) Julia had decided to tour some of the historic sites during her visit to Prague.

(87) Julia had decided to learn about local history during her visit to Prague.

I also manipulated the Scale Type, which was either *some/all* or an ad-hoc scale, as in (88). The ad-hoc scales were constructed with adjectives that could be construed as scalar in context. The presence of a scale was reinforced by explicitly marking the upper bound (with *even*) or lower bound (with *at least*) when it was introduced in the context sentence. The full set of conditions for a sample item is given in Table 3-1.

(88) Julia had decided to {tour even the obscure historic sites/tour at least the famous historic sites/learn about local history} during her visit to Prague.
She spent a week walking around with her guidebook.

	***Some/all* scale**	**Ad-hoc scale**
Sentence 1	Julia had decided to…	
Upper bound	tour all of the historic sites	tour even the obscure historic sites
Lower bound	tour some of the historic sites	tour at least the famous historic sites
Neutral	learn about local history	learn about local history
		…during her visit to Prague.
Sentence 2	She spent a week walking around with her guidebook.	
Sentence 3	On her travel blog, she described how she had…	
	toured some of the historic sites	toured the famous historic sites
		…in her whirlwind visit, but the rest were too difficult to find.

Table 3-5 Experiment 2: Sample item.

On her travel blog, she described how she had toured the famous historic sites in her whirlwind visit, but the rest were too difficult to find.

Procedure

Participants were tested individually in a quiet room in one 60-minute session. Eye movements were recorded using an EyeLink 1000 eye-tracker (SR Research, Toronto, Ontario, Canada) interfaced with a PC computer. Participants were seated with their chin and forehead stabilized by the eye-tracker apparatus, 32'' from an LCD monitor which displayed the stimuli. At this distance, 4.5 characters were displayed per degree of visual arc. The eye-tracker has an angular resolution of 0.25-0.5 degrees. Viewing was binocular, but only the dominant eye (for most participants, the right eye) was recorded. The sampling rate for recordings was 1,000 Hz.

Stimulus presentation and interface with the eye-tracker was implemented with the EyeTrack software suite (University of Massachusetts, Amherst; www.psych.umass.edu/eyelab/software). Sentences were presented in 14-point fixed-width Courier font in 3-4 lines.

A calibration procedure was performed before the experiment, and re-calibration was carried out between trials as needed. Before the experiment began, each participant was instructed to read for comprehension as naturally as possible. Each trial began with a gray square on the left edge of the display, aligned with the location of the first character of the text. Participant triggered the appearance of the text by fixating on the square, and pressed a button on a game controller to indicate when they had finished reading. On half of the items, a comprehension question appeared after the text. Participants responded to the question by pressing the trigger buttons on the game controller—left for

"yes" and right for "no". The words "yes" and "no" were printed on the left and right of the screen under the question to remind participants which button to press.

Data analysis

Each trial was visually inspected to correct fixations for small vertical drift and delete fixations identified as blinks by the software. Fixations less than 80 ms in duration and within one character of the previous or following fixation were incorporated into the neighboring fixation. All remaining fixations shorter than 80 ms were excluded, since readers do not extract much information during such short fixations (Rayner, 1998). Fixations longer than 800 ms were also excluded, as these are likely to be due to track losses.

All critical regions for analysis were on the third line of the presented passage. The line began with two words before the verb immediately preceding the scalar trigger, and ended at least four words after the phrase 'but the rest'. There were four critical regions, illustrated in Table 3-6: (1) the triggering scalar expression, (2) a spillover region following the trigger, (3) the phrase 'but the rest', introducing the complement set, and (4) a spillover region following the complement set.

For each region, five eye-tracking measures were calculated. The first two measures, *first-pass time* and *right-bound time*, are limited to the reader's "first pass" through the sentence—fixations within a region that occur before the reader has fixated

Trigger	Trigger-spillover	Complement	Complement-spillover
some of the historic sites the famous historic sites	in her whirlwind visit,	but the rest	were too difficult

Table 3-6 Experiment 2: Critical regions for analysis.

on any later region. *First-pass time* is the sum of fixation times starting with the first fixation inside the region until the next fixation in an earlier or later region. *Right-bound reading time* is the sum of fixation times starting with the first fixation inside the region until the next fixation in a later region. Right-bound time thus includes any fixations in the region after a regression to an earlier region, but before any fixations on later regions.

Regression path time includes reading times not only on the region of interest, but also on earlier regions that are fixated in the course of regressions. It is the sum of all fixation times in the region or any previous region, starting with the first fixation inside the region until the next fixation in a later region.

Total time and *reread time* reflect later processing of the information in a region. Total time is the sum of all fixations in the region during the entire trial. Reread time is the sum of all fixations in the region that occur after a later region has been fixated: that is, the total time minus the first pass time.

For all reading time measures, the data for a particular region were excluded if the reading time was zero.

Reading time measures for each region were analyzed using linear mixed-effects models. Results for each SCALE TYPE were modeled separately, since the words in the critical regions were substantially different for the two types. Each model had a fixed effect for CONTEXT and random by-subject and by-item intercepts. The 3-level factor CONTEXT was coded orthogonally, yielding two contrasts. The first contrast compares the mean of the *upper bound* contexts to the mean of the *lower bound* and *neutral* contexts. This contrast is useful for determining whether *upper bound* contexts lead to implicature-related costs at the trigger and reduced cost at the complement set. The second contrast

compares the mean of the *neutral* contexts to the mean of the *upper bound* and *lower bound* contexts. This contrast is useful for determining whether the *neutral* context differs from the other two because the scalar expression has no antecedent in the context.

Given the large number of observations (~430 for each region/measure), the significance of fixed effects can be estimated using the t-statistic (Baayen, 2008). Fixed effects are considered significant at the 5% level if the absolute value of their t-statistic is greater than 2, and marginally significant (at the 10% level) if it is greater than 1.67.

3.7.3.2 Results

Mean reading times for each measure in the critical regions are given in Table 3-7.

Some/all scale

In the trigger region (see Figure 3-4), there was no significant contrast between the *upper bound* contexts and the other two contexts in any measure. However, reading times for the *neutral* contexts were significantly higher than in the other two contexts in several measures: right-bound time (t = 2.45), regression path time (t = 2.15), and total time (t = 2.95). There were no significant differences between conditions in any measure in the spillover region following the trigger.

In the complement set region (see Figure 3-5), reread times for *upper bound* contexts were marginally shorter than for the other two contexts (t = -1.92). Reread times for *neutral* contexts were significantly shorter as well (t = -3.35); thus, here the *lower bound* context was the outlier in having longer reread times. However, regression path times were marginally longer for *neutral* contexts in this region. In the spillover region

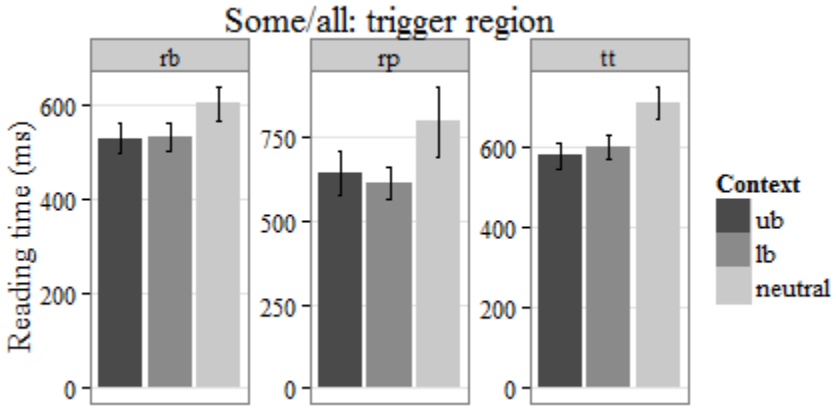


Figure 3-4 **Experiment 2: *Some/all* scale, trigger region.**
By-subject means. Error bars represent standard error. rb = *right-bound time*; rp = *regression path time*; tt = *total time*.

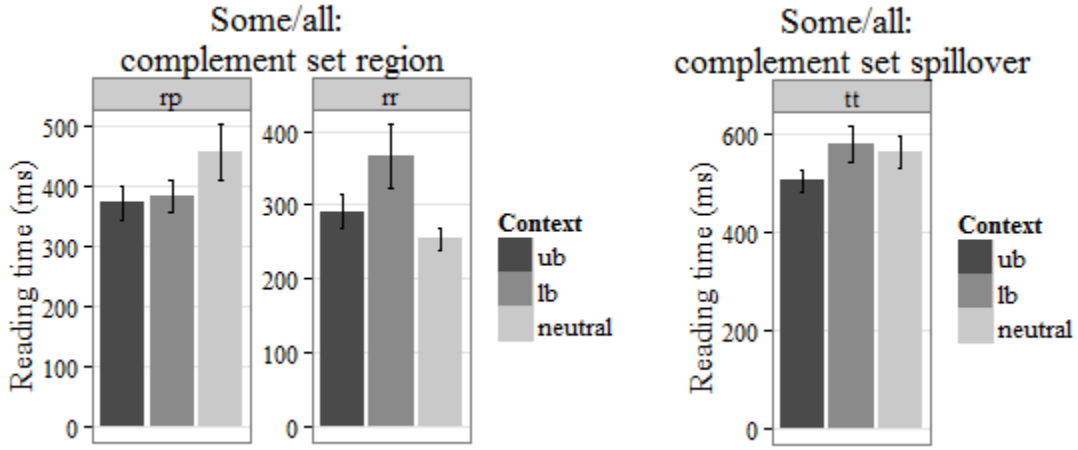


Figure 3-5 **Experiment 2: *Some/all* scale, complement set region and spillover.**
By-subject means. Error bars represent standard error. rp = *regression path time*; rr = *reread time*; tt = *total time*.

following 'but the rest', total reading time was significantly shorter in the *upper bound* contexts compared to the other two contexts (t = -2.05).

Ad-hoc scales

In the *ad-hoc* scale condition, the *upper bound* contexts did not differ from the other two contexts in any region, in any measure. However, there were some places where the *neutral* contexts differed from the other two.

In the trigger region (see Figure 3-6), reading times for the *neutral* contexts were significantly longer than the other two contexts in several measures: first pass time (t = 2.48), right-bound time (t = 2.99), and total time (t = 3.68). There were no significant differences between conditions in any measure in the spillover region following the trigger.

In the complement set region, reread times for *neutral* contexts were marginally longer than for the other two contexts (t = 1.88). There were no other significant differences between conditions in this region or in the following spillover region.

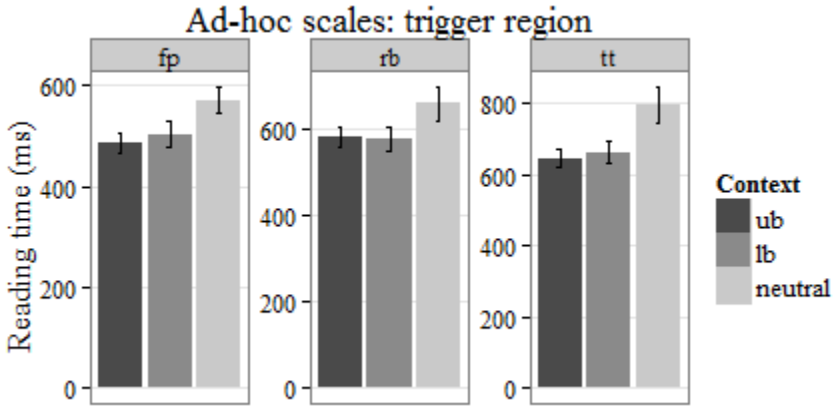


Figure 3-6 **Experiment 2: Ad-hoc scales, trigger region.**
By-subject means. Error bars represent standard error. fp = *first pass time*; rb = *right-bound time*; tt = *total time*.

		Trigger	Trigger spillover	Complement	Complement spillover
First-pass time					
Some/all	*UB*	435 (22)	693 (39)	333 (21)	417 (20)
	LB	449 (20)	736 (35)	339 (17)	436 (25)
	Neutral	483 (24)	732 (44)	339 (16)	406 (26)
Ad-hoc	*UB*	486 (20)	766 (37)	320 (17)	421 (26)
	LB	504 (26)	724 (32)	345 (21)	414 (27)
	Neutral	**572 (27)**	705 (36)	348 (24)	419 (23)
Right-bound time					
Some/all	*UB*	530 (32)	781 (46)	344 (23)	439 (21)
	LB	533 (30)	804 (32)	359 (20)	457 (25)
	Neutral	**603 (36)**	793 (40)	359 (17)	455 (29)
Ad-hoc	*UB*	581 (22)	835 (38)	324 (16)	458 (29)
	LB	576 (29)	803 (36)	350 (21)	457 (32)
	Neutral	**659 (39)**	849 (43)	360 (26)	460 (27)
Regression path time					
Some/all	*UB*	640 (64)	954 (99)	372 (28)	508 (37)
	LB	609 (47)	877 (40)	383 (26)	508 (33)
	Neutral	**792 (105)**	944 (49)	**456 (45)**	591 (56)
Ad-hoc	*UB*	687 (59)	938 (58)	411 (72)	549 (44)
	LB	694 (59)	902 (57)	363 (21)	513 (40)
	Neutral	709 (46)	1003 (73)	425 (72)	542 (48)
Reread time					
Some/all	*UB*	471 (53)	584 (92)	**291 (23)**	350 (37)
	LB	554 (79)	558 (42)	366 (43)	410 (67)
	Neutral	547 (60)	499 (53)	**254 (14)**	372 (35)
Ad-hoc	*UB*	559 (50)	697 (95)	361 (39)	412 (31)
	LB	552 (57)	714 (68)	294 (31)	409 (32)
	Neutral	500 (49)	622 (57)	**399 (45)**	415 (35)
Total time					
Some/all	*UB*	579 (32)	821 (44)	377 (27)	**507 (24)**
	LB	602 (30)	844 (33)	412 (28)	581 (36)
	Neutral	**710 (41)**	880 (38)	398 (20)	565 (31)
Ad-hoc	*UB*	646 (26)	895 (45)	403 (26)	569 (33)
	LB	662 (31)	868 (38)	393 (25)	577 (31)
	Neutral	**797 (50)**	917 (42)	432 (23)	568 (25)

Table 3-7 **Experiment 2: Average reading time for each measures.**
By-subject averages. Standard errors in parentheses.

3.7.3.3 Discussion

This experiment has several critical findings which help to clarify some of the previous results in the literature.

For the *some/all* scale, reading times at or immediately following the region containing 'but the rest' indicate that participants computed an upper-bounded interpretation for 'some' in the *upper bound* contexts, but not the *lower bound* or *neutral* contexts. The shorter reading times in these regions observed in the *upper bound* contexts replicated previous studies. However, there was no evidence that the upper-bounded interpretation of 'some' was more costly to compute. There was no region or measure where reading times were longer in the *upper bound* contexts. This finding supports the hypothesis that the costs observed in some of the previous experiments were due to peculiarities of the context, not to an inevitable feature of the pragmatic inference.

For the ad-hoc scales, reading times at 'but the rest' did not differ between conditions, so there was no evidence that participants computed upper-bounded interpretations of the scalar expression before they finished reading the sentence. This finding suggests that the processing costs that are so readily observable for scalar quantifiers are unusually early. Computing scalar implicatures which rely on less readily available relationships between lexical items may take considerably longer, or require even more robust contextual support. This finding also casts doubt on the findings I discussed in section 3.7.2.1, on processing costs exclusive readings of 'or' and scalar implicatures with ad-hoc scales. Both experiments on 'or' used contexts with no previous mention of the scalar alternative (Breheny, Katsos, & Williams, 2006, Experiment 1; Katsos, Breheny, & Williams, 2005, Experiment 2). Thus, the cost may be related to the

need to infer the relevant alternative, not compute the upper-bounding implicature. Similarly, the previous experiment with an "ad-hoc" scale required comprehenders to infer the intended relation between the weak and strong members of the ad-hoc scale (e.g. <'paint the house', 'paint the roof'>). It seems likely that the costs reported in all three of these experiments actually reflect the difficulty of identifying the relevant alternatives in the context, rather than computing the implicature itself. In fact, the implicature may not have been computed at all before the end of the trial.

Finally, for both scale types, processing the scalar trigger was significantly more costly when the previous context did not explicitly mention a member of the scale. This finding suggests that the processing costs observed in studies with less explicit contexts may be attributable to more general pragmatic processes that are not related to the implicature of interest.

3.8 General discussion

The goal of this chapter was to gain some insights on the computation of scalar implicatures during real-time comprehension. I discussed a dishearteningly diverse array of potential algorithms for computing implicatures, in the hopes that at least some of the options could be eliminated based on evidence from previous studies and Experiments 1 and 2. I focused on two aspects of the algorithm: the computation of the implicated meaning, and the use of context in determining the intended meaning. What did we accomplish?

On the first aspect—computing the implicated meaning—I did not manage to narrow down the options very much. Based on previous studies and the results of Experiment 1, I argued that some of the costs for implicature that have been observed in

previous studies are actually attributable to the upper-bounded meaning itself. The remaining results demonstrate that literal meanings become available earlier than pragmatically-enriched meanings. This generalization is compatible with most algorithms for computing implicated meanings.

Let's consider the three options I proposed in section 3.4.1.1: (1) implicated meanings are accessed directly based on certain trigger lexical items, (2) implicated meanings are computed through a domain-specific process, and (3) implicated meanings are computed through a domain-general inferential process. The first option—direct access to implicated meanings for certain lexical items—has the least support, but is not inconsistent. If we divorce the idea of direct access to implicated meanings from the idea that implicatures arise by default, then it is quite compatible with the results. Suppose that each meaning of a quantifier is stored with the lexical item, so that either literal or enriched interpretations can be accessed directly. Suppose that one of the meanings is chosen once there is enough evidence that it is the intended meaning. If sufficient evidence for a literal interpretation arrives faster than for an enriched interpretation, we would expect literal interpretations to become available earlier than enriched interpretations, even if both are accessed directly. All this is to say that direct access is still possible, as long as we are willing to push all of the cost of implicature onto accessing contextual information. The second and third options—domain-specific and domain-general multi-step computations—are more compatible with the observed timing of literal and pragmatic interpretations. Marty and Chemla's (2013) finding that implicature computation is dampened under conditions of high non-linguistic cognitive load suggests that at least part of the process is domain-general.

On the second aspect of the algorithm—using context—I focused on explaining the findings that already exist, without trying to propose a more general account for how context is used in real time comprehension. I found that the accessibility of relevant alternatives in the context had a substantial effect on whether costs were observed for computing upper-bounded interpretations of scalar expressions.

4 Children's understanding of implicature

It has been noted for some time that non-literal interpretation is an area of weakness for young children, both in the developmental literature and as a kind of "common knowledge." Everyone has an anecdote about a child misinterpreting someone's utterance by taking it too literally (although I have never heard one about a failure to compute an upper-bounded interpretation of 'some'). My goal in the rest of this dissertation is to explain why non-literal language is difficult. What aspects of pragmatic interpretation are in place early on, and what develops over time and with experience?

One reason to try to understand children's pragmatic competence better is to get a handle on how children solve one of the most general learning problems in language acquisition. Children must bootstrap their way into structure-meaning mappings without initially knowing much about structure or meaning. They need to be able to guess the intended meanings of particular strings of sounds, in order to make inferences about the structure and contents of the string. On the other hand, they need information from the linguistic input in order to guess what the intended meaning of an utterance might be. How do children break into this problem?

One possibility is that children have such a sophisticated understanding of the world around them that they are able to guess at the intended meaning of an utterance more often than you might think. Early on, most of their "interpretation" is really just making guesses based on the context, with the linguistic signal contributing very little. Eventually, they gather enough data—mappings from raw-ish strings to intended meanings—to infer how the syntactic/semantic representations of those strings must work.

An alternative extreme possibility is that children have a very poor understanding of the world around them, but a very highly developed ability to extract information from linguistic signals. Early on, most of their "interpretation" is really just decoding the syntactic/semantic structure, with the context contributing very little. In this case, it's harder to imagine what "data" about the intended meaning they might be collecting so as to eventually infer the mapping between the linguistic signal and the intended meaning.

Of course, the reality is somewhere in between these two extremes, but different literatures have made suggestions more on one side than the other. For example, the word-learning literature reports that young infants are able to infer the meanings of words based on subtle social cues and reasoning about the intention behind the social cues (Diesendruck & Markson, 2001). The literature on children's early sentence comprehension suggests that infants and young children often override the information in the linguistic signal in favor of their expectations based on world knowledge (Strohner & Nelson, 1974; Chapman & Kohn, 1978). On the other hand, studies of children's early parsing abilities suggest that children are unable to use contextual cues to constrain their parsing decisions (Trueswell, Sekerina, Hill, & Logrip, 1999; Snedeker & Trueswell, 2004). And of course, there is abundant and growing evidence that children struggle with indirect language of various kinds—metaphor, sarcasm, and scalar implicature, to name a few—until at least adolescence.

One lesson that I draw from the fact that children's apparent pragmatic competence varies substantially across different domains is that focusing on a single type of implicature would be a poor research strategy. With that in mind, we will explore several types of implicature in depth. In this chapter, I review the sizeable literature on

children's understanding of scalar implicatures, as well as the much less sizeable one on relevance implicatures. In the following chapters, I discuss two types of relevance implicature that have been largely ignored in the recent surge of interest in developmental pragmatics: indirect requests, and "parenthetical" interpretations of belief reports.

In any case of implicature, there are several types of competence that we need to consider. First, do children understand the principles underlying the implicature—for example, that conversational contributions should be relevant and sufficiently informative? There is good reason to suppose that something like Grice's Cooperative Principle is completely universal—indeed, communication would hardly be possible without it—and most likely need not be learned. This notion is made more plausible by the possibility that principles of communication are not specific to language.

Second, are children capable of the kinds of inferences that may be necessary for computing implicatures? There is no doubt that general inferential abilities develop over an extended period (Markovits & Barrouillet, 2002; Markovits & Barrouillet, 2004), and children will become capable of more complex, far-flung inferences as they get older. The important question is whether children have enough of this inferential capacity and the processing resources that support it to get off the ground with simple implicatures relatively early on.

Finally, are children capable of accessing the relevant contextual information? This question subsumes a number of others. Do children who understand the maxims of conversation necessarily understand what counts as relevant in a particular conversation? Do children encode the right contextual information so that it will be accessible later for

the purpose of implicature calculation? Are there limits on the contextual information that children can retrieve in the course of computing an interpretation?

In all cases I discuss, I come to similar conclusions about children's pragmatic competence. Children have the principles, reasoning abilities, and processing capacity they need to compute pragmatic inferences. Where they experience difficulty is in accessing relevant contextual information. The exact nature of this difficulty is unclear: children's behavior is consistent with any of the options mentioned above: (1) failure to properly encode contextual information, (2) difficulty retrieving information that they did encode, or (3) failure to grasp what information it is necessary to retrieve.

Finally, a note on the age range of interest. The studies I review below tested children as young as 3 and as old as 10 years of age, but mostly in the range of 5-6 years. I am interested in the possibility that the fundamental competence displayed by 5 year-olds is also present in children as young as 2-3, although perhaps obscured more by their greater ignorance about the world and social conventions, as well as their limited language processing capacities. Since 3-year-olds are capable of participating in conversation (although perhaps not conversations that anyone would call adult-like), they are old enough to know about how conversations work. Even younger children engage in contingent interactions that would require them to infer the speaker's communicative intention—following instructions, for example (which I discuss in the next chapter). No studies of implicature that I know of show 3-year-olds failing a task that 4-5 year-olds succeed at. The generally poor performance displayed by older children seems to have discouraged researchers from extending what successes have been observed to younger age groups. 3-year-olds are quite capable of participating in complex truth-value

judgment tasks (as my Experiment 7 demonstrates), and I know of no other potential barriers. My findings with 3-4 year-olds on indirect requests (Chapter 5) and belief reports (Chapter 6) suggest that work with younger children on scalar implicature might be fruitful.

4.1 Scalar implicature

There have been many experimental investigations of children's ability to compute scalar implicatures, using a wide range of methods and manipulations of scale and context. The general strategy is to present children with a statement, question, or instruction designed to evoke different responses depending on whether it is interpreted in terms of its literal meaning or an implicature-enriched meaning. Overall, children seem to generate pragmatic interpretations at a much lower rate than adults until quite late in childhood. However, manipulations of the context and the structure of the task reveal that even 4-year-olds are capable of computing implicatures under the right circumstances.

I begin my review with Noveck's (2001) study, which was largely responsible for setting off the research program on children's comprehension of scalar implicatures. It establishes a baseline for children's low rate of responding based on upper-bounded interpretations. I organize the rest of the review around two potential explanations for children's poor performance. First, children may fail to grasp the experimenter's intention to test implicated meanings rather than literal meanings. Second, children may fail to access relevant alternatives that are needed for computing the implicature. This problem may have multiple sources, and also be exacerbated by children's difficulty grasping the experimenters' intention.

4.1.1 Noveck (2001): Low rates of implicature in children

Noveck (2001, Experiment 3), following Smith (1980), tested French-speaking 8- and 10-year-olds using the same statement evaluation task he used with adults. Recall that the critical sentences are logically true, but underinformative and thus false under the pragmatic interpretation, as in (89), repeated below.

(89) Some giraffes have long necks.

While adults rejected such statements 59% of the time on average, 8- and 10-year-olds rejected them only 11% and 15% of the time, respectively. Both adults and children appeared to individually choose either a semantic or pragmatic strategy and apply it consistently. However, while only a third of the adults chose a semantic strategy, 95% of the children did so. Noveck concluded that children's competence with any particular "weak" scalar term is initially limited to literal or "semantic" interpretations; pragmatic interpretations follow later in development.

There are some concerns with this version of the statement evaluation task. One potential problem is that participants must evaluate the statements against their general world knowledge. Even if it's safe to assume that all the participants had the relevant knowledge—a risky assumption for children—there were no constraints on how they used it. For example, some participants who appeared to behave "semantically" may have been imagining a situation in which the upper-bounded meaning of the proposition was true. For example, (89) could be felicitous if one considers that baby giraffes do not have particularly long necks.

However, it seems that this problem cannot explain children's low rate of rejecting underinformative sentences. Pouscoulous and colleagues (2007, Experiment 1)

provided a more constrained context for statement evaluation. Participants were presented with four boxes and a variety of different plastic animals. In the critical test trial, all of the turtles are in the boxes, and participants were asked whether they agreed with the statement in (90), repeated below.

(90) Some turtles are in the boxes.

Neither children nor adults rejected underinformative statements like (90) at a higher rate than in Noveck's task. 9-10 year-old children rejected the sentence only 9% of the time, adults 47% of the time. Thus, children's low rate of pragmatic interpretations cannot be solely due to the role of world knowledge in the statement evaluation task.

A second potential problem, more serious than the first, is that since the statements are presented with no conversational context, participants have no basis to judge whether an upper-bounded interpretation was intended. As Katsos and Bishop (2011) point out, in a task involving three types of sentences (true, true-but-infelicitous, and false) but only two judgment choices (true or false), it is not at all obvious how the middle sentence type should be classified. It may be that adults, but not children, infer that the experimenter intends them to judge the felicity of the sentences, rather than their literal truth. In the next section, I review several studies which suggest that this factor does in fact partially explain children's low rate of upper-bounded interpretations.

4.1.2 Understanding the intention of the task

Guasti and colleagues (2005) attempted to highlight the importance of informativeness by providing training before administering the statement evaluation task. In their Experiment 1, they replicated Noveck's finding: Italian-speaking 7-year-olds

rejected the underinformative statements only 12% of the time. In Experiment 2, they provided training in which participants were presented with a picture (e.g. a grape) and two possible descriptions (e.g. ‘grape’ and ‘fruit’), and asked to indicate which description was “better.” The 7-year-olds had no difficulty choosing the more restrictive term for the picture. Then, in the statement evaluation task, they were more likely to give “pragmatic” responses (52% rejection of infelicitous statements) than when they had no training—a rate comparable to that observed in adults in the statement evaluation task (Noveck, When children are more logical than adults: Experimental investigations of scalar implicature, 2001). However, the effect of the training did not persist one week later (Experiment 3). Thus, the training did not teach children something new—they already knew that some descriptions are more informative than others. The training simply suggested to children that the following experiment might be about felicity, rather than truth.

Another way to help children recognize the intended meaning is to provide a richer context, so that the conversational contribution of the critical utterance is clear. Guasti and colleagues (2005) used a story-based truth value judgment task in their Experiment 4 to determine whether children would benefit from a richer context. Participants were presented with a story including visual supports. The target sentence was a description of “what happened” in the story, uttered by a puppet. For example, in one story a group of five soldiers are deciding to travel by motorbike or horse. After discussing the benefits of each option, all the soldiers eventually choose to ride a horse. When the puppet was asked to say what was happening in the story, she uttered the target sentence in (91).

(91) Some soldiers are riding a horse.

(91) is clearly infelicitous, since an active question under discussion in the story is how many soldiers will end up riding a horse rather than a motorbike. Evidently 7-year-olds noticed the infelicity, because they were much more likely to reject underinformative sentences in this task than in the statement evaluation tasks from Experiments 1-3. Their rate of rejection (75%) was comparable to adults' (83%). This result suggests that children's relatively low rate of "pragmatic" interpretations was not due to a difficulty with computing the implicatures, but rather a difficulty licensing the implicatures without a supportive context.

Papafragou and Musolino (2003) investigated whether a richer context would help even younger children compute scalar implicatures. In their Experiment 1 they tested Greek-speaking 5-year-olds with a task very similar to that used by Guasti et al. (2005) for 7-year-olds. Instead of asking children to judge the truth of the sentence, they asked whether the puppet had "answered well", to encourage children to focus on the felicity of the utterance rather than its literal truth. It should be noted that this removes the need for children to compute scalar implicatures: they can succeed on the task by simply noticing whether a given utterance is informative enough in context. Thus, this task can test whether children recognize a precondition for an implicature—underinformativity—but not whether children can actually compute the implicature.

Papafragou and Musolino found that 5-year-olds did not perform as well as the 7-year-olds in the Guasti study: they rejected underinformative sentences with 'some' only 12.5% of the time, while adults did so 92.5% of the time.

In Experiment 2, Papafragou and Musolino introduced two modifications to the task to encourage pragmatic interpretations. First, the judgment task was preceded by a training task involving informative descriptions of objects, very similar to that used in Guasti and colleagues' Experiment 3. Second, the stories were revised to emphasize the main character's performance on a particular task. The puppet was asked to comment on how the character did, rather than on the story in general. For example, in one story a character brags that he is very good at throwing hoops around a pole. He challenges Mickey to try it with three hoops. Mickey tries really hard and manages to put all the hoops around the pole. The puppet is then asked, "How did Mickey do?" She responds with the target sentence in (92).

(92) Mickey put some of the hoops around the poll.

This story strongly emphasizes the upper bound: it's important whether Mickey will succeed in putting all the hoops around the poll. (It's not clear from the description of the task whether the upper bound "all of the hoops" is ever explicitly mentioned in the story.) Thus, (92) is clearly infelicitous, because it's relevant that Mickey actually got all of the hoops around the pole. This contextual manipulation improved performance for 5-6 year-olds' performance: children rejected underinformative sentences 47.5% of the time (compared to 12.5% in Experiment 1). The results of these two experiments suggest that younger children need more contextual support to license scalar implicatures. They are less likely than older children or adults to infer that the upper-bounded interpretation is relevant.

Papafragou and Tantalou (2004) also elicited a relatively high rate of pragmatic interpretations from younger children by providing a supportive context and using action-

based responses rather than judgments. They presented Greek-speaking 4-6 year-olds with stories in which characters were assigned particular "jobs." The character would then report on what they had accomplished, and the child was asked to reward them with a prize if they had completed their task. For example, in one trial an elephant is given four paper stars and instructed to color them. The elephant goes into a dollhouse to do the coloring. When he emerges from the house, the experimenter asks, "Did you color the stars?" The elephant utters the target sentence in (93) in response.

(93) I colored some.

These stories, like those of Papafragou and Musolino's Experiment 2, strongly emphasize the importance of the upper bound, making the underinformative sentences clearly infelicitous. 4- to 6-year-olds refused to reward the character 77.5% of the time in such cases, indicating that they had computed the upper-bounding scalar implicature. They almost always justified their decision by invoking the stronger term *all*.

Papafragou (2006) directly compared children's performance on the action-based "reward" task to a more traditional judgment task. She found that 5-year-olds were much more likely to compute upper-bounded interpretations of aspectual expressions in the action-based task (Experiment 2) than the judgment task (Experiment 1).

Katsos and Bishop (2011) took a different approach. Rather than enriching the context to push children toward the upper-bounded interpretation, they provided a third response option. Children rewarded "Mr. Caveman's" utterances with a small, big, or huge strawberry, depending on how well he described a simple illustrated story. The third response option allows children to show that they recognize the infelicity of the underinformative utterances, without having to make a decision about which

interpretation was most likely intended. They found that although 5-6 year-olds rejected only 29% of underinformative utterances in a binary judgment task (Experiment 1), they showed a clear distinction between underinformative and fully-informative utterances in the ternary judgment task (Experiment 2). 16 of 18 children consistently rewarded underinformative utterances with the "big" strawberry, while rewarding fully-informative utterances with the "huge" one. The remaining two children were even stricter, rewarding the underinformative utterances with only a small strawberry. These results demonstrate that children are aware that the critical utterances are underinformative. However, in a binary judgment task without contextual support to indicate which interpretation is more appropriate, they err on the side of being tolerant of pragmatic infelicity. Katsos and Bishop point out that this willingness to accommodate infelicitous utterances also plays a role in adult judgments. In natural conversation, or experiments that allow more latitude in responses, adults tend to respond to true-but-infelicitous utterances with informative corrections: "Yes, but…"

To summarize, a large part of children's poor performance in experimental tasks may be due to the unnatural conversational situations involved, which often fail to provide needed cues to the intended meaning of the critical utterances. A variety of modifications help to rectify this problem and improve the performance of children as young as 4-5 years old. Providing training that highlights the felicity of a description improves children's performance, perhaps by giving them a clue that the intended utterances are intended to be felicitous, not just technically true (Guasti, et al., 2005; Papafragou & Musolino, 2003). Providing a story context, especially one that emphasizes the relevance of the stronger scalemate (the upper bound) makes it clear how the critical

utterance is intended to be evaluated (Guasti, et al., 2005; Papafragou & Musolino, 2003). Using an action-based task may make the purpose of the critical utterance more transparent (Papafragou & Tantalou, 2004; Papafragou, 2006). Finally, providing a third response option allows children to demonstrate their understanding of the informativeness of the utterance without having to determine what the speaker's intention was (Katsos & Bishop, 2011).

It should be noted that all of these manipulations except the pre-test training have also been shown to increase adults' rate of rejecting underinformative utterances and deriving upper-bounded interpretations. Thus, the pattern of children's performance does not differ from adults' qualitatively. The main difference would seem to be that when there is no context to license an upper-bounded interpretation, adults are approximately evenly divided on which interpretation they choose, while the majority of children choose a literal interpretation as their default. Although this difference may initially seem deep, it is put in context by the different default responses used by children in interpreting indirect requests and belief reports, which we discuss in the next two chapters. In the final chapter, I speculate that children's behavior may be driven by the distribution of literal and enriched uses in the input.

4.1.3 Availability of scalar alternatives

A second potential problem for children is that they may have trouble accessing relevant scalar alternatives in order to compute the upper-bounding implicature (Papafragou & Skordos, to appear). This problem is at least partially addressed by the contextual manipulations discussed in the previous section: making the upper bound

relevant in the context makes it easier for children to access it as an alternative. Another way to help children access scalar alternatives is to make them explicit in the task.

Chierchia and colleagues (2001) found that even with relatively rich contexts, only about half of 3- to 6-year-old English-speaking children generated exclusivity implicatures for disjunction (Experiment 2). They introduced the "felicity judgment task" to determine whether children who provide literal responses area actually insensitive to the fact that the critical utterances are informative. In the new task, after hearing a story, children are presented with two alternative descriptions. Both descriptions are logically true, but one is underinformative given the context. For example, in one story several farmers are deciding which of their animals to clean. After considering all the animals, each farmer decides to clean a horse and a rabbit. Then two puppets each provide one of the alternative sentences, as in (94).

(94) a. Every farmer cleaned a horse or a rabbit

b. Every farmer cleaned a horse and a rabbit.

3-5 year-old children correctly chose the more informative statement 93.3% of the time, suggesting that their difficulty with the truth value judgment task was not due to a lack of ability to compare the informativeness of two alternatives.

Similarly, Papafragou and Ozturk (2007) found that both children and adults were more likely to reject underinformative sentences with epistemic modals like (95) when they were presented side by side with their stronger alternatives, like (96).

(95) The mouse may be in the pink box.

(96) The mouse has to be in the pink box.

Skordos and Papafragou (2012) manipulated the availability of scalar alternatives by manipulating how the different kinds of test sentences were organized in the experiment. This makes the alternatives more or less salient without directly involving them in the interpretation of the critical sentences by making them part of the story. Children were tested on four sentence types: true and false 'all' sentences, and felicitous and infelicitous 'some' sentences. The task was to say whether a puppet "said it well" when using the sentence to describe a simple scene. One group of children heard the sentences in blocks, such that they encountered all of the infelicitous 'some' sentences first, then the felicitous 'some' sentences, then the 'all' sentences. The second group of children also heard the sentences in blocks, but this time the felicitous and infelicitous 'some' sentences were intermixed in the first blocks. The final group of children heard all types of 'some' and 'all' sentences intermixed throughout the experiment. The scalar alternative for 'some' ('all') should be maximally accessible for the third group of children, and minimally accessible for the first group, who heard all of the infelicitous sentences first before having anything else to compare them to. For adults, the ordering of the different trial types made no difference—they were at ceiling in all conditions regardless. For 5-year-olds, the order made no difference to their performance on the 'all' sentences or the felicitous 'some' sentences, but it had a huge effect on their performance on the infelicitous 'some' sentences. 19 of the 21 children who had heard all the trial types intermixed were able to correctly reject infelicitous 'some' sentences. By contrast, only 4 of the 20 children who heard all of the infelicitous 'some' sentences first were able to correctly reject them. Children in the middle group, who heard intermixed 'some'

sentences and then the ‘all’ sentences, showed middling performance: 11 of the 20 children correctly rejected underinformative utterances.

Another way of manipulating the accessibility of scalar alternatives is to change the scale. It is generally assumed that ‘all’ and ‘some’ form a scale in virtue of the entailment relation between them, as I discussed in section 2.2.1 (Horn, 1972). The scale is “context-independent”: ‘all’ is always a potential alternative to ‘some’, even when it is not strongly suggested by the context. Although a context-independent scale may make implicature computation easier for adults, it may actually be more difficult for children who have not yet learned that the scale is conventional.

Barner, Brooks, and Bale (2010) tested English-speaking 4-year-olds with questions involving either the quantifier ‘some’, as in (97), or a “contextually-specified” alternative—in (98), ‘the dog and the cat’ vs. ‘the dog and the cat and the cow’. Rather than attempting to license an implicature, they used ‘only’ to trigger an upper-bounded reading (cf. (a) vs. (b)). In the critical test trials for (97)-(98), children would be presented with a picture in which three animals—a dog, a cat, and a cow—are sleeping.

(97) a. Are some of the animals sleeping?

b. Are only some of the animals sleeping?

(98) a. Are the dog and the cat sleeping?

b. Are only the dog and the cat sleeping?

Children answered “yes” to quantificational sentences like those in (97) about two-thirds of the time, regardless of the presence of “only”. However, with sentences involving context-specific alternatives, children were much less likely (14%) to accept sentences with ‘only’ than sentences without ‘only’ (93%). The authors concluded that

children at this age are able to compute upper-bounded interpretations, but they lack knowledge of context-independent scales like the quantifiers.

This difficulty with context-independent scales may only be a problem for the youngest children, however. With 5-6 year-olds, Katsos and Bishop (2011) found no difference in performance between quantifier scales and a similar context-dependent conjunction-based scale in either the binary or ternary judgment task.

To summarize, making the scalar alternatives more salient in the task makes children more likely to compute upper-bounded interpretations. This can be accomplished by making the critical response into a comparison of two alternatives rather than a judgment about a single utterance, as in the "felicity judgment task" (Chierchia, Crain, Guasti, Gualmini, & Meroni, 2001; Papafragou & Ozturk, 2007). However, it should be noted that these results only show that children are aware of the different levels of informativeness of different alternative utterances, not that they are able to compute implicatures based on that knowledge. The distribution of descriptions in the experiment also makes a difference: children are more likely to get upper-bounded interpretations if scalar alternatives (e.g. 'all') are intermixed with the critical underinformative utterances (Skordos & Papafragou, 2012). Finally, ad-hoc scales may be easier for children—somewhat counterintuitively—since they do not require knowledge of a conventional scale like <'all', 'some'> (Barner, Brooks, & Bale, 2010). However, older children (5-6 year-olds) who perhaps have more experience with the conventional scales do not show improved performance on ad-hoc scales (Katsos & Bishop, Pragmatic tolerance: Implications for the acquisition of informativeness and implicature, 2011)

4.1.4 Summary: Scalar implicature

The wide variability in children's performance across different studies reveals areas of competence as well as difficulty. Children are clearly aware that some utterances are more informative than others, and under the right circumstances they use this knowledge to license and compute a scalar implicature. However, in contexts where the relevance of scalar alternatives is less obvious, children may not realize that an implicature is intended.

Another important lesson to take away from these studies is that binary judgment tasks are quite problematic for testing understanding of pragmatic felicity, and may seriously underestimate children's competence. Children, and even adults, are disposed to accommodate apparently infelicitous utterances if the context does not make it obvious that they are unacceptable, and no intermediate response option is available.

4.2 Relevance implicature

Although most of the developmental literature has focused on scalar implicature, there have also been a smattering of studies on relevance implicature. In this domain, children's performance has generally been reported to be even worse than with scalar implicatures.

Peter de Villiers and colleagues (2009) investigated the understanding of relevance implicatures in children aged 3;6 to 10;11. Children were presented with short conversations accompanied by pictures. In one story, the picture shows "a man with an amazed look on his face and a fork and tongs in his hands looking at an empty roasting pan on the kitchen counter. A boy is pointing to the dog lying under the table." Children heard the conversation in (99).

(99) The dad said: ‘What happened to the ham?’

The boy said: ‘The dog looks happy.’

Children were asked about the boy’s intended meaning: “What did the boy mean? Why did he say that?” Their answers were coded as adequate or inadequate. Examples of adequate and inadequate answers are given in (100) and (101), respectively.

(100) a. Because the dog ate the ham… the dog pooped. (4;4)

b. He meant that the dog looks happy because he ate it. (6;4)

(101) a. It made a mess. (3;7)

b. The dog has a smiley face. (3;11)

c. Because the dog just looked happy. (4;9)

Young children’s performance on this task was quite poor. 5-year-olds gave adequate answers less than half the time, and 4-year-olds less than 30% of the time. Children did not consistently provide adequate answers until nearly 10 years of age. However, this task is much too difficult to serve as a criterion for whether children understand relevance or how to compute relevance implicatures. Some of the critical utterances were extremely obscure. For example, in one story, the conversation in (102) was accompanied by a picture of a boy wearing glasses and standing in front of his mother, who is holding up three fingers.

(102) The boy asks: ‘Where are my glasses?’

His mom says: ‘How many fingers am I holding up?’

Understanding the mother's utterance in this conversation requires not only world knowledge that may not be shared by most preschoolers, but also an inference of many steps to get from the literal meaning to the intended meaning. Aside from the obscurity of the stories, preschool-aged children may have substantial difficulty answering metalinguistic questions like "Why did he say that?", even if they understand the conversation.

Bernicot, Laval and Chaminaud (2007) tested 6-, 8-, and 10-year-old children on relevance implicatures, among other types of indirect language, using a story completion task. In one story, Donald and Daisy Duck are in the yard. Donald asks Daisy, "Should I mow the lawn?" Daisy replies, "The nephews are taking a nap." Children chose between two story completions. In one, Donald Duck mows the lawn; in the other, he waters the flowers. If children understood Daisy's intended meaning—that Donald shouldn't mow the lawn because it would wake up the nephews—they should complete the story with Donald watering the flowers. This task allows children to demonstrate their understanding without having to provide any explicit judgments about the intended meaning. Even the 6-year-olds answered correctly 92% of the time; the 8- and 10-year-olds were near ceiling.

Verbuk and Shultz (2010) tested 5-7 year-old children on relevance implicatures and corresponding non-verbal inferences. For example, in one story, Cat and Dog are standing next to a table with a parrot on top, and a bowl full of food underneath. Cat tells dog, "Now it's your turn to do something. Feed the parrot, please. I'll sit in the living room and read." After a while, Dog comes into the living room. In the verbal condition, Dog says, "I put the empty bowl back under the table." In the non-verbal condition,

children are shown a second picture, in which the bowl under the table is now empty. In both conditions, children were asked a yes/no question about the story ("Do you think Dog fed the parrot?") and asked to justify their answer. This design controls for the difficulty of the inference by comparing performance in the verbal relevance implicature condition to the non-verbal inference condition. Overall, children performed significantly better on non-verbal inferences than verbal relevance implicatures. 6-year-olds provided correct responses 76% for non-verbal inferences, but only 42% of the time for the verbal relevance implicatures. 7-year-olds, however, achieved 80% accuracy on the relevance implicatures. The authors do not report accuracy separately for the youngest age group.

These three studies report a range of different success rates for relevance implicature for children aged 3 to 10 years. What are we to make of this variability? Unfortunately, unlike in the case of scalar implicatures, there is are so many differences across these studies—not only in the task, but also in the particular relevance implicatures tested—that it is difficult to draw general conclusions. These tasks could have been difficult for children for any number of reasons.

In future work on relevance implicatures, it would be useful to constrain study to specific types of relevance implicature, so that other factors can be varied more systematically. In the next chapter, I propose that indirect requests are an excellent place to begin.

4.3 Summary

Children's low rate of generating scalar implicatures has been held up as a prime example of their general difficulty with non-literal language. I have argued that their poor performance is due to difficulty grasping the experimenter's intention in using certain

utterance, and difficulty identifying and accessing relevant alternatives in context. I have suggested that these problems may be limited to experimental situations, which make use of contexts that are severely impoverished compared to those of real conversation. Since there is no evidence on how well children understand scalar implicatures in their day to day lives—and indeed it would be rather difficult to collect the relevant data—I can only speculate on how children's performance in experiments relates to their ability to infer implicated meanings outside experimental contexts. Since children seem to have the necessary principles and procedures for computing implicatures in favorable contexts, it may be that they do so quite well in the real world. Real conversations may provide exactly the kind of favorable context that they need. Alternatively, it may be the case that the greater richness and complexity of real conversations are lost on children if they lack the processing capacity to make use of all of it.

As for relevance implicatures, there is more anecdotal evidence that children frequently fail to compute them in real conversations. However, failure to understand extremely far-flung inferences, sarcastic jokes, and metaphorical language, among other things, does not entail that children are hopeless at understanding how the maxim of relevance licenses indirect speaker meanings. In the next two chapters, I investigate two more constrained cases of relevance implicature, and show that 3-4 year-old children show surprising readiness to compute non-literal meanings.

5 Children's interpretation of indirect requests

In the previous chapter, I argued that the developmental literature that has been most visible to those with a Gricean approach to pragmatics—the literature on scalar implicatures—underestimates children's pragmatic competence. For a more well-rounded view, it is important to look at children's performance with other kinds of implicatures.

This chapter focuses on young children's understanding of indirect requests (or "indirect directives"). Indirect requests are utterances that are intended to act like imperatives, inducing the listener to perform some action, without having that intention encoded in the literal meaning. In fact, in most indirect requests, the illocutionary force of the utterance taken literally is quite different from the intended one. For example, the commands in (103) can be expressed by a variety of different questions (104) or assertions (105)-(106).

(103) a. Send me a draft by Monday.

b. I {demand/request/ask} that you send me a draft by Monday.

(104) a. {Can/Could/Would/Will} you send me a draft by Monday?

b. {Can't/Couldn't/Shouldn't} you send me a draft by Monday?

c. Do you want to send me a draft by Monday?

(105) a. I {need/want/expect/would like} you to send me a draft by Monday.

b. You'd better send me a draft by Monday.

c. (I think) you should send me a draft by Monday.

d. You'll want to send me a draft by Monday.

(106) a. I look forward to seeing a draft by Monday.

b. Monday is the absolute latest I can look at a draft.

c. I'll be checking my inbox on Monday.

Although some forms of indirect requests are somewhat conventionalized—the question forms in (104), for instance—others are arbitrary and context dependent, as in the statements in (106).

All indirect requests, regardless of conventionality, can be argued to arise from a relevance implicature. In a conversation where the listener's ability or willingness to send a draft by Monday are not under discussion, the questions in (104) are irrelevant, if interpreted literally. If the listener assumes that the speaker is in fact attempting to make a relevant conversational contribution, then he can infer that the speaker's intended speech act is a request.

Searle (1975) argues that indirect requests usually achieve their intended purpose by asking or commenting about the felicity conditions for performing a directive illocutionary act, or the reasons for doing so, or some combination of both. A speaker cannot felicitously issue a directive unless he wants the listener to perform the requested action, and thinks it would be possible for the listener to do so. Thus, some indirect requests involve a statement of the speaker's desire (e.g. 'I want a draft by Monday') or a questions about the possibility of the action (e.g. 'Could you send me a draft by Monday?'). In issuing a directive, a speaker predicates a future act of the listener. Thus, some indirect requests involve a statement or question about that future action (e.g. 'Will you send me a draft by Monday?'). Indirect requests based on the reasons for the request

can take any number of forms, stating or asking about the reasons for performing the action (e.g. 'I can't read the draft after Monday.').

For several reasons, indirect requests are a very promising domain for investigating children's ability to compute relevance implicatures. Children are very familiar with the communicative intention of a request. Parents ask (or order) their children to do things quite frequently. Newport, Gleitman, & Gleitman (1977) argue that since mothers have little in common with young infants, requests for action are essentially the only topic of conversation available. Few quantitative estimates of the rate of directives in maternal speech are available; they range from 15% to about a third (Hoff-Ginsberg, 1986; Pratt, Kerig, Cowan, & Cowan, 1992; Schaffer & Crook, 1980). Older children capable of producing utterances of the necessary complexity make frequent requests of their own (Garvey, 1975). Thus, unlike other cases of relevance implicature that have been tested previously (reviewed in section XX), children should be quite able to represent the intended meaning. The only potentially missing component is the ability to link the literal meaning of the utterance to the indirectly-communicated request.

Second, indirect requests display a broad range of conventionalization, as I mentioned above. Some forms are at least as conventional as the upper-bounded meaning of 'some', if not more so. Others are completely arbitrary. It is therefore possible to hold the intended meaning constant and manipulate the degree of conventionality of the utterance form. However, it is important to note that the kind of conventionality at work in indirect requests is somewhat different from the conventionality of the <*all*, *some*> scale. While the upper-bounded meaning of an utterance containing 'some' can usually

be tied to the quantifier, the request meaning of a conventional polite request cannot really be pinned on, say, the modal auxiliary or the interrogative functional head. This difference presumably has to do with the fact that quantity implicatures tend to restrict the meaning of utterances, while relevance implicatures expand them. It seems plausible for the word 'some' to be interpreted as *some but not all* in some contexts. It's simply not possible to pull the same trick with 'could' in (107)—it would have mean something like *I request*. At minimum, the enriched meaning would have to be tied to a construction—some larger chunk of structure.

(107) Could you send me a draft by Monday?

Finally, as with other relevance implicatures, indirect requests can involve inferences of varying levels of complexity, and make reference to relevant information that is easy or difficult to access. For example, the utterances in (108) and (109) communicate the same implicated meaning (110) in basically the same way, but (109) requires the listener to have more background knowledge.

(108) It's time to go to school.

(109) It's 7:40.

(110) Pack your bags and head for the door.

Given all this variation, it is possible to test a wider range of hypotheses about what makes implicatures more or less difficult for children to compute in comprehension. In the next section, I will review the previous literature on children's understanding of indirect requests. In section 5.2, I report a series of experiments designed to explore a

question that has not been addressed by the previous literature: do 3-4 year-old children know what counts as relevant in context?

5.1 Evidence from previous literature

Children's understanding of indirect directives has been studied for a variety of different purposes. For a developmental linguist, the utterance forms that mothers use to express different conversational functions is of primary interest. For a clinical psychologist, parental strategies for ensuring children's compliance is of utmost importance. Nevertheless, there are several generalizations to draw from this diverse body of research. First, parents use both indirect and direct requests with their children, and even very young children understand and comply with both types at least some of the time. Second, the more indirect a request is, the less likely children are to understand it, but understanding does not necessarily improve with age. It is not consistently the case that younger children fail to understand indirect requests that older children can understand. Also, since the more conventional forms of indirect requests tend to be less indirect, it is not always possible to determine whether children's different levels of success are due to directness or conventionality. Finally, children clearly rely on both the context and the linguistic content to interpret indirect requests, although it is not always clear how critical each source of information is in any given case.

Since the studies I review below make use of a range of different types of indirect request, it will be useful to establish standardized names for the most common types. Since compliance is generally not optional in the situations studied in the developmental literature, requests are more often referred to as "directives." *Question directives* are of two types: *explicit action questions* explicitly state the desired action (111), while

situation questions only imply the desired action by mentioning some problematic state of affairs (112). *Declarative directives* are also distinguished by whether they explicitly state the desired action: *explicit action declaratives* (113) and *situation declaratives* (114). *Situation declaratives* are also called *hints*. Note that the studies I report below may not have used these labels for the different types.

(111) *Explicit action question:* Could you close the window?

(112) *Situation question*: Is the window open?

(113) *Explicit action declarative*: I can't close the window.

(114) *Situation declarative*: The window is open.

Of these four types, only the explicit action questions are conventional. Explicit action questions are the most direct, followed by explicit action declaratives, then the situation questions and declaratives.

5.1.1 Indirect requests in natural parent-child interactions

Shatz (1978) observed relatively unstructured play sessions with three mothers and their children, aged 2;1-2;4. She found that children attempted to comply with their mothers' requests about half the time, regardless of whether the request was phrased as an imperative or as a question. Children responded to questions literally—a verbal or non-verbal "yes" with no attempt to complete the requested action—only about 9% of the time. Children also responded appropriately equally often to *explicit action questions* and less direct, less conventional *situation questions* like (115). About 40% of mothers' question directives were *situation questions*.

(115) Are there any more suitcases?

Schaffer & Crook (1980) investigated mothers' (n=24) ability to elicit compliance from their 15 and 24 month-old children. Like Shatz, they found that mothers used both direct and indirect directives, and there was generally little difference in the efficacy of each type, although imperatives had an edge in some situations. Lytton & Zwirner (1975), in a much larger study (n=136) of parents interacting with their 2.5 year-old boys, found that although indirect requests were a relatively small proportion of verbal requests, they tended to be more effective than direct imperatives.

These studies confirm the anecdotal intuition that young children are not completely flummoxed by indirect requests. However, they are not informative as to how children go about interpreting them. One possibility is that children assume that all of their parents' utterances are requests (not an unwarranted assumption, perhaps) and attempt to respond with an action as often as possible. The fact that their actions are often appropriate could be an artifact of external constraints on possible actions. For example, a child operating under this heuristic might respond appropriately to a request like (115) simply by noting that the utterance involves suitcases, and there is nothing else to do in the situation except find another toy suitcase. A more sophisticated possibility is that although children do not assume that people's utterances are universally requests, they understand the demands of the physical context well enough to anticipate potential requests and align them with their parents' utterances.

Alternatively, it may be the case that children go about interpreting indirect requests in the same way that adults are assumed to: by decoding the literal meaning of the utterance and inferring its intended meaning and illocutionary force. Shatz (1978) was reluctant to attribute this degree of linguistic and pragmatic sophistication to 2-year-olds:

"One would hardly want to explain the children's behavior by granting them the ability to analyse linguistic messages in terms of their literal meaning, their relation to context, and the speaker's obligation to follow maxims of conversation, and then to infer the appropriate illocutionary force of indirect utterances. Hypotheses less demanding in terms of cognitive and social knowledge are required" (p. 45). However, this assessment may be too pessimistic. 2-year-olds are certainly capable of decoding the literal meaning of simple questions. As for the pragmatic inference, describing it in terms of "maxims of conversation" and "illocutionary force" makes it sound very complicated—so much so as to be implausible even for adults under the time pressures of normal conversation. But as we have already discussed, the algorithm for adult-like pragmatic inference need not take this literally Gricean form.

5.1.2 Experimental investigation of indirect request understanding

5.1.2.1 Naturalistic experimental paradigms

Several experiments have tested understanding of indirect requests in naturalistic situations, where children may not realize they are being tested. In this way the context can be more controlled to determine whether children succeed in complying with indirect requests based on non-linguistic cues alone. The form of the utterance can also be manipulated to determine how much children rely on information from the linguistic signal.

Ervin-Tripp and colleagues (1987) tested 3-, 5-, and 7-year-olds on five requests that were naturally interspersed throughout the testing session. (Their main experiment will be discussed in the next section.) Some requests were contextually predictable in that they were required by the situation: for example, the experimenter "accidentally" knocks

some cards off the table, and requests that the child pick them up so they can continue the experiment. Other requests were less predictable: for example, the office door is left open at the beginning of the session, and the experimenter requests that the child close it midway through. There were five forms of request, and each child heard only one of each during the session. These included (in order of increasing indirectness) explicit action questions (116), explicit action declaratives, situation declaratives (118), situation questions (119), and exclamatory mentions of the object (120). Only the explicit action questions were conventional forms for indirect requests.

(116) Can you find my watch?

(117) I can't find my watch.

(118) My watch is missing.

(119) Is my watch over there?

(120) Oh, my watch!

For most forms, the context of the request made no difference, suggesting that children were not relying exclusively on the demands of the situation to infer that a request was intended. For the least informative utterance type—the exclamatory object mentions like (120)—children were less likely to help out if the request was not contextually predictable.

5- and 7-year-olds complied with the requests regardless of the utterance form. The probability of compliance decreased somewhat for more indirect requests, but children were still more likely than not to help out in response to the exclamatory object mention (about 70% compliance, compared to 100% for explicit action questions). There was no striking difference between the conventional forms (explicit action questions) and

non-conventional forms. By contrast, 3-year-olds were much less likely to comply with the more indirect requests. Even the explicit action declaratives like (117) elicited only 40% compliance, compared to 82% for the explicit action questions. This pattern may reflect an effect of both the conventionality of the request and the level of indirectness.

To determine whether 5-year-olds were at all sensitive to the relevance of the utterance when interpreting it as a request, some children were tested on anomalous requests like (121).

(121) My pen is blue.

This statement, uttered in a situation where there is only one pen and it is obviously blue, cannot reasonably be used to request the pen. Although 5-year-olds were significantly less likely to "help" in response to these odd statements, 3 of the 7 children tested did simply hand over the mentioned object. This suggests that children may be less sensitive than adults to what makes an utterance appropriate to the situation and the intended meaning. We will return to this intriguing finding with our experiments in section 5.2.

Spekman and Roth (1985) took a more structured approach, but the task was still relatively naturalistic. Children (3-, 4-, and 5-year-olds) played with toys in simple pretend scenarios (e.g. giving a baby doll a bath), following instructions from the experimenter. They tested six types of requests: direct imperatives (122), explicit action questions (123), permission questions (124), desire/need statements (125), situation questions (126) and situation declaratives (127). Both the explicit action questions and the permission questions are conventional forms of indirect request; the other three were non-conventional.

(122) Wash the baby.

(123) Could you give me the baby's pajamas?

(124) May I have the baby?

(125) I need a towel to dry the baby.

(126) What happened to the baby powder?

(127) Some water spilled on the table.

All three age groups complied with requests at the same rate. However there were significant differences based on utterance type. Children were less likely to comply with direct imperatives at the first utterance, although they almost always complied after a repetition. Whenever they did not comply, it was because they decided to perform the intended action themselves (e.g. putting the hat on the doll instead of giving it to the experimenter to put on). Children were also less likely to comply with the most indirect requests—the situation questions and declaratives—but their compliance did not improve much after repetition. Noncompliance consisted in a literal verbal response or no response at all. (Unfortunately the rate of literal responses is not reported.) However, it should be noted that even 3-year-olds complied over 60% of the time with the most indirect utterance types (situation questions and declaratives)—strikingly better performance than the 20-30% observed by Ervin-Tripp and colleagues (1987) for the same utterance types.

Both of these studies demonstrate that at least by 3 years, children are sensitive to the form of an utterance used as a request: they do not simply assume that all utterances are requests for action. In general the level of indirectness rather than conventionality determined children's success rate. Children are also reliant on context. 3-year-olds'

different levels of success in the two studies may be attributable to their better understanding of the structure of the situation and discourse context in Spekman and Roth's experiment. Finally, the 5-year-olds' willingness to treat irrelevant utterances as requests in Ervin-Tripp's study suggests that they may not yet understand what level of relevance is expected from conversational contributions.

5.1.2.2 Third-person story paradigms

One limitation of naturalistic compliance tasks is that children may choose not to comply for reasons unrelated to comprehension. Studies that use third-person stories instead of involving the child directly avoid this problem. They also have the advantage of providing relatively little situational support to help children guess what the speaker might want. However, as I discussed in section 4.2 on relevance implicatures, metalinguistic judgments are so difficult that treating them as criterial almost certainly underestimates children's competence.

The story-based tasks used by different studies have very similar structures. The experimenter tells a story accompanied by pictures, in which a speaker—usually a parent figure—issues a request in some form to the child protagonist. For example, in one story used by Elrod (1983; 1987), "Bill and his friend are playing. They come into the kitchen for a drink. Bill's mom has been cleaning the kitchen." Then Bill's mom says either the direct imperative in (128) or the indirect situation declarative in (129).

(128) Please stay out of the kitchen.

(129) I just waxed the floor.

The child participant is asked, "Why did she say that?" Children's responses were coded for whether they referenced Scott's mom's *intention* (e.g., "She wants him stay out of the kitchen") or a potential *consequence* (e.g. "The floor would get dirty if they walked on it"). Then the child is asked to finish the story: "Let's see what you think Bill and his friend will do." The child is presented with three pictures corresponding to potential story completions: "Will Bill and his friend (a) stay out of the kitchen, (b) come in to see how shiny the floor is, or (c) come into the kitchen to get a drink?" Option (a) is considered an appropriate (compliant) *nonliteral* response; option (b) is an inappropriate (non-compliant) *literal* response; option (c) is a non-compliant response without a literal component. Success on both of these questions requires children to take first the speaker's and then the listener's perspective.

Elrod (1983) found that performance on this task significantly improved with age: 5-6 year-olds were more likely than 3-year-olds to provide appropriate explanations and story completions. However, there was no effect of the request type: all age groups performed equally well for both direct and indirect requests. In a very similar study, Elrod (1987) found that story completions improved with age, but there was no interaction with the request type. However, while older children were equally good explaining direct and indirect requests, younger children provided better explanations for direct than indirect requests. This finding certainly demonstrates that older children are better at expressing why people say certain things when they have certain intentions, but it does not necessarily indicate that younger children are less competent at understanding indirect language. A further interesting finding from this study is that when presented with the story pictures without narration, older children were more likely than younger

children to guess that the character in the story would be making a request. This finding again speaks against the hypothesis that the youngest children expect all speakers' utterances to be requests: that expectation apparently requires a relatively rich context.

Ervin-Tripp and colleagues (1987) used a very similar paradigm to test comprehension of requests in 3-, 5-, and 7-year-olds. All of the requests were indirect, either situation questions like (130) or situation declaratives like (131). There was also a "silent" condition, in which the story ended before the speaker uttered the indirect request. Half of the stories were about "helping": the target response was a helpful action. The other half were about "prohibition": the target response was abstinence from a "naughty" action. As in Elrod's studies, children were asked to complete the story and explain the speaker's utterance. The story completions were not constrained by specific choices.

(130) Is the door open?

(131) The door is closed.

In the "helping" stories, 3-year-olds were about half as likely as the older children to complete the story with the protagonist acceding to the indirect request. In the "silent" condition, no children suggested that the protagonist would help. By contrast, in the "prohibition" stories, children of all ages suggested that the protagonist should stop his naughty action in both the indirect request and silent conditions. Clearly, children understood the prohibition contexts better and thus were more able to respond appropriately to indirect request. Children's better understanding may be attributable to their greater familiarity with social scripts involving prohibition: it's plausible that children have more experience being prohibited from actions than with being asked to

help. Thus, the knowledge that it's bad to use crayons on the wall would be more accessible than the knowledge that someone might want the door opened for them if they're carrying a large load of groceries. However, even if we accept this assumption, there are still two possible explanations for the younger children's different behavior in each condition. Younger children might respond appropriately in the prohibition contexts by relying exclusively on their knowledge of prohibition-type social scripts, essentially ignoring the linguistic input. Alternatively, children's algorithm for interpretation in both types of story might rely on linguistic information and an adult-like inference process, but break down in the "helping" contexts because of a lack of relevant knowledge to feed the inference.

Bernicot & Legros (1987) used similar stories, but rather than having children complete the stories, they asked children to judge the speaker's emotional reaction when the listener fails to comply with his direct or indirect request. Children chose between three options illustrated with pictures: "[The requester] is very angry," "[The requester] is unhappy", or "[The requester] finds everything's all right." Children were scored using a very strict criterion: they were considered to have given "directive" responses if they chose either the "very angry" or "unhappy" picture, and provided an appropriate justification for their choice. Thus, this task only requires children to take the speaker's perspective.

The critical utterances at the end of the stories were of three types: direct imperatives, indirect situation declaratives, or non-directives. Half of the stories provided a strong context for a request, and half only weakly suggested the possibility of a request, if at all. 5-6 year-olds showed sensitivity to both of these factors: they were more likely

to provide directive-type responses in strong contexts and more direct utterances. 3-4 year-olds, on the other hand, rarely provided directive-type responses, and showed significant differences between conditions (although the differences mostly trended in the right direction). The authors concluded that the 3-4 year-olds were much less able than the older children to understand indirect requests. However, since these children also performed quite poorly with direct requests, their difficulty seems to lie more with the task than with the form of the request.

One potentially important difference between the stories used by Bernicot and Legros and those used by Elrod, Ervin-Tripp and colleagues is that the requester was a peer of the child protagonist, rather than an authority figure. Thus, the stories may not have fit as neatly into social scripts that the child participants were most familiar with. Once again, it is difficult to interpret 3-4 year-olds' failure. Do they fail to derive request interpretations of certain kinds of utterances because they lack a pragmatic mechanism for doing so? Or do they simply have difficulty accessing the relevant knowledge to feed the pragmatic inference process?

5.1.3 Indirect requests in forced choice action-based tasks

Naturalistic compliance tasks can underestimate children's comprehension because children can choose not to comply for other reasons. Story-based forced choice tasks avoid that problem, but are more demanding because they require children to adopt both the speaker's and the listener's perspective. Action-based forced choice tasks can combine the best features of both. A speaker makes an indirect request of the child in a situation where the child has a limited number of response options.

Only one previous study that I know of has used this type of task to investigate children's understanding of indirect requests. In their Experiment 3, Schulze, Grassmann, & Tomasello (2013) had 3-year-olds give objects to a puppet based on hints having to do with whether the preconditions of use for the object were fulfilled. For example, in one trial, a puppet says, "What do we have for breakfast?" The experimenter presents some cereal and a roll, and asks the puppet which one she wants. Then the puppet utters a statement indicating either that the preconditions for using one of the objects (the "target object" are fulfilled (132) or not fulfilled (133). The child's task is to hand over one of the objects. In the "fulfilled" condition the correct response is the target object; in the "unfulfilled" condition it is the alternative object.

(132) I bought milk.

(133) The milk is all gone.

Adult controls selected the intended object 100% of the time in the *fulfilled* condition and 93% of the time in the *unfulfilled* condition. 3-year-olds chose correctly in the *unfulfilled* condition 73% of the time, but they were at chance in the *fulfilled* condition. 4-year-olds chose the correct object more often than chance in both conditions (~65%), but still much less often than adults.

This task removes the need for the child to infer that the utterance is a request: it's clear that they are meant to hand over one of the objects. It provides a better test of how well children understand the relevance and intended meaning of the utterance, because they cannot succeed at the task by simply offering the mentioned object. The results demonstrate that even 3-year-olds were able to understand the "hints" provided by the puppets. Interestingly, 3-year-olds—in contrast to 4-year-olds and adults—did not

consider statements about a fulfilled precondition to be relevant to the choice of object. This behavior stands in contrast to Ervin-Tripp and colleagues's (1987) observation that at least some 5-year-olds were willing to consider even extremely infelicitous utterances as requests.

This discrepancy is difficult to interpret, since there is no other evidence available on how much children are willing to accommodate infelicitous or underinformative requests. Both results might be seen as similar to children's behavior in studies of scalar implicature, but in different ways. On the one hand, children's willingness to accept underinformative utterances shows tolerance of infelicity. Ervin-Tripp's accommodating 5-year-olds were also tolerant of infelicity, and tried to make the best of the input they got. On the other hand, children's willingness to accept underinformative utterances may also reflect a tendency towards literal interpretation in impoverished contexts. Schulze and colleague's 3-year-olds also tended toward a literal, underinformative interpretation of the fulfilled-precondition hints.

In Experiments 3-5, we attempt a more systematic investigation of what 3-4 year-olds are willing to accommodate in an indirect request.

5.2 Experiments 3-5[1]

Experiments 3-5 used a simple forced choice action-based task, similar to that of Schulze and colleagues. The situation made clear that the critical utterance should be

[1] These experiments were carried out in collaboration with Kaitlyn Harrigan, with additional assistance from Myles Dakan and Ilina Stojanovska.

interpreted as a request. The participant's task was to infer the specific intended meaning of the utterance by understanding how it related to the context.

Experiment 3 was a pilot study to establish that 3-year-olds were able to interpret oblique statements about the objects as requests. Experiments 4 and 5 investigated whether 3-4 year-olds are able to distinguish utterances that are appropriate as indirect requests from those that are not.

5.2.1 Experiment 3

The goal of Experiment 3 was to establish that young children could interpret non-conventional indirect requests in a fairly simple experimental paradigm. Since the critical utterances were similar to the "situation declaratives" of previous studies, there was some question as to whether 3-year-olds would understand them as requests.

5.2.1.1 Methods

Participants

15 children aged 3 years (3;0) to 4 years, 2 months (4;2) participated in the study (3.02-4.16 years, mean = 3.7 years, 7 girls). Participants were recruited from the Center for Young Children preschool or the Infant Studies Database at the University of Maryland. All participants were typically-developing monolingual English-speakers.

Design and materials

In each trial, the experimenter offered a choice of two toys to a puppet named Froggy. Froggy responded with a statement hinting at which toy he wanted, by stating something *positive* or *negative* related to one of the toys without mentioning it by name. The adult-like response is to interpret Froggy's utterance as an indirect request, giving

him the hinted object in the *positive* condition and the other object in the *negative* condition.

For example, in one trial, the two toys were a boat and a dump truck. To suggest that he wanted the dump truck, Froggy would utter (134) in the *positive* condition and (135) in the *negative* condition.

(134) *Positive*: I like playing with dirt!

(135) *Negative*: I don't like playing in the water.

For each pair of objects, we created four sentences, manipulating the STATEMENT TYPE and target object. The four versions of the sample trial are given in Table 5-1. We created four lists with 16 experimental items, eight each of *positive* and *negative* statements. The four versions of each item were distributed across these four lists.

In addition, we created 4 practice items which did not vary across the four scripts. In these items, Froggy's utterances explicitly mentioned one of the toys, as in (136) (*positive*) and (137) (*negative*).

(136) I like play-doh: it's squishy!

(137) I don't like aliens: they're scary.

Statement Type	**Target object**	
	Boat	*Dump Truck*
Positive	I like playing in the water!	I like playing with dirt!
Negative	I don't like playing with dirt.	I don't like playing in the water.

Table 5-1 Experiment 3: Target sentences for sample item.

Procedure

Sessions took place in a quiet room with the child seated at a table next to one experimenter. The other experimenter sat across the table and operated the puppet. The first experimenter introduced the game: *Today we're going to play with our friend Froggy. We want him to have a good time, so we brought lots of toys and other stuff for him. We'll give him some choices, and it will be your job to give him the one he wants.*

After obtaining the child's assent to participate, the first experimenter began the first trial placing two objects on the table in front of the child: *Check it out! We have a [boat] and a [dump truck]*. The experimenter then addressed Froggy: *Froggy, which one do you want to play with?* The experimenter operating the puppet would then utter the target sentence. If the child did not immediately hand over one of the objects, the first experimenter would prompt her to do so: *Can you give him the one he wants?* When the child handed over one of the objects, the puppet responded enthusiastically regardless of which object the child chose. The experimenter operating the puppet recorded the child's response.

Data analysis

Children's responses were coded as the "related" or "unrelated" object. In the *positive* condition, the "related" object is considered correct; in the *negative* condition the "unrelated" object is correct. Accuracy rates were calculated for each condition. Accuracy was modeled using mixed effect logistic regression, with fixed effects for the statement type and the participant's age, as well as the maximal by-subject and by-item random effects structure (random by-subject and by-item intercepts, random by-subject

and by-item slopes for the statement type). Binomial tests were used to compare accuracy to chance levels.

5.2.1.2 Results

Children were highly accurate in both conditions: 82% for *positive* statements, 78% for *negative* statements (both above chance, p's $<< 0.001$). The small difference between the two statement types was marginally significant ($p = 0.084$). There was no significant effect of participant age.

5.2.1.3 Discussion

Children overwhelmingly chose the intended object in response to both positive and negative statements. Knowing that the statement was intended as a request, they were able to infer the content of the request.

With this high level of performance as a baseline, we can now ask whether children can distinguish utterances that are appropriate as indirect requests—in this context, by expressing clear positive or negative sentiments about one of the objects—from utterances which cannot act as indirect requests.

5.2.2 Experiment 4

The goal of Experiment 4 was to test whether children are sensitive to what counts as relevant for the purpose of an indirect request.

5.2.2.1 Methods

Participants

17 children aged 4 years (4;0) to 5 years, 1 month (5;1) participated in Experiment 4 (4.0-5.12 years, mean = 4.74 years, 13 girls). Participants were recruited

from the Infant Studies Database at the University of Maryland. All participants were typically-developing monolingual English-speakers.

Design and materials

Experiment 4 included *irrelevant* statements, in addition to the *positive* and *negative* statements tested in Experiment 3. The *irrelevant* statements also contained words which were related to one of the objects, but did not express a clear positive sentiment. For example, the statement in (138) mentions water, just like the positive statement in (139), and thus is somewhat related to the toy boat. However, swimming in the water is less relevant than playing, since playing could involve the boat. The *irrelevant* statements sound as though the object reminded Froggy of a fact he wanted to share, unrelated to whether or not he wanted the object.

(138) I go swimming in the water.

(139) I like playing in the water.

For each of 15 pairs of objects, we created a *positive*, *negative*, and *irrelevant* statement, as in Table 5-2. All three of the statements contained a first person pronoun, so that the *irrelevant* statements did not seem less personal than the other two types. We did not counterbalance the target object, since there was no evidence in Experiment 3 that the

Statement Type	**Example**	**Target Object**
Positive	I like playing in the water!	boat
Irrelevant	I go swimming in the water.	sticker
Negative	I don't like playing in the water.	dump truck

Table 5-2 Experiment 4: Target sentences for sample item in main experiment.

children responded differently for different objects. The three versions of each trial were distributed across three lists.

We created 6 practice items, two in each of the conditions. The *positive* and *negative* practice items were similar to experimental items. The first *irrelevant* practice item was more irrelevant than the experimental items, in that it did not mention anything related to one of the objects. When faced with a choice between a black scarf and a purple scarf, Froggy irrelevantly utters, "Green is my favorite color." The second practice item was similar to the experimental items.

Since it is possible that children would not be able to give Froggy a sticker for infelicitous comments in general, we created a control experiment to see how children would respond to more obvious infelicity. The setup of the experiment was identical, except that all of the objects were toy food items. Half of the trials contained violations of the existence presupposition of definite reference. For example, when offered a tomato or a pineapple, Froggy said, "Give me the blue one." Since neither of the items was blue, children were to give him a sticker instead. The control experiment included 2 practice items, 3 *felicitous* items and 3 *infelicitous* items, given in Table 5-3. The items were presented in the same pseudo-random order to each participant.

Condition	**Items** (*mentioned*, unmentioned)	**Statement**
Felicitous	*banana*, egg	Give me the yellow one
	big corn, little corn	Give me the big one.
	watermelon, donut	Give me the fruit.
	donut with sprinkles, cookie	Give me the colorful one.
Infelicitous	tomato, pineapple	Give me the blue one.
	bread, donut	Give me the vegetable.
	melon, cucumber	Give me the purple one.
	square cookie, French fries	Give me the round one.

Table 5-3 Experiment 4: Target sentences for control experiment.

Procedure

The setup was nearly identical to that of Experiment 3, except the experimenter's introduction to the game included instructions on when to give Froggy a sticker instead of a toy: … *We'll give him some choices, and it will be your job to give him the one he wants. Sometimes Froggy isn't very helpful. If you can't tell which toy he wants, you can give him a sticker instead. Stickers always make him happy, right Froggy?*

The experimenters provided some feedback on the child's choice on the practice items, but on experimental items Froggy responded equally enthusiastically regardless of the child's choice.

The experimenters first presented all the items of the main experiment, then the items from the control experiment on presupposition violations. Before beginning the control experiment, the experimenter reminded the child that she could give Froggy a sticker if Froggy wasn't being helpful and she couldn't tell which food he wanted.

Data analysis

Children's responses were coded as the "related" or "unrelated" object, or a sticker. Accuracy was modeled using mixed effect logistic regression. Comparisons to a hypothetical chance-level performance were less useful in this experiment, since the three response options are not really equally probable.

5.2.2.2 Results

Control experiment

First we examined responses in the control experiment, to ensure that children were able to use the sticker to indicate infelicitous statements. The related object was considered correct for the *felicitous* condition; the sticker for the *infelicitous* condition.

Children responded with the correct object on 84.3% of the *felicitous* trials, and correctly responded with a sticker on 68.6% of the *infelicitous* trials. The model of children's accuracy included a fixed effect for STATEMENT TYPE, a random by-subject intercept, and a random by-subject slope for STATEMENT TYPE. In this model, the difference in children's accuracy based on STATEMENT TYPE was not significant ($p = 0.44$). When the random by-subject slope was removed from the model, the difference based on STATEMENT TYPE was significant ($p = 0.035$). This dramatic difference between the two models suggests that performance may have varied substantially across participants.

As shown in Figure 5-1, while most children got 2 or 3 trials correct in the *felicitous* condition, performance was more varied in the *infelicitous* condition, with three children providing no correct responses. These children may not have understood the point of the sticker response. Indeed, the same three children also failed to respond with a

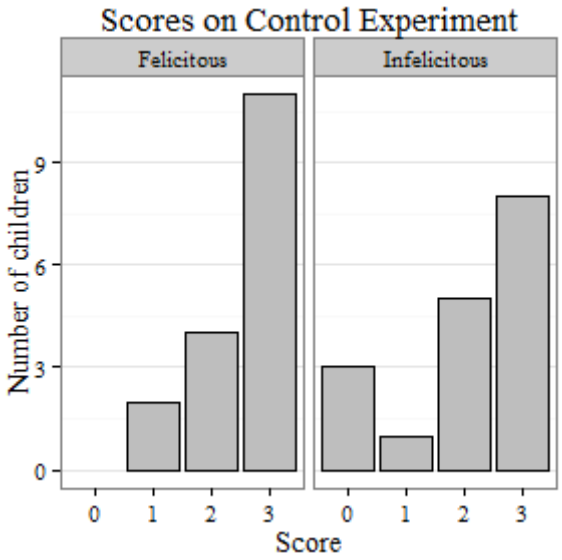


Figure 5-1 Experiment 4: Distribution of scores on control experiment.

sticker on any *irrelevant* trials in the main experiment. Data from these children were excluded from further analysis, leaving 14 remaining participants.

Main experiment

For the main experiment, accuracy rates are less informative: it is important to know which incorrect response children provided. Thus, for each condition, we modeled the probability of each response type (related object, unrelated object, or sticker) separately, using a logistic mixed effects model with a fixed effect for STATEMENT TYPE, random by-subject and by-item intercepts, and random by-subject and by-item slopes for STATEMENT TYPE. The 3-level factor STATEMENT TYPE was coded orthogonally, resulting in two contrast variables. The first contrast compared performance on *positive* statements to the mean of *irrelevant* and *negative* statements. The second contrast compared performance on *irrelevant* statements to *negative* statements.

See Table 5-4 and Figure 5-2 for summaries of the results. Children chose the related object significantly more often for *positive* statements (91%) than *irrelevant* (59%) or *negative* (9%) statements (*positive* > *negative/irrelevant*, $p << 0.0001$; *irrelevant* > *negative*, $p = 0.001$). They chose the unrelated object significantly more often for *negative* statements (67%) than for *positive* (3%) or *irrelevant* (17%) statements (*negative* > *irrelevant*, $p << 0.0001$; *negative/irrelevant* > *positive*, $p = 0.087$). They chose the sticker equally often for *irrelevant* (24%) and *negative* (24%) statements, which was significantly more often than for *positive* statements (6%) (*negative/irrelevant* > *positive*, $p = 0.0019$).

Condition	Response Type		
	Related object	Unrelated object	Sticker
Positive	0.91	0.029	0.057
Irrelevant	0.59	0.17	0.24
Negative	0.086	0.67	0.24

Table 5-4 Experiment 4: Results for main experiment.

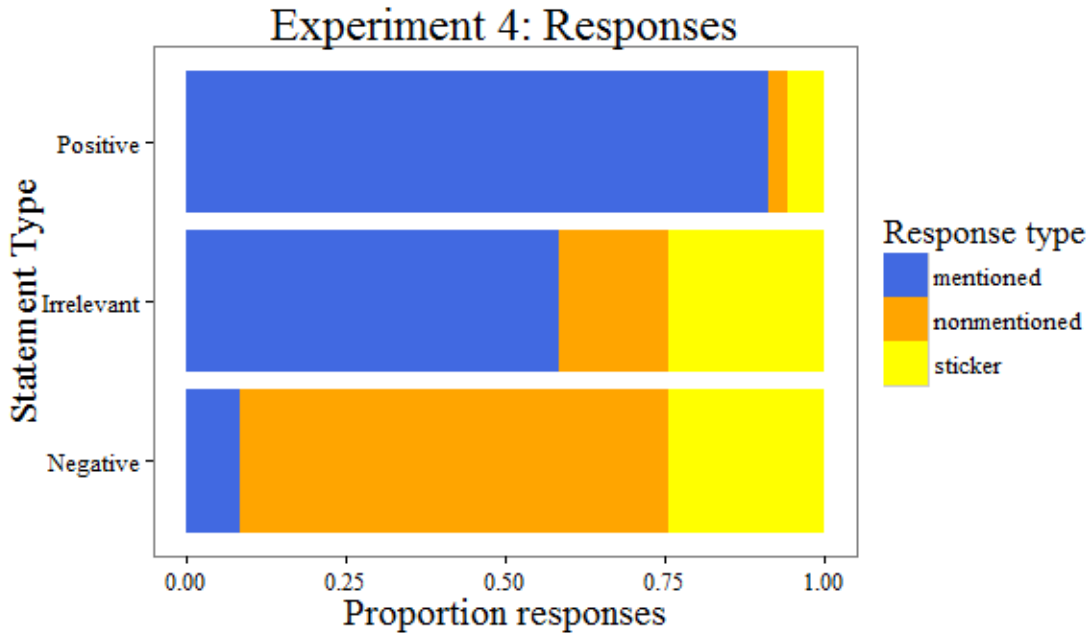


Figure 5-2 Experiment 4: Results for main experiment.

5.2.2.3 Discussion

In the critical *irrelevant* condition, the most common response was to hand over the related item. However, children did so significantly less often in this condition than in the *positive* condition, so they were sensitive to the difference in relevance between the two conditions.

In the *negative* condition, children usually chose the unrelated object. However, they also chose the sticker as often in this condition as in the *irrelevant* condition. This response is not entirely inappropriate, since a negative sentiment towards one object does

not necessarily entail a positive sentiment towards the other. That inference is licensed in this situation only because the statement follows the explicit question, "Froggy, *which one* do you want to play with?" Thus, responding with a sticker in the *irrelevant* condition reflects sensitivity to how negative statements can be used generally, but insufficient weight being given to the immediate conversational context.

The results of this experiment indicate that 4-year-olds have some awareness of what counts as relevant for an indirect request. However, since the sticker was a less salient response option than the two objects, their low rate of choosing the sticker may reflect task difficulty rather than difficulty determining what is relevant. Responding with the sticker requires a metalinguistic judgment about the adequacy of Froggy's utterance, and the whole point of using an action-based task was to avoid the need for such judgments. We remedied this problem in Experiment 5.

5.2.3 Experiment 5

The goal of Experiment 5 was to gauge children's sensitivity to the relative relevance of different statements without requiring them to explicitly reject irrelevant statements. We achieved this by having Froggy utter two statements on each trial instead of one. This method is reminiscent of the "felicity judgment task" used in previous studies (Chierchia, Crain, Guasti, Gualmini, & Meroni, 2001; Papafragou & Ozturk, 2007), but instead of having to explicitly choose between the two statements, children were free to base their response on whichever statement was most relevant.

5.2.3.1 Methods

Participants

29 children aged 3 years (3;0) to 4 years, 2 months (4;2) participated in Experiment 5 (3.04-4.19 years, mean = 3.72, 15 girls). Participants were recruited from the Infant Studies Database at the University of Maryland. All participants were typically-developing monolingual English-speakers.

Design and materials

The setup of Experiment 5 was identical to that of Experiment 3, except that Froggy uttered two statements in response to the offer of a choice between the two toys: *positive/negative*, *positive*/*irrelevant*, or *negative/irrelevant* (see Table 5-5).

The critical condition is *positive*/*irrelevant*. Each of the sentences mentions something related to one of the objects (e.g. 'dirt' and 'water'). If children cannot determine which statement is more relevant to the situation, then they will choose at random. If children can tell that the *irrelevant* statement is less relevant than the *positive* statement, they should choose the toy related to the *positive* statement more often.

We used the same set of 15 object pairs and associated statements as in Experiment 4. The different versions of each item were distributed across three lists. The order of the pair of sentences was counterbalanced.

Condition	**Example**	**Target object**
Positive/ *Negative*	I like playing with dirt. I don't like playing in the water.	truck
Positive/ *Irrelevant*	I like playing with dirt. I go swimming in the water.	truck
Negative/ *Irrelevant*	I don't like playing dirt. I go swimming in the water.	boat

Table 5-5 Experiment 5: Target sentences for sample item.

Procedure

The procedure was identical to that of Experiment 3, except that Froggy uttered two sentences on each trial instead of one.

Data analysis

Binomial tests were used to compare accuracy to chance levels. We modeled accuracy using a logistic mixed effects model with fixed effects for STATEMENT TYPE, order (whether the target object was related to the first or second statement), and their interaction, as well as random by-subject and by-item intercepts, and random by-subject and by-item slopes for STATEMENT TYPE. The 3-level factor STATEMENT TYPE was coded orthogonally, resulting in two contrast variables. The first contrast compared performance on *positive/negative* trials to *negative/irrelevant* trials. The second contrast compared performance on *positive/irrelevant* statements to the mean of the other two trial types.

5.2.3.2 Results

Although children's accuracy in the *positive*/*irrelevant* condition (67%) was significantly above chance ($p << 0.0001$), it was significantly lower than in the *positive/negative* (88%) and *negative/irrelevant* (82%) conditions ($p << 0.0001$). The difference between the *positive*/*negative* and *negative*/*irrelevant* conditions was also significant ($p = 0.004$). Although there was a numerical difference in the *positive/irrelevant* condition between trials where the *positive* statement came first (62%) and second (72%), neither the main effect for order nor any interactions were significant.

5.2.3.3 Discussion

In the critical condition, where the *positive* and *irrelevant* statements suggested conflicting response options, children were able to determine which of the two statements

to pay attention to about two-thirds of the time. This above-chance performance means that even 3-year-olds are able to process the literal content of a "hint" and interpret it with reference to the demands of the conversational context.

5.3 General discussion

We have seen that even 2-3 year-old children are able to interpret indirect requests appropriately most of the time. Their behavior across different contextual manipulations and utterance forms demonstrates that their interpretations reflect both their understanding of context and the literal form of the utterance. Although the youngest children's rate of success inferring indirect requests is certainly not as high as older children's or adults', their use of information from both the linguistic signal and the context suggests that it is possible that their interpretation procedures are qualitatively similar to those of adults.

Children's difficulty in some studies can be attributed to two main sources. First, when children are not familiar with the demands of the situation—for example, when they do not know a social "script" for the situation—they are less able to infer what the speaker wants them to do. This, of course, is not surprising. Second, children may not have enough experience to determine what level of relevance is required in any given situation. The "irrelevant" statements that they allowed in Experiments 4-5 might be considered sufficiently relevant to be used to make requests in different contexts. Even the statement "My pen is blue," from Ervin-Tripp and colleague's (1987) study, could be used to make a request—for example, in a context where there are pens of multiple colors within reach of the listener, and the speaker is obviously lacking a pen. Given the wide variety of non-conventional indirect requests that children may experience, it might take

them some time to sort out what features of any given utterance made it relevant to the situation at hand. In cases where they cannot immediately determine the connection between an utterance and the context, they err on the side of accommodating—assuming that the contribution is relevant against all odds. This tendency is actually quite similar to what adults do, although adults have the advantage of a greater store of experience and knowledge. In general when adults encounter apparently irrelevant statements—the materials from Experiments 4-5, for instance—they immediately come up with various possible interpretations that would make the statement relevant. When adults are not in a situation which presupposes that the speaker is unreliable (like an experiment where a puppet's utterances are "helpful" or "unhelpful"), they tend to accommodate utterances at all costs. Kids do too.

6 Children's interpretation of belief reports

In the last two chapters, I have argued that children are quite sophisticated pragmatically. They are able to use both the literal content of utterances and the context to estimate the speaker's intended meaning. Of course, they are also limited by their lack of knowledge about the world and underdeveloped expectations about how conversations proceed. In this chapter, I apply this view of children's pragmatic strengths and weaknesses to explain a pattern of non-adult-like behavior that has previously been attributed to quite different sources.

It has been repeatedly observed that children seem to have non-adult-like interpretations of *belief reports* like (140), where the reported belief conflicts with reality. 3-4 year-olds tend to say such sentences are false.

(140) John thinks that Baltimore is in Virginia.

These judgments seem to closely parallel young children's performance on false belief tasks. In the now-traditional change-of-location false belief task introduced by Wimmer and Perner (1983), children are asked how a character with a false belief will behave. In a representative story, Maxi is helping his mom put away groceries. He puts some chocolate in the blue cupboard before going out to the playground. While he is gone, Maxi's mom uses the chocolate to make a cake, and puts the leftovers into the green cupboard instead of back into the blue cupboard. Then Maxi returns to eat some chocolate. Children are asked the test question in (141). This version of the task is considered "non-linguistic" because verbs of belief are not used and children can point rather than responding verbally.

(141) Where will Maxi look for the chocolate?

In Wimmer and Perner's original experiment, 3-year-olds' performance was abysmally poor. They provided almost no correct responses to questions like (141), saying that Maxi would look for the chocolate in its actual location (the green cupboard) rather than where he put it (the blue cupboard). 4-5 year-olds were somewhat better: about half of the children provided at least one correct response. Even 5-6 year-olds still struggled with the task. Many additional replications and expansions of the experiment followed. Wellman, Cross and Watson (2001) performed a meta-analysis of 178 studies and found that the age trend is robust: children perform significantly below chance until about 3;5, and above chance beginning at about 4;0. Very few experimental factors have any effect on children's performance.

Thus, in both linguistic and "non-linguistic" tasks involving false belief, children seem to be unduly influenced by reality, instead of using beliefs to make their judgments. It makes sense to suppose that there is some relationship between the two tasks, and children's poor performance in each has the same cause. However, multiple underlying deficits could explain children's behavior.

Two hypotheses have been prominent in the literature. The "conceptual hypothesis" is that children are unable to understand false mental states at a conceptual level (or through the operation of some "Theory of Mind" module) until 4-5 years of age. The "syntax/semantics hypothesis" is that children have not yet acquired adult-like syntactic/semantic representations for belief reports. Both of these hypotheses have compelling evidence in their favor, but also run into serious challenges.

As an alternative to these two types of account, I propose that children's difficulties with both tasks are fundamentally pragmatic, not conceptual or syntactic/semantic. I argue that the conceptual and semantic prerequisites for belief report interpretation are in place by 3 years of age (if not earlier), and children of this age therefore understand the literal meaning of belief reports. However, children often fail to understand the role of beliefs in the discourse context, which affects their inferences about speaker meaning.

In the next section I explain the properties of belief reports in the adult language. Then in section 6.2 I review the previous evidence about children's understanding and use of belief reports. In section 6.3 I outline the conceptual hypothesis and the syntax/semantics hypothesis, and explain the challenges for each of these accounts. In section 6.4 I introduce the "pragmatic hypothesis"—that children are able to represent the literal meaning of belief reports, but have difficulty determining the intended speaker meaning in context. Finally, I report two experiments that support the pragmatic hypothesis. I conclude by discussing how the pragmatic hypothesis might handle some of the problems encountered by the conceptual and syntactic/semantic hypothesis.

6.1 Properties of belief reports

Adult-like use and interpretation of belief reports comprises multiple domains of competence, and there is no reason to suppose that children acquire all of them simultaneously. It is misleading to state the question as, "At age X, have children acquired 'think' or not?" Children need not perform exactly equivalently to adults to demonstrate some kind of competence. With this in mind, I want to outline the

components of adult-like competence before reviewing the previous evidence about children's production and comprehension of belief reports.

6.1.1 Literal meaning

Belief reports allow us to describe belief states—ours and those of others. A belief state is some kind of representation of what is true in the real world. Linguistically, we can describe the real world by enumerating the propositions that are true and false. If I want to tell you a new fact about the world, I can utter a simple declarative sentence that expresses a proposition. By uttering it, I am endorsing the proposition as true in the actual world (i.e., I'm *asserting* the proposition). If I utter (142), and I mean it literally, I'm asserting that John is working from home today. I intend you to update your model of the world with the new fact that John is working from home today.

(142) John is working from home today.

A proposition embedded in a belief report is not asserted by the speaker. If I utter (143), and I mean it literally, I'm asserting something about Mary's belief state. I intend you to update your model of the world only by changing your model of what Mary believes to be true in the world. One way of formalizing this is to distinguish "belief worlds" from the "actual world" (Hintikka, 1971). Propositions expressed by simple declarative sentences are evaluated with respect to the actual world; propositions expressed by embedded clauses in belief reports are evaluated with respect to belief worlds.

(143) Mary thinks that John is working from home today.

To understand the learning problem for belief reports, it's important to consider how they fit in to children's growing grammatical knowledge. Under a traditional view of the grammar, clausal complements are nothing special. If you can represent a clause as a grammatical unit, there's no reason not to embed that unit under a verb, just as you would with a simple noun phrase. Thus, we assume that what has to be learned are the constraints on this representational capacity. Not all verbs allow clausal complements. Even verbs that do allow them have particular requirements about the grammatical properties of the clause, as demonstrated in (144)-(149).

(144) Mary {thinks/believes/hopes/*wants/*wishes} (that) John is working from home.

(145) Mary {thinks/believes/*hopes/*wants/*wishes} that John should work from home.

(146) Mary {thinks/believes/hopes/*wants/wishes} that John would work from home.

(147) Mary {*thinks/believes/*hopes/wants/wishes} John to be working from home.

(148) Mary {*thinks/*believes/hopes/wants/wishes} to be working from home.

(149) John is working from home, I {think/believe/hope/*want/*wish}.

Although the complexity of these patterns is daunting, we assume that children must be able to use them (Gleitman, 1990; Fisher, Gleitman, & Gleitman, 1991; Gleitman, Cassidy, Nappa, Papafragou, & Trueswell, 2005). In any case, the question of how and when children acquire this syntactic knowledge will not be my focus here.

The real challenge for acquiring verbs that take sentential complements is learning their meanings. All verbs that embed sentential complements express a relation between a proposition and some individual. Usually, this relation holds between the

embedded proposition and the main clause subject, but not always. In (150), for example, the relation is between the speaker and the proposition *John is working from home*.

(150) John seems to be working from home.

Most of the variation in the meanings of these verbs is in the type of relation they express. Very broadly, the relations can involve mental attitudes (perceptions, beliefs, desires, emotions) or communications. Thus, we can characterize the problem of acquiring an attitude verb as that of determining which individual in the sentence or the context to relate to the proposition embedded in the sentence, and the nature of the relation between them.

Determining which relation is expressed by an attitude verb is even more difficult than the average case, since the relation is not observable by the usual means. You cannot see, hear, feel, smell, or taste someone believing or desiring something. At best, you can infer what they believe or desire based on their behavior. It would obviously be difficult for children to learn the different relations expressed by 'think', 'realize', 'want', and 'hope' with no further evidence to go on than hearing those words used in situations where thinking, realizing, wanting, and hoping are happening. To make matters worse, adults quite frequently use attitude verbs in situations that fail to satisfy even that criterion, as I explain in the next section.

6.1.2 Non-literal uses of belief reports

Belief reports, like any other kind of sentence, can be used to express different speaker meanings in different contexts. Consider the different implications of the same belief report in the different contexts in (151)-(152).

(151) A: Why didn't Mary invite John to the meeting?

B: She thinks he's working from home.

(152) A: Where is John? It's time to start the meeting.

B: Mary thinks he's working from home.

The exchange in (151) is about explaining Mary's behavior, so her mental states are highly relevant. B's utterance is therefore intended a straightforward report of Mary's belief, which may or may not be true in the actual world. One way to gauge the main point of an utterance is to see how it could be denied. In this situation, a third party could deny B's utterance with (153), suggesting an alternative belief to explain her behavior, but the denial in (154) would be considered irrelevant.[2]

(153) C: No, she thinks he's too busy to attend today.

(154) C: No, he's actually in the conference room.

The exchange in (152) is about John's whereabouts, so Mary's mental states are less relevant. The main point of B's utterance is to provide information about John, but it also notes that the source of the information is Mary. In this situation, a third party could felicitously deny B's utterance with (154), suggesting alternative information about John's whereabouts. By contrast, the denial in (153) is possible, but somewhat odd, since it seems to continue the commentary on Mary's beliefs instead of focusing on John's

[2] Of course, it is always possible to accommodate seemingly irrelevant utterances, as I discussed at the end of Chapter 5. Here, C's utterance in (154) could be taken to implicate that Mary couldn't think that John is working from home, since he's actually in the conference room, so there must be some other explanation for her behavior.

whereabouts. This suggests that B was offering the content of the complement clause as his main contribution to the conversation. Given that conversational participants generally try to contribute only what they know to be true (Grice's maxim of Quality), B implicitly endorses the truth of the complement clause in the actual world.

It has long been noted that certain attitude verbs can be used for conversational functions other than describing an attitude. However, much of the discussion in the literature has focused on the peculiar syntactic properties of these "parenthetical" uses. Perhaps the most striking characteristic of parentheticals is their fairly free placement in a sentence (Urmson, 1952; Ross, 1973). They seem to be similar in distribution and meaning to certain adverbials: compare (155) and (156).

(155) (I believe) John (I believe) is (I believe) working from home (I believe).

(156) (Most likely) John (most likely) is (most likely) working from home (most likely).

However, they show systematic patterns of restriction (Urmson, 1952; Bolinger, 1968; Bresnan, 1968; Ross, 1973; Hooper, 1975; Rooryck, 2001a) which could not be predicted by the hypothesis that they are simply reanalyzed as adverbials. For example, the subject of the parenthetical verb is restricted, as in (157)-(158), but different subjects can modulate the meaning, as demonstrated by the different roles of the *say* parentheticals in (159). There are also restrictions on tense and aspect which vary across parentheticals, as in (160)-(161).

(157) John is working from home, {I think / *you think / ?he thinks}.

(158) John is working from home, {you know / *I know / *he knows}.

(159) John is working from home, {I'd say / you say? / he says / they say}.

(160) John is working from home, {I think / I thought / ?I'm thinking / *I have thought / *I will think}.

(161) Jessica has tickets, {I believe / *I believed / *I am believing / *I have believed / *I will believe}.

These syntactic restrictions have prompted syntactic accounts of parenthetical interpretations in which the attitude verb sits at the head of a functional projection for evidential markers (Rooryck 2001a; 2001b) or a sentence adverbial (Bresnan 1968). Yet parenthetical uses do not require (overt) parenthetical syntax. The comparison of (151) and (152) demonstrates that sentences with standard word order can receive either mental state or parenthetical interpretations depending on what is most relevant in context. I will not be discussing further the syntactic properties of parenthetical attitudes. The important point for my purposes is that belief reports with the same surface string can be associated with different speaker meanings in different contexts. Pragmatic reasoning is therefore required to determine the speaker's intended meaning, regardless of whether there is syntactic ambiguity (Simons, 2007).

Thus, I am working under the assumption that "parenthetical" interpretations of belief reports, in which the speaker endorses the truth of the complement clause in the actual world, can be derived through purely pragmatic means. Let's walk through the example in (152), repeated below as (162), so it's clear how this works.

(162) A: Where is John? It's time to start the meeting.

B: Mary thinks he's working from home.

The literal content of B's response is a description of Mary's belief state. A comment about Mary is not, a priori, a relevant response to A's question. The proposition most closely related to the literal meaning that would provide a relevant response is the proposition expressed by the complement clause. The listener therefore infers that the speaker intends to proffer the content of the complement clause as an answer to the question. Why, then, did the speaker embed this proposition inside a belief report? Why are Mary's beliefs relevant? The listener infers that if the speaker (B) had direct evidence for the truth of the complement clause, he would have simply asserted it directly. B's source of evidence must be Mary. B is not fully committed to the truth of the proffered proposition: he will vouch for its truth only so far as he will vouch for Mary.

There are actually several inferences here: (1) the complement clause carries the relevant proposition for addressing the Question Under Discussion, (2) Mary's beliefs are relevant as a source of evidence for the proffered proposition, and (3) the speaker is only committed to the truth of the proffered proposition to the extent that Mary is knowledgeable and trustworthy.

The implication that the speaker is only partially committed to the truth of the proffered proposition actually arises in other cases that do not involve a source of evidence. For example, the sentence in (163) in response to the same question would have the implicated meaning that the speaker is less than certain that John is working from home, because otherwise he would have simply said, "He's working from home." Belief reports with 'think' can be interpreted with upper or lower bounds on certainty, depending on what alternative utterances are available. For example, (163) can be upper-bounded, on the assumption that the speaker could have said (164). In some situations,

the contrast between ‘think’ and ‘know’ can lead to the implication that a belief is false, as in (165).

(163) I think he’s working from home.

(164) I’m sure he’s working from home.

(165) I know John’s at the beach, but Mary thinks he’s working from home.

Thus, in order to reliably compute appropriate speaker meanings for belief reports, listeners need, in addition to the pragmatic principles and inference mechanisms that are common to all implicatures, to be able to (1) determine whether beliefs are relevant to the Question Under Discussion, and (2) know that ‘think’ can be considered part of a scale of certainty with other belief verbs as well as bare assertions.

How do the non-literal interpretations of belief reports affect the learning problem for belief reports? It seems likely that they will only make matters worse. Learning the relation between the subject and the complement clause is already hard enough, given its unobservability in the environment, and the non-literal interpretations add an additional relation between the speaker and the complement clause. To interpret a belief report, the listener has to not only add a proposition to their representation of the subject’s belief worlds, but also consider whether or not to add that proposition to their representation of the speaker’s belief worlds and to their representation of reality.

6.1.3 Summary: Questions about acquisition

To summarize, I will enumerate the questions we now have about children’s acquisition of belief reports, based on their properties in the adult language.

First, at what age are children able to represent the syntactic structure of a belief report, with a finite sentence embedded under the verb? When this representation becomes available for one verb, are children immediately able to apply it to other verbs? How do children parse and represent belief reports (and other cases of sentential embedding) before they are able to represent their actual structure, if there is such a stage?

Second, at what age do children understand the belief relation between the subject and the complement clause in a belief report? This could be broken down into several stages. When do children understand that 'think' refers to some kind of mental state? When do they identify that mental state as one of belief? When can they reliably evaluate whether the belief relation holds?

Finally, at what age do children reliably assign appropriate speaker meanings to belief reports? Again, this could be broken down into stages. At what age do children recognize that belief reports are used to describe belief states and to provide information about the world? When can they reliably determine whether a conversation is about beliefs or the actual world? When do they understand that belief reports participate in a scale having to do with the speaker's certainty about the truth of the complement clause?

If we could answer all of these questions, we would be satisfied with our description of the trajectory of children's acquisition of belief reports. Most of the previous literature does not tackle these questions one at a time, but rather asks the broader descriptive question, are children's interpretations adult-like, or not? However, the range of tasks that have been used do allow us to draw more specific conclusions, as we will see in the next section.

6.2 Previous evidence

Now that we have well in mind what adult-like competence with belief reports entails, we are ready to review previous studies with children. I begin with studies of children's spontaneous production of 'think', which provide a general idea of the earliest age that children are able to use 'think' conversationally. Then I turn to comprehension studies, which provide a more rigorous test of children's understanding.

6.2.1 Spontaneous production

Shatz, Wellman and Silber (1983) examined uses of mental verbs in language samples from 31 children. They used both linguistic and non-linguistic context to determine the function of each use. Most children produced at least one mental verb, and used 'think' to refer to a mental state by 2 years, 8 months (2;8). However, relatively few of their utterances unambiguously referred to mental states. Over half were instances of 'I don't know'. The rest were mainly parenthetical, functioning to direct the conversation, as in (166), or modulate the degree of certainty, as in (167).

(166) I thoughted we'd eat some cake.

(167) I think this is a lamb.

Bloom, Rispoli, Gartner and Hafitz (1989) investigated uses of 'see', 'look', 'think', and 'know' in the speech of children aged 2;0-3;2. Although 83% of children's uses of 'think' had a clausal complement—a surprisingly high proportion, given the frequency of routine phrases like 'I think so'—their uses were still restricted compared to the input. For example, almost all instances of 'think' had a first person singular subject. Like Shatz et al. (1983), they concluded that in the majority of such utterances, the main

clause verb is parenthetical. ‘Think’ and ‘know’ were usually used to modulate the degree of certainty.

Diessel & Tomasello (2001) take a more skeptical view of the structural complexity of children’s early ‘think’ utterances. They note that although children’s earliest uses of ‘think’ do occur in multi-clausal sentences, there is little evidence that the main clause ‘think’ is any more than a formulaic parenthetical to indicate the speaker’s degree of certainty. As in the examples in (168)-(170), ‘think’ always occurs in the present tense with the first-person singular ‘I’ as its subject, without any auxiliaries, adverbial modifiers, negation, or the overt complementizer ‘that’. It is only later that children begin to use ‘think’ with a wider variety of subjects, tenses, and aspects, and in interrogative sentences. The wider variety of syntactic contexts corresponds to a greater proportion of uses to describe mental states, rather than simply express the degree of certainty.

(168) I think I’m go in here. [3;1]

(169) Think some toys over here too. [3;2]

(170) It’s a crazy bone I think. [3;5]

In summary, studies of children’s spontaneous production suggest that they begin using ‘think’ in conversation quite early, before age 3. Until age 4, however, they mainly use attitude verbs for parenthetical functions, most often to express uncertainty. It is difficult to determine the underlying the structure of their parenthetical uses. They could be frozen idioms, produced without any knowledge of sentential complementation, the literal semantics of ‘think’, or its place on a scale of certainty. However, parenthetical uses of attitude verbs are very common in adult speech: Diessel and Tomasello (2001)

note that in parents' speech in the transcripts they studied, formulaic uses were several times more frequent than other uses. Thus, children's usage of 'think' for a restricted function may simply reflect the distribution in the input, rather than a lack of underlying semantic and pragmatic competence.

6.2.2 Comprehension

Johnson and Maratsos (1977) investigated 3- and 4-year-olds' comprehension of 'think' and 'know' in situations involving true and false beliefs. In a representative story, "Sally played a trick on John. While John wasn't looking, Sally took his toy duck and hid it under this box [A]. But she played a trick on John. She told him it was under this box [B], and he believed her." Children were asked a series of questions about the beliefs of Sally and John. Only children who correctly answered the question *Where will John look for the toy?* on the first or second attempt were asked the belief questions, given in (171)-(176).

(171) Does John think it's under box B? (*Yes*)

(172) Does John know it's under box A? (*No*)

(173) Does John *think* it's under box B or does he *know* it's under box B? (*think*)

(174) Does Sally know it's under box A? (*Yes*)

(175) Does Sally think it's under box B? (*No*)

(176) Does Sally *think* it's under box A or does she *know* it's under box A? (*know*)

All 4-year-olds passed this criterion and performed fairly well on the belief questions. By contrast, less than half of the 3-year-olds answered this question correctly

on the first try.[3] 3-year-olds tended to answer 'yes' to both questions about John (about 80% of the time), as though he were in states of true and false belief simultaneously. They were slightly better with the questions about Sally, correctly rejecting questions like (175) 44% of the time. The authors concluded that a "sophisticated understanding of mental verbs" is emerging in 4-year-olds, but is quite limited in 3-year-olds. It is certainly notable that even children who were able to pass the criterial false belief question (*Where will John look for the toy?*) showed such poor performance on the belief questions. However, we should be cautious about interpreting the 3-year-olds' poor performance, since it seems that they were saying "yes" to most questions. They may have found the task—a short story followed by a barrage of questions—to be a bit overwhelming.

Moore, Bryant, and Furrow (1989) investigated children's understanding of 'think' and 'know' in cases where they are used to express the speaker's degree of certainty. In a hiding game, children looked for a piece of candy in one of two locations based on the advice of two puppets. One puppet would say, *I know it's in the blue box*, the other *I think it's in the red box*. A child's response was considered correct if they looked in the box suggested by the puppet who said he *knew*. 3-year-olds were at chance, while 4-year-olds mostly chose the correct location. These results suggest that 3-year-olds' use of 'think' to express uncertainty in their own productions may not reflect a deep pragmatic understanding of how 'think' gets its uncertainty implications.

Jill de Villiers designed a test of children's ability to represent false complements that minimizes conceptual demands as much as possible (de Villiers J. G., 1995; de

[3] Since this study came before the flurry of false belief studies in the eighties, 3-year-olds' poor performance was probably somewhat surprising.

Villiers & de Villiers, 2000; de Villiers & Pyers, 2002). Children are presented with a story in which a character made a mistake, tells a lie, or has a false belief. The sentence with an attitude verb and false complement is provided directly, as in (177); children only need to remember the complement to succeed.

(177) This girl saw something funny at a tag sale and paid a dollar for it. She thought it was a toy bird but it was really a funny hat. What did she think she bought?

De Villiers & Pyers (2002) tested 3-4 year-olds three times over about 7 months. In the first round, fewer than 30% passed; by the last round, about 90% passed. These results suggest that children are not able to represent false complements until around age 4. This could reflect either a syntactic or a semantic deficit.

Sowalsky, Hacquard, and Roeper (2009) used a truth-value judgment task to test 2-5 year-olds' understanding of sentences with 'think' and 'according to', like those in the examples in (178) and (179). In some stories a character had a false belief; in others it was unknown whether the belief was true or false.

(178) Puppy thinks that it is raining outside.

(179) According to Turtle, it is snowing outside.

2-3 year-olds had difficulty with both sentence types, but were better with 'according to' (66% correct) than 'think' (35%). 4-5 year-olds were near ceiling with 'according to' (90%), but still had difficulty with 'think' (4-year-olds: 56%; 5-year-olds: 67%). All age groups were less accurate with false belief stories (52% overall) than when the reality was unspecified (66%). The relatively poor performance of 4-5 year-olds on 'think' is somewhat surprising given their success in previous studies (de Villiers J. G.,

1995; Johnson & Maratsos, 1977), suggesting that the truth-value judgment task might be a more rigorous test.

6.3 Previous accounts and problems

6.3.1 Conceptual hypothesis

Children's non-adult-like interpretations of 'think' in false belief scenarios seem to correspond to their poor performance in traditional false belief tasks. It seems reasonable to assume that understanding a concept at some non-linguistic level is a prerequisite for mapping a word to that concept. Under this assumption, children's poor performance on the non-linguistic task is primary; the non-adult-like understanding of belief reports is just an inevitable consequence of the same problem. A likely explanation of both is that children lack an understanding of false belief at the conceptual level. The change that occurs around 4 years of age, causing improvements on both tasks, is that children become able to attribute false beliefs to others.

This hypothesis encounters two main empirical challenges. The first is that although performance on linguistic and non-linguistic tests of belief understanding are closely correlated in development, it seems that children succeed earlier on the linguistic tasks. For example, in a longitudinal study with 3-4 year-old children, de Villiers and Pyers (2002) found that there were strong correlations at each time point between false belief task performance and language measures, particularly the memory for false complements task described in section 6.2.2. Furthermore, children were much more likely to pass the language task before the false belief tasks, suggesting that language is the causal factor driving their improved performance. The strong correlation between

competence with tensed complements and performance on false belief tasks and the apparent causal role of language development have been replicated in several different populations, including children with autism (Tager-Flusberg & Joseph, 2005), children with SLI (de Villiers, Burns, & Pearson, 2003), and deaf children and adults with either delayed or normal language development (de Villiers P. A., 2005). In a meta-analysis of 107 studies, Milligan and colleagues found large effects for several different kinds of language measure on false belief performance; furthermore, early language was more likely to predict later false belief performance than vice versa (Milligan, Astington, & Dack, 2007). These results are unexpected if the primary deficit is conceptual: we would have expected false belief task performance to be the driving factor, not performance on the linguistic tasks. On the other hand, this pattern is expected under the syntax/semantics hypothesis, as I discuss in the next section.

The second main empirical challenge for the conceptual hypothesis is the increasing evidence that even very young children and infants show understanding of false belief and its implications when they are tested using more "implicit" methods. Many researchers have argued that traditional false belief tasks are fundamentally flawed as measures of young children's underlying competence since they require an explicit decision and response. Clements and Perner (1994) were the first to use eye-tracking to compare children's "implicit" and "explicit" responses to a traditional change-of-location story. In a representative false belief scenario, children were presented with a scene featuring two mouse holes connected by a tunnel. One mouse, Sam, places some extra cheese in a blue box at Hole A. Then he goes back down the tunnel to go to sleep. While he is asleep, another mouse moves the cheese to a red box at Hole B. When Sam wakes

up, he is hungry and wants to go get the cheese. At this point in the story, children's gaze direction was recorded as they heard the prompt in (180). A few seconds later they were asked a direct question, as in (181).

(180) I wonder where he's going to look?

(181) Sam wants to get the cheese. Which box will he open first?

The correct response was considered looking toward Hole A after the first prompt (the "implicit" measure), and pointing to it after the question (the "explicit" measure). 2.5-year-olds were below chance on both tasks. 3-year-olds (2;11-3;2) performed much better on the implicit measure (over 75% accurate) than the explicit measure (less than 25% accurate). The two older age groups (3;3-3;7 and 3;8-4;6) also showed substantially better performance on the implicit measure than the explicit measure, although their explicit responses improved with age. Since the implicit and explicit responses in this study convey the same information, it is quite striking that children's performance on each was so different.

Although Clements and Perner (1994) did not find evidence for 2-year-olds understanding false beliefs, other researchers (as would many parents) pointed out that 2-year-olds engage in numerous behaviors that would seem to require them to attribute belief states to others. For example, they attempt to deceive, which requires an understanding that it is possible for others to have false beliefs, and that others' false beliefs might lead to behavior that would be advantageous to the child (Chandler, Fritz, & Hala, 1989). They also attempt to help people that they know to have a false belief about the location of an object (O'Neill, 1996).

Given these abilities, it is perhaps less surprising that further research eventually uncovered evidence that younger children can show understanding that a character has a false belief in a change-of-location scenario. Onishi and Baillargeon (2005) found that 15-month-olds can predict actions based on false beliefs in change-of-location stories carefully adapted for looking-time paradigms. A protagonist places an object in a particular location, then leaves the scene. In the false belief condition, the object moves to a different location while the protagonist is absent. In the final scene, the protagonist searches for the object. Infants look longer at the scene when the protagonist searches in the actual location of the object rather than where she left it, suggesting that infants are surprised when a person's actions conflict with their beliefs. This finding was subsequently replicated with 13-month-olds (Surian, Caldi, & Sperber, 2007) and 24-month-olds (Southgate, Senju, & Csibra, 2007), and extended to demonstrate that young infants have a surprisingly sophisticated understanding of false beliefs and their effects in a variety of situations (Song & Bailargeon, 2008; Song, Onishi, Baillargeon, & Fisher, 2008; Baillargeon, Scott, & He, 2010; Kovács, Téglás, & Endress, 2010).

All together, this evidence convincingly demonstrates that the development occurring in the preschool years is not the emergence of a "new" concept of belief. Rather, it must involve some more subtle aspect of children's representational or processing abilities, such that they can access their knowledge in some tasks but not in others. One possibility that has been suggested is that attributing false beliefs is a cognitively demanding task that easily breaks down under stress. Standard false belief tasks prompt children to engage in an explicit reasoning process that contrasts someone else's (false) belief with their own (true) belief. Considering both belief states

simultaneously may overwhelm children's limited processing capacity, causing them to respond based on the most salient and available representation—their own belief state. This hypothesis gains support from the fact that children's performance on false belief tasks is improved when the conflict with their own beliefs is reduced (Wellman, Cross & Watson 2001)—for example, when the object is removed from the scene in Wimmer and Perner's (1983) "Disappear" condition.

Under this weaker version of the conceptual hypothesis, the relevant developments are in general processing capacity and the ability to resolve representational conflicts. I will not speculate about which non-linguistic capacities are important: I will simply assume that whatever factors make an explicit false belief task difficult should equally affect tasks testing 'think' sentences in false belief scenarios.

To summarize, the conceptual hypothesis is that children misunderstand belief reports for the same reason that they perform poorly on false belief tasks: they have difficulty representing, tracking, or otherwise reasoning about false beliefs. The complexity of the ever-growing literature on young children's Theory of Mind makes it impossible to maintain a simplistic version of this hypothesis. Nevertheless, it should stand as a null hypothesis for children's non-adult-like interpretations of belief reports. If children systematically misunderstand utterances related to a certain concept, we should seriously consider the possibility that the problem is non-linguistic.

6.3.2 Syntax/semantics hypothesis

The syntax/semantics hypothesis takes essentially the opposite approach from the conceptual hypothesis, by proposing that the problems with linguistic belief reports are

primary, rather than the problems with "non-linguistic" false belief understanding in explicit tasks.

This account is motivated by the findings (described in the previous section) that success on tasks testing belief report understanding precedes success on traditional false belief tasks. Training studies provide more direct evidence that understanding of belief reports is causally related to success on false belief tasks. Hale and Tager-Flusberg (2003) tested preschoolers (3;0-4;10) who initially failed pretests on false belief understanding and sentential complementation. Children were trained on tasks testing false belief, sentential complements, or relative clauses, and then given post-tests in all three areas. Children in the false belief and relative clause training groups improved only on the tasks they were trained in. In contrast, children in the sentential complements training group improved on both sentential complementation and false belief understanding. Lohmann and Tomasello (2003) tested 3-year-olds in five different training conditions. Four of the training conditions involved exposure to objects with deceptive appearances, accompanied by (a) no language, (b) language with no sentential complementation, (c) language with communication verbs with sentential complements, or (d) language with mental verbs with sentential complements. The fifth training condition involved language with sentential complements, but no exposure to deceptive objects or false complements. Children who received training on false sentential complements with either communication or mental verbs outperformed all other groups. These results suggest that the ability to represent false sentential complements is a causal factor driving children's improvement on false belief tasks.

Jill de Villiers has argued that the crucial development that occurs between the ages of 3 and 4 is the acquisition of the syntactic and semantic structures necessary to represent false propositional attitudes (de Villiers & de Villiers, 2000; de Villiers & Pyers, 2002; de Villiers J. G., 2005; 2007; de Villiers & de Villiers, 2009). Sentential complements alone are evidently not enough; in the Lohmann & Tomasello training study, only training on false complements was effective. Since the truth of a sentential complement is irrelevant at the syntactic level, the important development must be semantic in nature. The critical features of false complements seem to be that they are truth-evaluable ("realis" in de Villiers' terms) and that they represent a particular perspective on the world. De Villiers suggests that children's mastery of the semantics of communication verbs—which correspond to observable events in the world—allows children to bootstrap their way into an understanding of mental verbs. These semantic structures for representing mental attitudes are then crucially involved in successful performance on a false belief task.

The syntax/semantics hypothesis also encounters two main challenges. The first is the same as one facing the conceptual hypothesis: the recent evidence that very young children and infants have some concept of false belief. If it were true, as de Villiers argues, that a semantic representation of belief is a necessary precursor to belief attribution, pre-linguistic infants should not show any understanding of false beliefs. There are two ways out of this problem. One is to doubt the interpretation of the infant findings (Perner & Ruffman, 2005). The other is to explain how linguistic representations could affect the non-linguistic attribution of mental states other than by providing access to the concept.

One explanation of the latter type, suggested by Apperly (Apperly & Butterfill, 2009; Apperly, 2011), and consistent with the work of Carruthers (2002; 2009) is that pre-linguistic "implicit" understanding of false belief reflects the output of a qualitatively different mental system than the "explicit" understanding demonstrated in traditional false belief tasks. Humans may come equipped with an innate system for attributing and processing mental states. This system would be fast and informationally-encapsulated, and thus limited in scope. It might be able to generate simple expectations about behavior—which would be reflected in eye movements—but not the explicit, reasoned decisions required in the traditional false belief task. A later developing, more general system reliant on language and executive function would be able to use the outputs of the low-level system more flexibly. Language might play a crucial role in this high-level system by representing information in a way that is interpretable by multiple mental systems (Carruthers, 2002).

Although this type of account is intriguing, it is somewhat problematic for the narrowest versions of the syntax/semantics hypothesis. Language offers other tools besides sentential complementation for describing people's attitudes, as in (182). If the role of language in false belief understanding were merely that of a facilitator in a reasoning process, we might not expect such a close connection between false belief understanding and sentential complements in particular.

(182) According to Mary, John is working from home.

The second main challenge for the syntax/semantics hypothesis is to account for the consistency observed cross-linguistically in the acquisition of attitude verbs. Children learning any language (at least those that have been tested) seem to acquire these verbs in

approximately the same order: first desire verbs, then communication verbs, and finally belief verbs. The potential problem for the syntax/semantics hypothesis is that attitude verbs in different languages impose different grammatical constraints on their complements. For example, in German, both desire and belief verbs take finite complements, as illustrated in (183)-(184). Thus, if a child has the syntactic capacity to represent a desire report, he should also be able to represent a belief report. However, Perner and colleagues (2003) showed that German-speaking 2-3 year-olds answer desire questions like (184) correctly even in situations where the desire conflicts with reality, but still have extreme difficulty with belief questions like (183) when the belief conflicts with reality.

(183) Was glaubt die Mutter, dass Andreas tut?
what believes the mother, that Andreas does
What does the mother think that Andreas is doing?

(184) Was will die Mutter, dass Andreas tut?
what wants the mother, that Andreas does
What does the mother want Andreas to do?

In Mandarin and Cantonese, clauses are not marked for finiteness, and there is some question as to whether finiteness plays any role in the grammar. Thus, both desire and belief verbs take complement clauses that are identical in their surface form. Nevertheless, Chinese-learning children begin talking about desires before they begin talking about beliefs. Tardif and Wellman (2000) found that Mandarin-learning children began using desire verbs by 1;10, if not earlier. Some Mandarin-speaking mothers report

that ‘want’ is one of their child’s first words, occurring by 11 months, on average. Belief verbs come later, and are produced much more rarely.

While the conceptual hypothesis has no problem accounting for these cross-linguistic consistencies since the relevant conceptual development would presumably unfold in the same way cross-culturally (Callaghan, et al., 2005; Liu, Wellman, Tardif, & Sabbagh, 2008), they are quite problematic for any version of the syntax/semantics hypothesis that emphasizes tensed complements or any other single syntactic property. In response to these criticisms, Jill and Peter de Villiers now assume that the important feature of false complements must be semantic, having to do with how the different perspectives on a proposition are represented (de Villiers & de Villiers, 2009). Unfortunately, the need to rely on abstract semantic features rather than observable aspects of sentence structure makes their account less satisfying. A feature like “Point of View” is uncomfortably close to the conceptual representations it’s supposed to enable.

To summarize, the syntax/semantics hypothesis is that children’s non-adult-like understanding of belief reports and poor performance on false belief tasks can both be attributed to non-adult-like syntactic or semantic representations of belief. While this hypothesis runs into some problems explaining the poor performance on the false belief task, it should—like the conceptual hypothesis—still stand as a null hypothesis for children’s poor performance on linguistic tasks. If children systematically misunderstand a certain type of sentence, we should always consider the possibility that they are representing it incorrectly at the syntactic or semantic level.

6.4 Pragmatic hypothesis

It is no doubt possible to salvage the conceptual and syntax/semantics hypotheses in the face of the evidence discussed above. However, if we take the evidence at face value, we should consider rejecting the idea that 3-year-olds lack the conceptual or syntactic/semantic apparatus to represent the literal meaning of belief reports. Once we do that, the only thing left to explain children's non-adult-like behavior is pragmatic difficulty.

I propose that 3-year-olds' non-adult-like interpretations reflect the computation of inappropriate speaker meanings, not incorrect literal meanings. Under this "pragmatic" hypothesis, children tend to assume that the status of the complement clause in reality is the main point of the utterance—as in the adult "parenthetical" use of belief reports—even in situations where adults know that the conversation is about beliefs. This faulty understanding of the discourse leads children to judge belief reports based on reality, rather than the subject's beliefs.

To make the hypothesis more concrete, let's walk through an example. Suppose we walk into a room and observe my dog frantically sniffing and pawing at an empty treat box. I might utter the belief report in (185).

(185) Molly thinks that there are still treats in there.

As an adult, you understand that the implicit Question Under Discussion I just introduced is, "Why is Molly so interested in that empty treat box?" There are various ways you could deny my statement, including the options in (186). Any denial would focus on Molly's mental state, not the status of the box in reality. We both already know

that there are actually no treats in the box, and it would be irrelevant to comment on that fact.

(186) a. No, she knows it's empty, but she can still smell the treats.

b. No, she just likes tearing up boxes.

Now suppose that I'm having this conversation with a 3-year-old child instead of an adult. When we walk into the room and I utter (185), the child has to figure out why I would have said such a thing. One possibility is that I'm commenting on Molly's erroneous belief state, but this possibility may not be particularly salient. An alternative possibility is that I don't know whether there are any treats in the box, and I'm citing Molly's behavior as a source of evidence that there are. That is, the child assumes that I've just introduced a different Question Under Discussion: "Are there any treats in the box?" Under this assumption, the child might deny my statement with (187). This is the behavior that has been observed in previous studies: denying a belief report based on reality.

(187) No there aren't! (I already gave her the last one!)

The pragmatic hypothesis I am proposing is that children often fail to recognize the relevance of belief in context, and therefore assume that the speaker meaning has to do with reality. Why do children so often fail to recognize that beliefs are relevant? It may be because they do not track people's beliefs as automatically or as quickly as adults do. Alternatively, they may not understand when people are likely to be talking about beliefs, even if they accurately track beliefs. Young children do not seem to be exposed

to much discussion about belief states, so they may not be able to guess when a belief state is sufficiently noteworthy to be the topic of a conversation.

The pragmatic hypothesis does not depend on children going through a reasoning process every time they hear a belief report and "deciding" to compute the relevance implicature that gives rise to the parenthetical reading. It may be the case that since parenthetical-type uses of belief reports—where the speaker endorses the truth of the complement clause—are so much more common in the input that children assume by default that the main point is the complement.

The pragmatic hypothesis makes two clear predictions. First, if the parenthetical reading has not become completely grammaticized for children, it should be possible to draw their attention to the relevance of belief in a particular context, and induce them to judge a belief report based on beliefs. I test this prediction in Experiment 6.

Second, if children have access to the literal meaning of belief reports, they should be able to reject belief reports that are literally false, regardless of whether they compute an inappropriate speaker meaning. I test this prediction in Experiment 7.

6.5 Experiment 6: Context sensitivity (Lewis, Hacquard, & Lidz, 2012)

The goal of Experiment 6 was to test the prediction that if the relevance of belief is made more salient in the context, children will infer the correct speaker meaning for belief reports more often and show a more adult-like pattern of judgments. I presented children with stories about hide-and-seek and asked them to judge belief reports about the seekers' beliefs. In one condition, I enhanced the relevance of belief by introducing a conflict of belief between two seekers. If children are influenced by this contextual manipulation, they should show more adult-like responses in the critical false belief

conditions. Neither the conceptual nor the syntax/semantics hypothesis predicts that children should be influenced by context, so they would not predict any improvement.

To determine how much of children's difficulty could be attributed to difficulty inhibiting their own knowledge, I included a condition in which the child was ignorant of the location of the hider. If children's difficulty in the false belief condition is due to difficulty processing the conflict with their own knowledge, they should perform better in this "ignorance" condition. By contrast, the pragmatic hypothesis does not predict any improvement in the ignorance condition, since the children could still mistake the intended speaker meaning.

6.5.1 Methods

6.5.1.1 Participants

36 children aged 3 years, 10 months (3;10) to 4 years, 5 months (4;5) participated in the study (3.8-4.4 years, mean = 3.9, 17 girls). Participants were recruited from the Center for Young Children preschool or the Infant Studies Database at the University of Maryland. All participants were typically-developing monolingual English-speakers.

6.5.1.2 Design

Children were presented with stories about hide-and-seek. After each story, a puppet uttered a target sentence containing 'think', and the child was asked to judge whether the puppet was "right" about what happened.

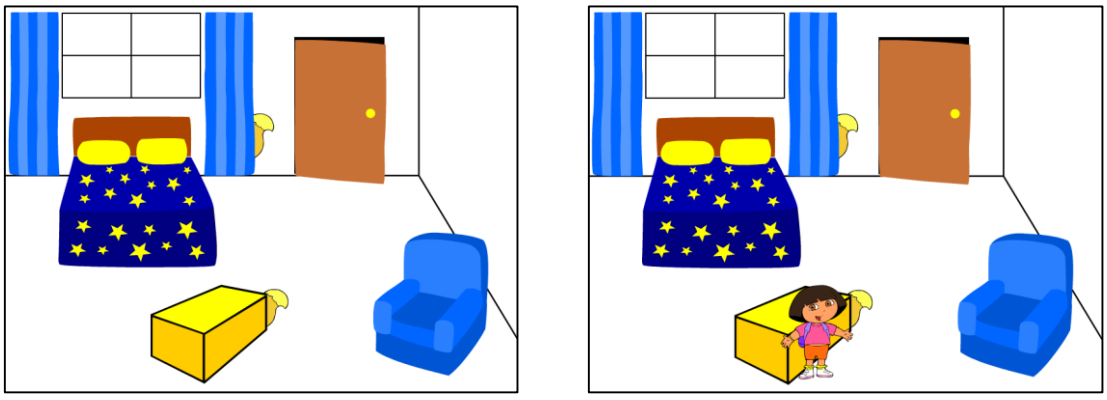
Sample story

All stories followed the same template, illustrated in the following sample story.

In the first scene, the characters (Swiper and Dora) are named and the experimenter confirms that the child can identify them. Swiper is identified as the Hider, Dora as the Seeker: *Swiper is gonna hide, and Dora will look for him. So she'll wait in the other room where she can't see.*

Dora leaves, and the child watches as Swiper hides behind the curtain. His yellow tail remains visible, protruding from behind the curtain. Then a squirrel (the Distracter) hides behind the toy box, leaving an identical yellow tail visible (Figure 6-1a). The experimenter points out the two clues to ensure that the child knows what evidence Dora will be using to guess Swiper's location.

Dora reappears to state a guess about Swiper's location based on one of the clues: *Hmm, where should I look? Oh! I see a yellow tail behind the toy box! I know--Swiper is there! I'll look for Swiper behind the toy box.* The Seeker's script is intended to establish that she is just guessing based on the first clue she noticed, but she is nevertheless



(a) (b)

Figure 6-1 **Experiment 6: Sample scenes.**
(a) Identical clues for the Hider (Swiper, behind the curtain) and Distracter (squirrel, behind the toy box). (b) The Seeker (Dora) guesses the location of the Hider.

confident—she believes what she's saying. Dora moves toward the toy box as she speaks and remains there for the rest of the story as a cue to her stated belief (Figure 6-1b).

At this point, the experimenter asks the puppet to say something about what's going on in the story. The puppet delivers a target sentence like (188). After the child responds, the puppet delivers a filler sentence. Once the child has responded to both the target sentence and the filler, the Hider and Distracter emerge from their hiding places.

(188) Dora thinks that Swiper is behind the toy box.

Manipulations

Within the stories, I manipulated whether the child had KNOWLEDGE of the Hider's true location. In the *knowledge* condition, the child watched as the Hider and Distractor hid in the scene (as in the sample story). In the *ignorance* condition, the screen was obscured during the hiding, so the child did not know which clue corresponded to the Hider until after responding to the sentences.

I also manipulated the BELIEF TYPE: whether the target sentence referred to a Seeker with a *true belief* or a *false belief*. In the sample story, the target sentences are about a seeker with a *false belief*. Note that in the *ignorance* condition, it is unknown at the point of the target sentence whether the Seeker has a true or false belief. The truth of the target sentences (i.e., the target response) was counterbalanced. Table 6-1 shows the set of possible 'think' target sentences for the sample story.

The most important manipulation for our pragmatic hypothesis was the NUMBER OF SEEKERS who looked for the hider. In the *2-seeker* stories, a second seeker guessed the other location, so that one seeker had a true belief and the other a false belief. The *2-seeker* stories were intended to heighten the relevance of belief in context by introducing

Sample sentence	Belief Type	Sent. Truth	Comp. Truth
Boots thinks that Swiper is behind the curtain.	TB	T	T
Boots thinks that Swiper is behind the toybox.	TB	F	F
Dora thinks that Swiper is behind the toybox.	FB	T	F
Dora thinks that Swiper is behind the curtain.	FB	F	T

Table 6-1 **Experiment 6: Target sentence types.**
Belief Type (TB = true belief; FB = false belief), truth of the sentence, and truth of the complement clause.

an important conflict of belief. The NUMBER OF SEEKERS was a between-subjects factor, so each child saw only *1-seeker* stories or only *2-seeker* stories.

6.5.1.3 Materials

I created 14 stories including a variety of scenes and characters. The locations of the Hider and Distracter were spread across the different hiding spots across trials, and the characters playing the Hider and Seeker rotated from story to story. I illustrated and animated the stories in Adobe Flash. I recorded narration for each story and added it to the animated videos.

I created two lists of target sentences. In each list, the order of sentences with respect to BELIEF TYPE and sentence truth was pseudo-randomized. Two filler sentences, one true and one false, were created for each story. The fillers did not involve belief. They were created using templates exemplified by (189)-(193). (190)-(191) were only appropriate in *knowledge* stories, and (193) in *ignorance*.

(189) Dora is looking for Swiper {behind the toy box/behind the curtain}.

(190) Swiper is really hiding {behind the curtain/behind the toy box}.

(191) There's really a squirrel {behind the toy box/behind the curtain}.

(192) We can see a yellow tail {behind the toy box/under the bed}.

(193) Swiper is {behind the curtain or behind the toy box/behind the door or under the bed}.

Each participant saw 2 practice trials, followed by 3 trials in each of 4 conditions (KNOWLEDGE × BELIEF TYPE). The distribution of true and false sentences was counterbalanced across conditions. Since there were an odd number of trials per condition, the distribution is only fully balanced when both lists are taken together.

6.5.1.4 Procedure

Sessions took place in a quiet room with the child seated in front of a laptop. The experimenter sat alongside the child, operating the puppet with one hand and coding responses with the other. Sessions were videotaped so that children's responses could be coded later by an independent viewer.

The experimenter began by explaining the task, introducing the puppet ("Drog", a baby dragon who wants to learn how to play hide-and-seek), and obtaining the child's assent to participate. To ensure that the child was comfortable telling the puppet whether he was right or wrong, the experimenter asked the puppet to label a few objects, and prompted the child to say whether the puppet was correct. Once the child had produced at least two *yes* and two *no* responses, the experimenter continued with the experiment.

In each trial, the child watched the animated video alongside the puppet. After the story, the puppet uttered the target sentence. The experimenter prompted the child to judge the sentence by asking, *Is Drog right?* For the two practice trials (which included only filler sentences), the experimenter provided feedback if the child responded incorrectly. The form of the feedback was flexible, but often involved pointing out relevant parts of the scene, repeating parts of the story, or modeling the correct response.

After the practice trials, the experimenter did not provide feedback. In general, the experimenter reacted to the child's response by giving feedback to Drog: *Good job, Drog—you got it right!* or *Silly Drog, you got that one mixed up!*

The filler sentence for each trial was chosen based on the child's response to the experimental sentence. If the child accepted the sentence, a false filler was chosen; if the child rejected it, a true filler was chosen.

6.5.1.5 Data analysis

Children's responses were coded online by the experimenter and again from the video recording by a different person. Responses were coded as *yes*, *no*, *I don't know*, or *unclear*. Only clear *yes* or *no* responses that were never revised were counted in accuracy rates. Video coders rejected trials in cases of experimenter error (3 out of 1420 trials), or when the child was clearly not attending or distracted (28 trials). Since most of the 4-year-old participants had fragile attention spans, coders only rejected trials in extreme cases where the child was out of her chair or talking over the story.

Accuracy rates for truth-value judgments were first analyzed separately for children in the *1-seeker* and *2-seeker* conditions, using logistic mixed effects models with fixed effects for BELIEF TYPE and KNOWLEDGE and the maximal by-subject random effects structure (random by-subject intercepts and random by-subject slopes for both main effects and the interaction). Binomial tests were used to compare accuracy to chance levels.

Accuracy in the *1-seeker* and *2-seeker* conditions was compared using a model with a fixed effect for NUMBER OF SEEKERS as well as BELIEF TYPE and KNOWLEDGE.

This model had the same by-subject random effects structure as the previous models, since the NUMBER OF SEEKERS did not vary within individual subjects.

6.5.2 Results

6.5.2.1 Filler accuracy

The fillers were designed to be easy to judge so they could be used as a criterion to exclude participants who could not understand or attend to the task. 4 participants who had accuracy rates below the predetermined cutoff of 65% were excluded from analysis. For the remaining 32 participants, filler accuracy ranged from 67% to 100% (mean = 87%, median = 90%). After exclusions, there were 16 participants each in the *1-seeker* and *2-seeker* conditions.

6.5.2.2 Accuracy on 'think' sentences

See Table 6-2 and Figure 6-2 for a summary of results.

In the *1-seeker* condition, there was a significant main effect of BELIEF TYPE (p = 0.0073): children were more accurate with *true belief* than *false belief*. There was also a significant interaction between BELIEF TYPE and KNOWLEDGE (p = 0.0038): the asymmetry based on BELIEF TYPE only held in the *knowledge* condition (as expected, since in the *ignorance* condition the belief type is in fact unknown). In *knowledge* stories, children were highly accurate in the *true belief* condition (83%: above chance, $p << 0.001$), and inaccurate in the *false belief* condition (36%: marginally below chance, p = 0.08). In *ignorance* (collapsing across BELIEF TYPE), children were just above chance (62%, p = 0.02). There was no significant difference in accuracy between *true belief* (69%) and *false belief* (56%) sentences.

Knowledge	Belief Type	*1-seeker*	*2-seeker*
knowledge	*true belief*	83%**	89%**
	false belief	36%°	52%
ignorance	*true belief*	69%*	81%**
	false belief	56%	82%**

Table 6-2 **Experiment 6: Accuracy rates by condition.**
Stars indicate that the accuracy rate was different from chance: ° p < 0.1, * p < 0.05, ** p < 0.001.

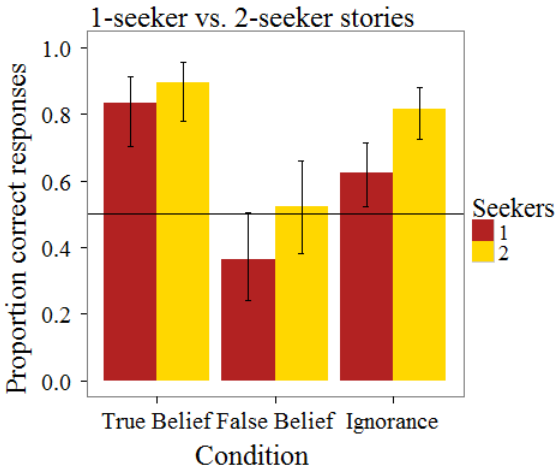


Figure 6-2 **Experiment 6: Accuracy rates by condition.**
Error bars represent 95% confidence intervals based on the binomial distribution.

In the *2-seeker* condition, there was a significant main effect of Belief Type (p = 0.023) and an interaction with Knowledge (p = 0.0015). In *knowledge* stories, children were highly accurate in the *true belief* condition (89%: above chance, $p << 0.001$), but no different from chance in the *false belief* condition (56%, p = 0.9). In *ignorance* stories, children were significantly above chance (82% overall, $p << 0.001$), with no significant difference between *true belief* (81%) and *false belief* (82%) sentences.

In the model for all the conditions together with NUMBER OF SEEKERS added as an additional fixed effect, the significant main effect of BELIEF TYPE remained (p = 0.0017), as well as the interaction between BELIEF TYPE and KNOWLEDGE (p << 0.001). There was

also a significant main effect of NUMBER OF SEEKERS ($p = 0.006$), but no interactions between NUMBER OF SEEKERS and any other factor. Thus, the overall pattern of responses was similar across the *1-seeker* and *2-seeker* stories, but children were more accurate across all conditions with the *2-seeker* stories.

6.5.3 Discussion

Children were highly accurate in the *true belief* condition, inaccurate in the *false belief* condition, and somewhere between in the *ignorance* condition. Their accuracy improved across conditions when the story involved two seekers instead of one, introducing a conflict of belief that highlighted the relevance of belief in the discourse context.

The most important finding for the pragmatic hypothesis is the improved performance in the *2-seeker* condition. I conclude that the heightened relevance of belief in the context helped children access a belief-based rather than a reality-based speaker meaning for the belief report.

Although all of the hypotheses predicted the difference in performance between the *true belief* and *false belief* conditions—that is what they were designed to do, after all—the middling performance in the *ignorance* condition is potentially informative. The intention of the *ignorance* condition was to eliminate the conflict between the characters' belief and the child's belief, which should make belief attribution easier. However, children showed lower accuracy in this condition than in the *true belief* condition. What made it more difficult? If children were attempting to evaluate the complement clause against reality—as part of a deviant literal meaning for the sentence or an inappropriate pragmatic reading—the *ignorance* condition would be confusing, because it makes that

evaluation impossible. Thus, the mild difficulty in the *ignorance* condition might provide some support for either of the linguistic hypotheses over the conceptual hypothesis.

Although the results of Experiment 6 demonstrate that children are sensitive to context when interpreting belief reports, and can provide more adult-like responses in some situations, it does not demonstrate that children compute the correct literal meaning for belief reports. We investigate this question in Experiment 7.

6.6 Experiment 7: Truth conditions for 'think'

The goal of Experiment 7 was to determine whether children are capable of rejecting a belief report based their knowledge of its literal meaning. We manipulated the literal truth of the belief report as a factor, rather than merely counterbalancing it as in our and others' previous studies. If children can evaluate belief reports based on their literal meaning, they should always reject sentences that are literally false. They may evaluate literally true sentences based on the truth of the complement clause. Thus, children's accuracy should be higher for literally false than literally true sentences.

In Experiment 7 we also tested a wider age range of children in order to investigate which conditions children show improvement in over the course of development.

6.6.1 Methods

6.6.1.1 Participants

50 children aged 3 years (3;1) to 4 years, 2 months (4;2) participated in the study (3.1-4.2 years, mean = 3.6, 26 girls). Participants were recruited from the Infant Studies Database at the University of Maryland. All participants were typically-developing

monolingual English-speakers. Data from an additional 6 children (3.1-4.0 years, mean = 3.5, 2 girls) were excluded because they could not complete the task.

6.6.1.2 Design and materials

Truth-value judgment task

As in Experiment 6, children were presented with stories about hide-and-seek, and asked to judge target sentences containing 'think'. All of the stories contained 2 seekers, to give younger children the best possible chance for adult-like responses.

To determine whether children are able to evaluate belief reports against the character's beliefs instead of reality, we manipulated the LITERAL TRUTH of the sentences as a factor (rather than simply counterbalancing it, as we did in Experiment 6). We collapsed BELIEF TYPE and KNOWLEDGE into a single BELIEF TYPE factor with 3 levels: *true belief*, *false belief*, and *unknown*.

We used the same 14 stories from Experiment 6. Rather than presenting the stories in animated videos, we illustrated each stories with a series of 8-9 still images.

We created two lists of target sentences. In each list, the order of sentences with respect to BELIEF TYPE and LITERAL TRUTH was pseudo-randomized. Two filler sentences, one true and one false, were created for each story, using the same templates as in Experiment 6. Each participant saw 2 practice trials, followed by 2 trials in each of 6 conditions (LITERAL TRUTH × BELIEF TYPE).

False belief task

In addition to the truth-value judgment task, most children also completed two trials of a standard change-of-location false belief task. The story was acted out by the experimenter using toys. The story for one of the trials was as follows:

This story is about Toby and his dad. Toby is playing with his cowboy hat and pretending to be a cowboy. He's having a great time. Then he decides to go outside and play, but he doesn't want his cowboy hat to get dirty. So he puts it under the bucket where he can find it later. [Toby leaves the scene.] *While Toby is playing outside, his dad comes in to clean up his room. He finds the cowboy hat under the bucket. He says, "Hey, this doesn't belong here! I'm going to put it in the toy box where it's supposed to be. …There. Much better."* [Toby's dad leaves the scene.]

After the story, children were asked the following series of questions.

(194) *Pre-test memory questions:*

a. Where did Toby put the cowboy hat (before he went outside)?

b. Where is the cowboy hat now?

(195) *Test question:* Toby is coming back inside, and he wants to play with his cowboy hat again. He remembers where he put it. Where is Toby going to look for the cowboy hat first?

(196) *Justification:* Why will he look there?

(197) Post-test memory question [depending on child's answer to Test question]:

a. Correct test: Where is the hat really?

b. Incorrect test: Where did Toby put the hat before he went outside?

6.6.1.3 Predictions

6.6.1.4 Procedure

Sessions took place in a quiet room. For the truth-value judgment task, the child was seated in front of an iPad. The experimenter sat alongside the child, operating the

iPad. Sessions were videotaped so that children's responses could be coded later by an independent viewer.

The experimenter began by explaining the task, introducing a little boy who appeared on the screen next to the story: *This little boy is only 2 years old, so he doesn't know how to play hide and seek. We're going to try to help him learn to play. After each story, he's going to try to say what happened in the story, and it will be your job to tell him if he's right or wrong.* After obtaining the child's assent to participate, the experimenter continued with the experiment.

In each trial, the experimenter narrated the story, swiping the screen to display each scene, then delivered the target sentence (in the voice of the little boy). The experimenter prompted the child to judge the sentence by asking, *Did the little boy get it right?* As in Experiment 6, the experimenter provided feedback for the two practice trials, but not the experimental trials. The filler sentence for each trial was chosen based on the child's response to the experimental sentence. If the child accepted the sentence, a false filler was chosen; if the child rejected it, a true filler was chosen. The experimenter recorded the child's responses using buttons on the iPad.

The false belief task always came after the truth-value judgment task. The iPad was removed, and replaced with the toys for the stories. If the child initially provided incorrect answers for the pre-test memory questions, the experimenter retold the story until the child responded correctly. No feedback was provided for the test question or post-test memory questions.

6.6.1.5 Data analysis

Children's responses were coded online by the experimenter and again from the video recording by a different person.

Responses for the truth-value judgment task were coded as *yes*, *no*, *I don't know*, or *unclear*. Only clear *yes* or *no* responses were counted in accuracy rates. Video coders excluded responses in cases of experimenter error (2 out of 1396 responses) or when the child was clearly not attending or distracted (1 response). As in Experiment 6, coders were conservative, only rejecting trials in cases of extreme and obvious inattention.

For the false belief task, all children provided correct answers on the pre-test memory questions on the first or second try. Children received a score of 1 for the trial if they provided correct answers for both the test question and the follow-up memory question. Each child was given a total FB score of 0, 1, or 2 for the number of correct trials.

Accuracy rates for truth-value judgments were analyzed using logistic mixed effects models with fixed effects for BELIEF TYPE, LITERAL TRUTH, and the subject's age, as well as a random by-subject intercept and random by-subject slopes for BELIEF TYPE, LITERAL TRUTH, and their interaction. All factors were coded orthogonally. The 3-level factor Belief Type was coded as two contrast variables: the first compared *true belief* to *false belief*; the second compared *false belief* to *unknown*.

To determine whether children's performance on the standard false belief task was predictive of their understanding of 'think', we used two different models. The first added children's FB Score as a fixed effect to the original model in place of the age effect. The second was a model of accuracy in the truth-value judgment task of the *false*

belief trials alone. The model had fixed effects for LITERAL TRUTH and FB Score, as well as a random by-subject intercept and a random by-subject slope for LITERAL TRUTH.

6.6.2 Results

6.6.2.1 Filler accuracy

The fillers were designed to be easy to judge so they could be used as a criterion to exclude participants who could not understand or attend to the task. 10 participants who had accuracy rates below the predetermined cutoff of 65% were excluded from analysis. These participants will distributed over the full age range (3.2-4.0 years, mean = 3.6, 5 girls). For the remaining 40 participants (3.1-4.2 years, mean = 3.6, 20 girls), filler accuracy ranged from 71% to 100% (mean = 86%).

6.6.2.2 Truth-value judgment task

See Table 6-3 and Figure 6-3 for summaries of the results.

There were significant main effects for both BELIEF TYPE contrasts: accuracy was higher in the *true belief* compared to the *false belief* condition ($p = 0.00083$), and higher in the *false belief* than *unknown* condition ($p = 0.011$). There was also a significant interaction between the first BELIEF TYPE contrast and LITERAL TRUTH ($p << 0.0001$): in the *true belief* condition accuracy was higher for *true* (85%) than *false* (69%) sentences, while in the *false belief* condition accuracy was higher for *false* (84%) than *true* (44%) sentences.

	Literal Truth	
Belief Type	T	F
true belief	85%**	69%*
false belief	44%	84%**
unknown	62%*	68%*

Table 6-3 **Experiment 7: Accuracy rates by condition.**
Stars indicate significant difference from chance: * $p < 0.05$, ** $p < 0.001$.

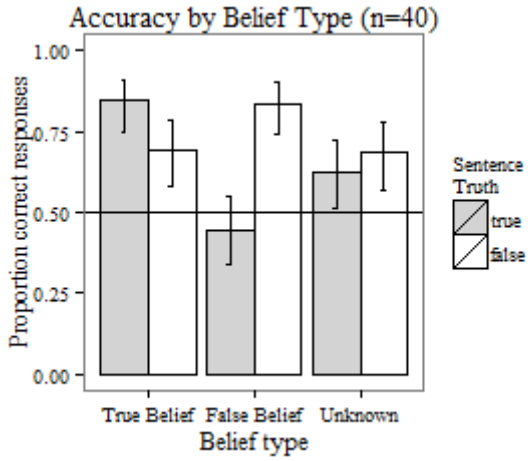


Figure 6-3 **Experiment 7: Accuracy rates by condition.**
Error bars represent 95% confidence intervals based on the binomial distribution.

There was no significant effect of the participant's age, whether it was entered as a continuous variable or as a 2-level factor based on a median split (about 3;8). Figure 6-5 shows accuracy by condition for the younger and older participants; the patterns are both qualitatively and quantitatively similar across the two groups.

Figure 6-4 Experiment 7: Results by age, median split.
Error bars represent 95% confidence intervals based on the binomial distribution.

6.6.2.3 False belief task

36 of the 40 participants in the analysis completed the false belief task. Children were grouped by their FB score (see Table 6-4). There was no significant difference in age between the three groups (one-way ANOVA, $p = 0.29$).

There were no significant effects of FB score on overall accuracy in the truth-value judgment task, or on the *false belief* trials alone. However, Figure 6-5 shows that while accuracy on the *false* sentences was similar across FB scores, there was a trend such that accuracy on *true* sentences was better for participants with perfect FB scores.

6.6.3 Discussion

Children's accuracy in the critical *false belief* condition was dramatically higher when the belief report was literally *false* compared to when it was literally *true*. This performance was remarkable given children's generally poor performance in false belief tasks, especially since even the youngest participants showed the same pattern. We

	FB Score		
	0	1	2
Number of children	11	8	17
Mean age in months (sd)	42.7 (3.7)	42.4 (2.4)	44.1 (4.1)

Table 6-4 Experiment 7: False Belief scores.

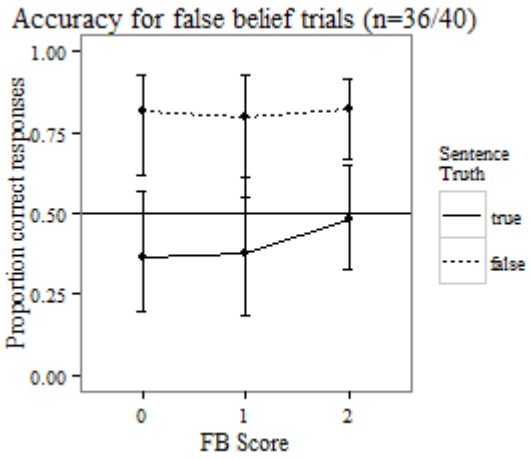


Figure 6-5 Experiment 7: Accuracy on *false belief* trials by FB score.
Error bars represent 95% confidence intervals based on the binomial distribution.

conclude that children are sensitive to the literal meaning of belief reports, and are able to evaluate whether the subject holds the stated belief by 3 years of age, if not earlier. Children's lower accuracy when the sentence is literally true suggests that children often default to a speaker meaning in which the complement clause is the main point of the utterance, and the speaker endorses its truth.

Performance on the false belief task did not predict children's performance on the truth-value judgment task. To the extent that there was a trend, it only applied to the literally *true* sentences in the *false belief* condition. It is possible that children's performance on the false belief task and their interpretation of true sentences in the *false belief* condition are both related to children's tendency to assume that the Question Under

Discussion involves reality instead of beliefs. However, it is difficult to interpret the results from the false belief task since participants completed it after having been exposed to twelve belief reports over the course of the truth-value judgment task. Although children did not receive any feedback on their responses, it is possible that this more concentrated exposure to mental state language would have affected their behavior on the false belief task. An effect of this sort might explain the surprising fact that there was no difference in age between children who failed the false belief task and those who passed it.

In this experiment as in Experiment 6, children's accuracy in the *unknown* condition (the *ignorance* condition of Experiment 6) was above chance but middling compared to accuracy in the *true belief* condition. Accuracy rate in this condition, like all the others, did not seem to change with age. As I suggested in my discussion of Experiment 6, middling accuracy in this condition might be expected in many of the children interpret the complement clause as the main point, since the truth of the complement clause in reality cannot be evaluated. However, the pragmatic hypothesis predicts that children should be able to reject literally false sentences in this condition just as in the *false belief* condition. One possible explanation is that the relevance of beliefs in the story is greater in the knowledge conditions, so children track them more carefully. Anecdotally, in the knowledge conditions children often make comments during the story about which seeker got it right or wrong. Even the shy ones will shake their heads as they observe a seeker making a wrong guess. In the ignorance condition, on the other hand, children are wholly focused on figuring out where the hider is; their spontaneous comments are mostly about their own guess about the hider's location (although we do

try to prevent participants from committing to a guess, and exclude any trials where they explicitly do so). They don't care much about what the seekers think, because they have the impression that the seeker's beliefs are just as arbitrary as their own would be. Thus, counterintuitively, removing reality from the equation reduces children's attention to the beliefs in the story, rather than enhancing it. They are less prepared to evaluate a belief report because they have not tracked the beliefs as carefully. Although it should be possible to reconstruct the beliefs by looking at the scene, since the seekers stand next to the location they guessed, some children may not go to the trouble, and simply guess.

Another result that deserves more scrutiny is that children's accuracy on *false* sentences in the *true belief* condition was slightly lower than for *true* sentences (69% compared to 85%). Even some of the oldest children in the study answered incorrectly in this condition. Although we might not be surprised in general when children have more difficult rejecting sentences than accepting them, this pattern stands in contrast to the other conditions in this study, where children's performance was as good or better for false sentences. Since this result was not predicted by any of our competing hypotheses, we can only speculate as to its source. Let's consider a sample scenario and sentence. Suppose Swiper is really hiding behind the curtain, and Boots also thinks he is (see Figure 6-6). A *false* sentence in the *true belief* condition would be (198).

(198) Boots thinks that Swiper is behind the toy box.

The sentence incorrectly attributes a false belief to Boots: Boots doesn't think that Swiper is behind the toy box, and in fact Swiper is not behind toy box. One might have expected the "double falsity" of the sentence to make it easier for children to reject: there's nothing temptingly right about it. Since the other seeker holds the opposite belief,

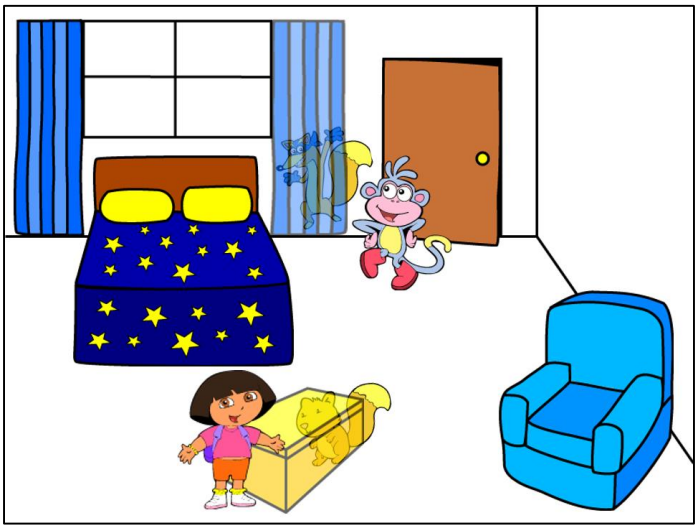


Figure 6-6 Experiment 7: Sample scene.

a contrasting true proposition is readily available regardless of whether the subject or the complement clause is taken to be in focus, as demonstrated in (199)-(200).

(199) No, DORA thinks that Swiper is behind the toybox.

(200) No, Boots thinks that Swiper is behind the CURTAIN.

What, then, makes it difficult for children to reject these sentences? Children might be taken aback by the utterance of such an obviously false sentence, leading them to second-guess their initial interpretation. My impression from seeing many children in this task is that they are not particularly concerned by the speaker's apparent obtuseness: the more obviously wrong the better, because they feel more confident in their answers.

There are two manipulations that might help determine which aspect of these sentences is difficult. The first manipulation is whether or not the belief attributed to the subject is one that someone else holds. The second manipulation is whether or not the belief attributed to the subject is one that was possible given the evidence. The sentences we tested attributed a belief that was plausible given the evidence and that someone else

in the discourse actually held. With the scene in Figure 6-6, one could also test the sentence in (201), which attributes a false belief to Swiper which no-one else holds, and which would not be plausible given the evidence. The other two combinations of these factors would require different scenes.

(201) Boots thinks that Swiper is behind the chair.

These manipulations affect whether the main clause and the complement clause express propositions that were at-issue in the story. In (198), both the belief attribution and the representation of reality in the complement clause were valid possibilities in the story. In (201), neither were. Although the standard assumption for truth-value judgment tasks is that false sentences are more felicitous—and thus easier to interpret—when they represent a valid possibility in the story, it may be that holding in mind all the possibilities can be overwhelming. Having too many ways to reject a sentence like (199) might actually make it harder.

Although the more unexpected findings from Experiment 7 warrant additional research, the main conclusions are strong. Children are capable of evaluating the literal meaning of belief reports, as evidenced by their rejection of literally false sentences even in the *false belief* condition. However, they tend to assume that the speaker meaning has to do with reality, as evidenced by their reality-based evaluation of literally true sentences.

6.7 General discussion

In this chapter I proposed that children's non-adult-like interpretations of belief reports should be attributed to a pragmatic problem, rather than a syntactic/semantic or

conceptual deficit. Children are less able than adults to determine what a speaker means by uttering a belief report because they do not always know that beliefs are under discussion in the discourse. Their default assumption is that beliefs are not relevant. This "setting" for the default could be attributed to children's general difficulty tracking beliefs or to the low frequency of conversations about belief states in their experience, or both.

An important issue that requires future research is exactly how children's linguistic experience affects their acquisition of belief reports. It is well attested that the belief language children hear has an effect on their understanding of beliefs and belief reports (Dunn, Brown, Slomkowski, Tesla, & Youngblade, 1991; Meins & Fernyhough, 1999; Ruffman, Slade, & Crowe, 2002; Peterson & Slaughter, 2003; Howard, Mayeux, & Naigles). There are several ways that this experience could have an effect, and most likely some combination of factors is at work.

First, children may learn about what mental states are and how they work by hearing their parents talk about them. While there are in principle opportunities to observe desires and beliefs affecting people's behavior without talking about them, conversations in which people's mental states are discussed explicitly would be a much more potent source of information. This factor involves children's non-linguistic ability to reason about belief states—inferring when people are in a particular belief state, and predicting what they will do because of it.

Second, children may be able to infer the potential meanings of attitude verbs from their syntactic distribution (Fisher, Gleitman, & Gleitman, 1991; Papafragou, Cassidy, & Gleitman, 2007; White, Dudley, Hacquard, & Lidz, 2012). If this is the case,

language input featuring different attitude verbs in a variety of frames would be highly informative. This factor affects children's acquisition of the literal meaning of belief reports.

Finally, children may base their expectations about the speaker meanings associated with belief reports on how they have heard them used in conversation. If most of the instances of belief reports in children's experience are "parenthetical" uses that serve to comment on reality, then children may infer that beliefs are an unlikely topic for conversation. This factor affects children's understanding of how belief reports are used pragmatically.

This final type of input effect is of most interest for us here, since we are focused on how children's pragmatic competence affects their interpretation of belief reports. There is some limited evidence that children whose mothers use 'think' most frequently in situations where they are actually endorsing the complement clause show worse performance on tests of false belief and mental verb understanding (Naigles, 2000; Howard, Mayeux, & Naigles). These findings would be consistent with the hypothesis that a greater proportion of "parenthetical" uses of 'think' in the input will strengthen children's expectation for such uses in conversation. However, there is an alternative explanation at the semantic level, which I turn to now.

One possibility that I have not yet mentioned is that the "parenthetical-like" pattern of children's judgments in our studies (and others) is generated by a semantically-encoded parenthetical reading, rather than a pragmatic-level speaker meaning. This alternative hypothesis is that children's semantic meaning for 'think' is something like *think correctly*. While this hypothesis generates very similar predictions for the pattern of

judgments across different scenarios, it cannot predict children's improved performance in the 2-seeker condition in Experiment 6. Still, it is a tempting hypothesis: if children constantly hear "parenthetical" uses of belief reports, in which the speaker endorses the truth of the complement clause, why shouldn't they conclude that 'think' means *think correctly*? However, when we consider the hypothesis in more detail, it becomes somewhat less plausible.

There are differences between the truth and felicity conditions of *think correctly* and those generated by a pragmatically-derived parenthetical interpretation, but they are extremely subtle (certainly too subtle to distinguish with the truth-value judgment tasks in Experiments 6-7). One thing to notice right away is that there aren't any verbs in English which simultaneously assert both an attitude attribution and the speaker's perspective on the truth of the content of the attitude. One of these meaning components is always presupposed. For a meaning close to *think correctly*, we have two options: 'know' (202) and 'be right' (203).

(202) Dora knows that Swiper is behind the curtain.

(203) Dora is right that Swiper is behind the curtain.

The utterance in (202) asserts that Boots believes that Swiper is behind the curtain, and presupposes that Swiper actually is behind the curtain. The utterance in (203) asserts that Swiper is behind the curtain, and presupposes that Swiper believes it. One way to diagnose the difference is to consider possible denials for the two sentences. For 'know', you can directly deny the belief claim, but not the reality claim, as in the denials of (202) in (204)-(205). The opposite pattern holds for 'be right', as shown in the denials of (203) in (206)-(207).

(204) No, she doesn't think that!

(205) a. #No he isn't!

b. Wait a minute, no he isn't!

(206) a. # No, she {doesn't think/didn't say} that!

b. Wait a minute, she {doesn't think/didn't say} that!

(207) No, he isn't.

The presupposed part of the meaning also projects out of negation. (208) still presupposes that Swiper is behind the curtain, although it's not true that Dora believes it. (209) presupposes that Dora thinks (or said) that Swiper is behind the curtain, although it's not true that he is.

(208) Dora doesn't know that Swiper is behind the curtain.

(209) Dora isn't right that Swiper is behind the curtain.

Of course it is possible to assert both parts simultaneously, but for that you need two words: 'think correctly'. The truth conditions for each of these three ways of encoding the meaning are represented in the truth table in Table 6-5, along with those of a pragmatic-level parenthetical interpretation. The "#" signifying infelicity in the presupposition-violating conditions can be thought of as the equivalent of a "Wait a

x thinks that *p*	*p*	'think'	'think correctly'	'know'	'be right'	*pragmatic parenthetical*
T	T	T	T	T	T	T
T	F	T	F	#	F	T/F
F	T	F	F	F	#	F
F	F	F	F	#	#	F

Table 6-5 Truth table for variations on 'think'.

minute!" response. The cell I have marked "T/F" under the pragmatic parenthetical is slightly different. Since we assume that a pragmatic representation of the speaker meaning goes "on top" of the literal meaning, the truth conditions of 'think' are preserved. However, in a situation where the literal meaning is satisfied but the implicated meaning—that the complement clause is true—is problematic, the response is often something like "Yes, but…", which seems subtly distinct from the "Wait a minute!" response to a presupposition violation.

Eliciting judgments of felicity is quite difficult, so we cannot easily test the hypothesis that children have one of these deviant semantic representations for 'think' by attempting to reproduce one of the truth tables from Table 6-5 in a truth-value judgment task. We can test whether children treat part of the meaning as presupposed by seeing how they interpret negated sentences. (Dudley, Orita, Moyer, Hacquard, & Lidz, 2013) have demonstrated that children do not treat 'think' like 'know'—they know that 'think' does not presuppose its complement. Children understood that in (210), the toy must be in the blue box, while in (211) it is likely to be in the red box.

(210) Lambchop doesn't know the toy is in the blue box.

(211) Lambchop doesn't think the toy is in the blue box.

It would be harder to test the alternative option, that children interpret 'think' as 'be right', presupposing the belief claim. Children would have to demonstrate understanding that (212) requires that Lambchop does think that the toy is in the blue box, while (211) requires that he doesn't. It's hard to imagine an experimental setup that would make such a judgment easy for children.

(212) Lambchop isn't right that the toy is in the blue box.

Thus, out of the four interpretation options considered, the only one we can eliminate on firm empirical grounds is 'know'. It is still logically possible that children's interpretation of 'think' is like 'be right' or *think-correctly*, instead of the pragmatic parenthetical.

One final source of evidence against these hypotheses is that children use 'think' with an uncertainty implication quite often in their own production (Shatz, Wellman, & Silber, 1983; Bloom, Rispoli, Gartner, & Hafitz, 1989; Diessel & Tomasello, 2001). This reading would be impossible if children assumed that the literal meaning of 'think' was *think correctly* or 'be right'. Indeed, children would have to be quite ignorant of pragmatic principles to assume that speakers frequently literally say, "I'm right that…" or "I think correctly that..." Since such statements would not yield any uncertainty implicatures, they only state redundantly what is already assumed: speakers say things that they believe to be true (the maxim of Quality). Although previous studies of younger children's understanding of the certainty scale in comprehension yielded lackluster performance (Moore, Bryant, & Furrow, 1989), it is possible that these studies underestimated children's understanding in the same way that the literature on scalar implicature initially underestimated children's understanding of the upper-bounded reading of 'some'. This question should be taken up again, in a study that applies the insights from the scalar implicature literature to maximize children's ability to show their understanding.

7 Conclusion

My goal in this dissertation was to investigate how listeners derive speaker meanings during comprehension. I wanted to integrate insights from the adult and developmental literatures to understand what exactly makes implicatures costly for adults and difficult for children. I also wanted to examine a diverse range of phenomena in order to gain a sense of how pragmatic enrichment works in the general case, rather than in the highly specific cases that have been heavily studied in the literature.

My main theme has been that accessing relevant information from the discourse is a critical component of pragmatic-level interpretation and a source of vulnerability for both adults and children.

7.1 Summary of findings

In Chapter 3 on adults' processing of scalar implicature in real-time comprehension, I reviewed the evidence demonstrating that adults do generate scalar implicatures in comprehension, but often at some cost. I considered three potential sources for the cost: the generation or verification of upper-bounded meanings compared to lower-bounded meanings, the process of inferring the implicated meaning, and the retrieval of relevant alternatives from the discourse context. I found evidence that all three factors play a role in making implicature-enriched interpretations costly. However, since the context varies radically across experiments, it deserves much more attention than it generally receives compared to debates about the cost of inference. Experiment 2 demonstrated that upper-bounded readings need not be costly to compute if the relevant alternative is readily available in the preceding discourse context.

In Chapter 4, I turned to children's ability to generate implicature-enriched meanings in comprehension. The review of the literature on scalar implicatures in children suggested that children understand the pressure of informativeness and are well able to compute upper-bounded interpretations based on this knowledge. Unsurprisingly, children's limited experience and world knowledge does limit their ability to infer speakers' intentions. More interestingly, they often fail to access relevant alternatives for scalar expressions when they are not made explicitly and saliently available in the immediate context.

In Chapter 5, I discussed children's understanding of indirect requests, a species of relevance implicature. I argued that indirect requests are ideal candidates for future study of relevance implicature, since their intended meanings are frequent in the input and well understood by children. The previous literature suggested that even very young children are capable of interpreting at least some forms of indirect request. Children of all ages are sensitive to both the literal meaning of the utterance and contextual factors that influence its interpretation. The limit on children's understanding seems to be the indirectness of the request, rather than non-conventionality. In my experiments I investigated a question that had not been posed in the previous literature: do children understand the limits on what can be considered relevant in a particular discourse context? I found that although children are sensitive to the fact that some statements are more relevant than others, they are still quite permissive of low-relevance statements.

Finally, in Chapter 6 I discussed children's understanding of belief reports. Children's non-adult-like interpretations of belief reports have generally been attributed to conceptual or syntactic/semantic difficulty. I argue that the problem is actually

pragmatic, and caused by a difficulty understanding what is relevant in context. Children overgenerate "parenthetical"-like readings of belief reports because they assume by default that discourses are about reality rather than about beliefs. This lack of attention to the relevance of belief in conversation may also explain children's poor performance on standard false belief tasks.

7.2 Integrating findings from adults and children

The most commonly assumed explanation for the processing costs observed in adults' real-time comprehension of scalar implicature and children's tendency to generate underinformative literal interpretations is that the pragmatic inference is costly. However, if the inference were the only or even the primary source of cost, we would not see so much variation across studies with both adults and children. I have proposed that a more important source of cost is the need to access relevant information in the discourse context.

To compute a scalar implicature, it is necessary to identify relevant alternatives to the scalar expression in the utterance. In studies of real-time comprehension, implicature-enriched interpretations are no more costly than literal interpretations when the relevant alternatives are made accessible—by including them explicitly in the discourse context, for example. Children's likelihood of generating upper-bounded interpretations increases when the alternative is explicit in the context and its relevance is clear and salient.

7.3 Integrating findings from different pragmatic phenomena

Although both the psycholinguistic and developmental literatures have focused almost exclusively on scalar implicature, I emphasize that it is important to investigate

other implicature-related phenomena alongside it. Scalar implicature is an outlier in two important ways: (1) scalar implicatures can be (or at least seem to be) conventionalized, while most implicature is highly context-dependent; (2) scalar implicatures can be local (tied to specific words or phrases), while most implicatures can only be derived by reasoning about the full utterance. It is impossible to determine whether the effects observed in adults and children are related to implicature (or perhaps Quantity-based implicature) in general, or scalar implicature in particular.

I examined two additional types of implicatures in children: indirect requests and "parenthetical" readings of belief reports. Both show some degree of conventionality, but in different ways from scalar implicature. Scalar implicatures are conventionally associated with individual words or the phrases containing them. Parenthetical readings of belief reports share this property: they are closely associated with the word 'think' (and in fact, part of the parenthetical interpretation can be attributed to a scalar implicature). By contrast, indirect requests are conventionally associated with larger constructions—questions of a particular form. On the other hand, scalar implicatures are so closely associated with some scalar expressions that the enriched meaning feels like part of the meaning of the word. This is certainly not the case for 'think'—naïve speakers tend not to notice the parenthetical reading of 'think' until it is pointed out to them. Indirect requests, however, do share this property of scalar expressions. The primary, most natural interpretation of a question like "Could you pass the salt?" is the implicature-enriched, indirect request interpretation: it takes a moment to realize that the question is literally about the listener's ability to pass the salt.

Despite these similarities, children's surface behavior with indirect requests and parenthetical belief reports is quite different from scalar implicature. In both cases, children readily compute the pragmatic enrichment. With belief reports, they in fact overgenerate the implicature-enriched reading, assuming a parenthetical interpretation when in fact the utterance was intended literally.

For indirect requests, the difference may be primarily methodological. Indirect requests have been tested in circumstances highly favorable to children's success: in action-based tasks involving scenarios that children understand well. Scalar implicature, on the other hand, was initially tested using context free truth-value judgments. Children need more contextual support than adults do to infer the experimenter's intention in using certain utterances. When that context is provided, as in some of the later more naturalistic studies on scalar implicature, children are much more likely to grasp the implicature-enriched intended meaning.

This methodological explanation is not available to explain the difference in children's treatment of scalar implicature and the parenthetical reading of 'think'. Even in very low-context tasks where the parenthetical reading is not at all relevant, such as de Villiers' memory-for-complements task, children tend to derive what we take to be an implicature-enriched meaning. Why would children assume a default literal meaning for scalar quantifiers and a default enriched meaning for 'think'?

There are several differences between the two cases that we could appeal to. The one I consider to be the most likely explanation actually flips the conventional wisdom about scalar implicatures on its head. Since scalar implicatures are supposed to be conventionalized, it is assumed that the most frequent intended meaning of certain scalar

expressions (certainly ‘some’, but others as well) is their enriched, upper-bounded meaning. However, the limited corpus work available suggests that this is actually not the case for ‘some’ (Degen, 2013, submitted). Upper-bounded readings constitute less than half of the uses of ‘some’. By contrast, the (also limited) work on the functions of ‘think’ in child-directed speech suggest that ‘think’ is overwhelmingly used with a parenthetical intended meaning. Thus, children’s “default” interpretations of scalar expressions and belief reports may actually match the input distribution much more closely than adults’ default interpretations do.

7.4 General conclusion

In this dissertation I have argued based on results from adults and children that the critical difficulty in generating implicatures in comprehension is accessing relevant contextual information. I have also demonstrated by investigating a wider range of phenomena that children’s difficulty with scalar implicature does not reflect general pragmatic incompetence. Even 3-4 year-olds generate implicatures based on a sophisticated understanding of pragmatic principles of conversation. Their abilities are limited by their lack of experience and their difficulty accessing relevant contextual information.

8 References

Ackerman, B. P. (1978). Children's understanding of speech acts in unconventional directive frames. *Child Development, 49*(2), 311-318.

Apperly, I. (2011). *Mindreaders: The Cognitive Basis of "Theory of Mind"*. New York: Psychology Press.

Apperly, I., & Butterfill, S. A. (2009). Do humans have two systems to track beliefs and belief-like states? *Psychological Review, 116*(4), 953-970.

Baayen, R. H. (2008). *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R.* Cambridge: University Press.

Bach, K. (1994). Conversational impliciture. *Mind & Language, 9*(2), 124-162.

Bach, K. (2000). Quantification, qualification and context: A reply to Stanley and Szabó. *Mind & Language, 15*(2-3), 262-283.

Bach, K. (2006). Impliciture vs. explicature: What's the difference? *Paper presented at the Workshop on Explicit Communication.* Granada.

Baillargeon, R., Scott, R. M., & He, Z. (2010). False-belief understanding in infants. *Trends in Cognitive Sciences, 14*, 110-118.

Barner, D., Brooks, N., & Bale, A. (2010). Quantity implicature and access to scalar alternatives in language acquisition. *Proceedings of SALT*, (pp. 525-543).

Bates, E. (1976). *Language and Context: The Acquisition of Pragmatics.* New York: Academic Press.

Bergen, L., & Grodner, D. (2012). Speaker knowledge influences the comprehension of pragmatic inferences. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38*(5), 1450-1460.

Bernicot, J., & Legros, S. (1987). Direct and indirect directives: What do young children understand? *Journal of Experimental Child Psychology, 43*, 346-358. doi:0022-0965/87

Bernicot, J., Laval, V., & Chaminaud, S. (2007). Nonliteral language forms in children: In what order are they acquired in pragmatics and metapragmatics? *Journal of Pragmatics, 39*, 2115-2132.

Bezuidenhout, A. L., & Morris, R. K. (2004). Implicature, relevance and default pragmatic inference. In I. A. Noveck, & D. Sperber (Eds.), *Experimental Pragmatics* (pp. 257-282). New York: Palgrave Macmillan.

Bloom, L., Rispoli, M., Gartner, B., & Hafitz, J. (1989). Acquisition of complementation. *Journal of Child Language, 16*, 101-120.

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International, 5*(9/10), 341-345.

Bolinger, D. (1968). Post-posed main phrases: An English rule for the Romance subjunctive. *Canadian Journal of Linguistics, 14*(1), 3-30.

Bott, L., & Noveck, I. A. (2004). Some utterances are underinformative: The onset and time course of scalar inferences. *Journal of Memory and Language, 51*(3), 437-457.

Bott, L., Bailey, T. M., & Grodner, D. (2012). Distinguishing speed from accuracy in scalar implicatures. *Journal of Memory and Language, 66*, 123-142.

Breheny, R. (2008). A new look at the semantics and pragmatics of numerically quantified noun phrases. *Journal of Semantics, 25*, 93-139. doi:10.1093/jos/ffm016

Breheny, R., Ferguson, H. J., & Katsos, N. (2013). Investigating the timecourse of accessing conversational implicatures during incremental sentence interpretation. *Language and Cognitive Processes, 28*(4), 443-467.

Breheny, R., Katsos, N., & Williams, J. (2006). Are generalised scalar implicatures generated by default? An on-line investigation into the role of context in generating pragmatic inferences. *Cognition, 100*, 434-463.

Bresnan, J. (1968). *Remarks on adsententials.* Manuscript, MIT.

Bucciarelli, M., Colle, L., & Bara, B. G. (2003). How children comprehend speech acts and communicative gestures. *Journal of Pragmatics, 35*, 207-241.

Callaghan, T., Rochat, P., Lillard, A., Claux, M. L., Odden, H., Itakura, S., . . . Singh, S. (2005). Synchrony in the onset of mental-state reasoning. *Psychological Science, 16*(5), 378-384. doi:10.1111/j.0956-7976.2005.01544.x

Carpenter, M., Call, J., & Tomasello, M. (2002). A new false belief test for 36-month-olds. *British Journal of Developmental Psychology, 20*, 393-420.

Carruthers, P. (2002). The cognitive functions of language. *Behavioral and Brain Sciences, 25*, 657-726.

Carruthers, P. (2009). An architecture for dual reasoning. In J. S. Evans, & K. Frankish (Eds.), *In two minds: Dual processes and beyond* (pp. 109-127). Oxford: Oxford University Press.

Carston, R. (1988). Implicature, explicature, and truth-theoretic semantics. In R. M. Kempson (Ed.), *Mental Representations: the Interface Between Language and Reality* (pp. 155-181). Cambridge University Press.

Carston, R. (2000). Explicature and semantics. In *UCL Working Papers in Linguistics* (Vol. 12, pp. 1-44).

Carston, R. (2004). Relevance Theory and the saying/implicating distinction. In L. Horn, & G. Ward (Eds.), *Handbook of Pragmatics* (pp. 633-656). Oxford: Blackwell.

Chandler, M., Fritz, A. S., & Hala, S. (1989). Small-scale deceit: Deception as a marker of two-, three-, and four-year-olds' early theories of mind. *Child Development, 60*(6), 1263-1277.

Chapman, R. S., & Kohn, L. L. (1978). Comprehension strategies in two and three year-olds: Animate agents or probable events? *Journal of Speech and Hearing Research, 21*, 746-761.

Chemla, E., & Spector, B. (2011). Experimental evidence for embedded scalar implicatures. *Journal of Semantics, 28*(3), 359-400.

Chierchia, G. (2004). Scalar implicatures, polarity phenomena and the syntax/pragmatics interface. In A. Belletti (Ed.), *Structures and beyond* (pp. 39-103). Oxford: Oxford University Press.

Chierchia, G., Crain, S., Guasti, M. T., Gualmini, A., & Meroni, L. (2001). The acquisition of disjunction: Evidence for a grammatical view of scalar implicatures. *Proceedings of the 25th Annual Boston University Conference on Language Development* (pp. 157-168). Sommerville, MA: Cascadilla Press.

Chierchia, G., Fox, D., & Spector, B. (2008). The grammatical view of scalar implicatures and the relationship between semantics and pragmatics. In C. Maienborn, K. von Heusinger, & P. Portner (Eds.), *Handbook of Semantics.* Mouton de Gruyter. Retrieved from http://semanticsarchive.net/Archive/WMzY2ZmY/CFS_EmbeddedSIs.pdf

Clements, W. A., & Perner, J. (1994). Implicit understanding of belief. *Cognitive Development, 9*, 377-395.

De Neys, W., & Schaeken, W. (2007). When people are more logical under cognitive load: Dual task impact on scalar implicature. *Experimental Psychology, 54*(2), 128-133.

de Villiers, J. (2007). The interface of language and theory of mind. *Lingua, 117*, 1858-1878.

de Villiers, J. G. (1995). Questioning minds and answering machines. In D. MacLaughlin, & S. McEwen (Eds.), *Boston University Conference on Language Development (BUCLD) 19* (pp. 20-36). Somerville, MA: Cascadilla Press.

de Villiers, J. G. (2005). Can language acquisition give children a point of view? In J. W. Astington, & J. A. Baird (Eds.), *Why language matters for theory of mind* (pp. 186-219). New York: Oxford University Press.

de Villiers, J. G., & de Villiers, P. A. (2000). Linguistic determinism and the understanding of false beliefs. In P. Mitchell, & K. J. Riggs (Eds.), *Children's Reasoning and the Mind* (pp. 191-228). Hove, U.K.: Psychology Press.

de Villiers, J. G., & de Villiers, P. A. (2009). Complements enable representation of the contents of false beliefs: the evolution of a theory of theory of mind. In S. Foster-Cohen (Ed.), *Language Acquisition.* New York: Palgrave Macmillan.

de Villiers, J. G., & Pyers, J. E. (2002). Complements to cognition: a longitudinal study of the relationship between complex syntax and false-belief understanding. *Cognitive Development, 17*, 1037-1060.

de Villiers, P. A. (2005). The role of language in Theory of Mind development: what deaf children tell us. In J. W. Astington, & J. A. Baird (Eds.), *Why Language Matters for Theory of Mind* (pp. 266-297). New York: Oxford University Press.

de Villiers, P. A., Burns, F., & Pearson, B. (2003). The role of language in the Theory of Mind development of language-impaired children: complementing theories. In B. Beachley, A. Brown, & F. Conlin (Ed.), *Proceedings of the 27th Annual Boston University Conference on Language Development* (pp. 232-242). Somerville, MA: Cascadilla Press.

de Villiers, P. A., de Villiers, J. G., Coles-White, D., & Carpenter, L. (2009). Acquisition of relevance implicatures in typically-developing children and children with autism. In J. Chandlee, M. Franchini, S. Lord, & G.-M. Reiner (Ed.), *BUCLD 33 Proceedings* (pp. 121-132). Somerville, MA: Cascadilla Press.

Degen, J. (2013, submitted). A corpus-based study of 'some' (but not 'all') implicatures. *Ms., University of Rochester*.

Degen, J., & Tanenhaus, M. K. (2011). Making inferences: The case of scalar implicature processing. In L. Carlson, C. Hoelscher, & T. F. Shipley (Ed.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 3299-3304). Austin, TX: Cognitive Science Society.

Degen, J., & Tanenhaus, M. K. (2013, under review). Naturalness of alternatives and the processing of scalar implicatures: A visual world eye-tracking study. Retrieved from http://www.bcs.rochester.edu/people/jdegen/docs/DegenTanenhaus_resubmitted_b.pdf

Diesendruck, G., & Markson, L. (2001). Children's avoidance of lexical overlap: A pragmatic account. *Developmental Psychology, 37*(5), 630-641.

Diessel, H., & Tomasello, M. (2001). The acquisition of finite complement clauses in English: A corpus-based analysis. *Cognitive Linguistics, 12*, 97-141.

Dudley, R., Orita, N., Moyer, M., Hacquard, V., & Lidz, J. (2013). Three year olds' understanding of 'know' and 'think'. *Talk at the 49th annual meeting of the Chicago Linguistics Society.*

Dunn, J., Brown, J., Slomkowski, C., Tesla, C., & Youngblade, L. (1991). Young children's understanding of other people's feelings and beliefs: Individual differences and their antecedents. *Child Development, 62*, 1352-1366.

Elrod, M. M. (1983). Young children's responses to direct and indirect directives. *The Journal of Genetic Psychology, 143*, 217-227.

Elrod, M. M. (1987). Children's understanding of indirect requests. *Journal of Genetic Psychology, 148*(1), 63-70.

Ervin-Tripp, S. M., Strage, A., Lampert, M., & Bell, N. (1987). Understanding requests. *Linguistics, 25*, 107-143.

Feeney, A., Scrafton, S., Duckworth, A., & Handley, S. (2004). The story of some: Everyday pragmatic inference by children and adults. *Canadian Journal of Experimental Psychology, 58*(2), 121-132.

Feigenson, L., Dehaene, S., & Spelke, E. (2004). Core systems of number. *Trends in Cognitive Sciences, 8*(7), 307-314. doi:10.1016/j.tics.2004.05.002

Fisher, C., Gleitman, H., & Gleitman, L. R. (1991). On the semantic content of subcategorization frames. *Cognitive Psychology, 23*, 331-392.

Garvey, C. (1975). Requests and responses in children's speech. *Journal of Child Language, 2*, 41-63.

Gazdar, G. (1979). *Implicature, Presupposition and Logical Form.* New York: Academic Press.

Geurts, B. (2006). Take "five": the meaning and use of a number word. In S. Vogeleer, & L. Tasmowski (Eds.), *Non-Definiteness and Plurality* (pp. 311-329). Amsterdam: Benjamins.

Geurts, B. (2009). Scalar implicature and local pragmatics. *Mind and Language, 24*(1), 51-79.

Geurts, B., & Pouscoulous, N. (2009). Embedded implicatures?!? *Semantics & Pragmatics, 2*(Article 4), 1-34.

Gleitman, L. R. (1990). The structural sources of verb meanings. *Language Acquisition, 1*(1), 3-55.

Gleitman, L. R., Cassidy, K., Nappa, R., Papafragou, A., & Trueswell, J. C. (2005). Hard words. *Language Learning and Development, 1*(1), 23-64.

Gordon, P. C., & Chan, D. (1995). Pronouns, passives, and discourse coherence. *Journal of Memory and Language, 34*, 216-231.

Gordon, P. C., Grosz, B. J., & Gilliom, L. A. (1993). Pronouns, names, and the centering of attention in discourse. *Cognitive Science, 17*, 311-347.

Grice, H. P. (1957). Meaning. *The Philosophical Review, 66*(3), 377-388.

Grice, H. P. (1975). Logic and Conversation. In P. Cole, & J. L. Morgan (Eds.), *Syntax and semantics 3: Speech arts* (pp. 41-58). New York: Academic Press.

Grodner, D. J., Klein, N. M., Carbary, K. M., & Tanenhaus, M. K. (2010). "Some," and possibly all, scalar inferences are not delayed: Evidence for immediate pragmatic enrichment. *Cognition, 116*, 42-55.

Gualmini, A., Crain, S., Meroni, L., Chierchia, G., & Guasti, M. T. (2001). At the semantics/pragmatics interface in child language. *Proceedings of Semantics and Linguistic Theory XI* (pp. 231-247). Ithaca, NY: CLC Publications.

Guasti, M. T., Chierchia, G., Crain, S., Foppolo, F., Gualmini, A., & Meroni, L. (2005). Why children and adults sometimes (but not always) compute implicatures. *Language and Cognitive Processes, 20*, 667-696.

Hale, C. M., & Tager-Flusberg, H. (2003). The influence of language on theory of mind: a training study. *Developmental Science, 6*, 346-359.

Hartshorne, J. K., & Snedeker, J. (submitted). The speed of inference: Evidence against rapid use of context in calculation of scalar implicatures.

Hintikka, J. (1971). Semantics for Propositional Attitudes. In L. Linsky (Ed.), *Reference and Modality* (pp. 145-167). London: Oxford University Press.

Hirschberg, J. (1985). *A Theory of Scalar Implicature.* Ph.D. dissertation, University of Pennsylvania.

Hoff-Ginsberg, E. (1986). Function and structure in maternal speech: Their relation to the child's development of syntax. *Developmental Psychology, 22*(2), 155-163.

Hooper, J. B. (1975). On assertive predicates. In J. P. Kimball (Ed.), *Syntax and Semantics* (Vol. 4, pp. 91-124). New York: Academic Press.

Horn, L. R. (1972). *On the semantic properties of logical operators in English.* Ph.D. Dissertation, University of California at Los Angeles.

Horn, L. R. (1984). Toward a new taxonomy for pragmatic inference: Q-based and R-based implicature. In D. Schiffrin (Ed.), *Meaning, Form, and Use in Context: Linguistic Applications (GURT '84)* (pp. 11-42). Washington: Georgetown University Press.

Horn, L. R. (2005). Current issues in neo-Gricean pragmatics. *Intercultural Pragmatics, 2*(2), 191-204.

Horn, L. R. (2006). More issues in neo- and post-Gricean pragmatics: A response to Robyn Carston's response. *Intercultural Pragmatics, 3*(1), 81-93.

Howard, A. A., Mayeux, L., & Naigles, L. R. (n.d.). Conversational correlates of children's acquisition of mental verbs and a theory of mind. *First Language, 28*(4), 375-402. doi:10.1177/0142723708091044

Huang, Y. T., & Snedeker, J. (2009). Semantic meaning and pragmatic interpretation in 5-year-olds: Evidence from real-time spoken language comprehension. *Developmental Psychology, 45*, 1723-1739.

Huang, Y., & Snedeker, J. (2009). Online interpretation of scalar quantifiers: Insight into the semantics-pragmatics interface. *Cognitive Psychology, 58*, 376-415.

Huang, Y., & Snedeker, J. (2011). Logic and conversation revisited: Evidence for a division between semantic and pragmatic content in real-time language comprehension. *Language and Cognitive Processes, 26*(8), 1161-1172.

Johnson, C. N., & Maratsos, M. P. (1977). Early comprehension of mental verbs: Think and know. *Child Development, 48*, 1743-1747.

Katsos, N., & Bishop, D. V. (2011). Pragmatic tolerance: Implications for the acquisition of informativeness and implicature. *Cognition, 120*, 67-81.

Katsos, N., Breheny, R., & Williams, J. (2005). Interaction of structural and contextual constraints during the on-line generation of scalar inferences. *Proceedings of GLOW 28.*

Kovács, Á. M., Téglás, E., & Endress, A. D. (2010). The social sense: Susceptibility to others' beliefs in human infants and adults. *Science, 330*, 1830-1834.

Levinson, S. C. (1983). *Pragmatics.* Cambridge: Cambridge University Press.

Levinson, S. C. (2000). *Presumptive Meanings: The Theory of Generalized Conversational Implicature.* Cambridge, MA: The MIT Press.

Lewis, S., Hacquard, V., & Lidz, J. (2012). The semantics and pragmatics of belief reports in preschoolers. *Proceedings of SALT 22*, (pp. 247-267).

Liu, D., Wellman, H. M., Tardif, T., & Sabbagh, M. A. (2008). Theory of Mind development in Chinese children: A meta-analysis of false-belief understanding across cultures and languages. *Developmental Psychology, 44*(2), 523-531.

Lohmann, H., & Tomasello, M. (2003). The role of language in the development of false belief understanding: A training study. *Child Development, 74*(4), 1130-1144.

Lytton, H., & Zwirner, W. (1975). Compliance and its controlling stimuli observed in a natural setting. *Developmental Psychology*, 769-779.

Markovits, H., & Barrouillet, P. (2002). The development of conditional reasoning: A mental model account. *Developmental Review, 22*, 5-36. doi:10.1006/drev.2000.0533

Markovits, H., & Barrouillet, P. (2004). Introduction: Why is understanding the development of reasoning important? *Thinking & Reasoning, 10*(2), 113-121.

Marr, D. (1982). *Vision.* San Francisco: W.H. Freeman.

Marty, P. P., & Chemla, E. (2013). Scalar implicatures: working memory and a comparison with 'only'. *Frontiers in Psychology, 4*(403).

Marty, P., Chemla, E., & Spector, B. (2013). Interpreting numerals and scalar items under memory load. *Lingua, 133*, 152-163.

Matsumoto, Y. (1995). The conversational condition on Horn scales. *Linguistics and Philosophy, 18*, 21-60.

Meins, E., & Fernyhough, C. (1999). Linguistic acquisitional style and mentalising development: The role of maternal mind-mindedness. *Cognitive Development, 14*, 363-380.

Milligan, K., Astington, J. W., & Dack, L. A. (2007). Language and Theory of Mind: Meta-analysis of the relation between language ability and false belief understanding. *Child Development, 78*(2), 622-646.

Moore, C., Bryant, D., & Furrow, D. (1989). Mental terms and the development of certainty. *Child Development, 60*, 167-171.

Musolino, J. (2003). The semantics and acquisition of number words: Integrating linguistic and developmental perspectives. *Cognition, 93*, 1-41.

Naigles, L. (2000). Manipulating the input: Studies in mental verb acquisition. In B. Landau, J. Sabini, J. Jonides, & E. L. Newport (Eds.), *Perception, cognition, and*

language: Essays in honor of Henry and Lila Gleitman (pp. 245-274). Cambridge, MA: MIT Press.

Neale, S. (1992). Paul Grice and the philosophy of language. *Linguistics and Philosophy, 15*(5), 509-559.

Newport, E. L., Gleitman, H., & Gleitman, L. R. (1977). Mother, I'd rather do it myself: Some effects and non-effects of maternal speech style. In C. E. Snow, & C. A. Ferguson (Eds.), *Talking to Children: Language Input and Acquisition* (pp. 109-149). Cambridge: Cambridge University Press.

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

Noveck, I. A., & Sperber, D. (Eds.). (2004). *Experimental Pragmatics.* New York: Palgrave Macmillan.

Ochs, E., & Schieffelin, B. B. (Eds.). (1979). *Developmental Pragmatics.* New York: Academic Press.

O'Neill, D. K. (1996). Two-year-old children's sensitivity to a parent's knowledge state when making requests. *Child Development, 67*(2), 659-677.

Onishi, K. H., & Baillargeon, R. (2005). Do 15-month-old infants understand false belief? *Science, 308*, 255-258.

Papafragou, A. (2006). From scalar semantics to implicature: Children's interpretation of aspectuals. *Journal of Child Language, 33*, 721-757.

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

Papafragou, A., & Ozturk, O. (2007). Children's acquisition of epistemic modality. In A. Belikova (Ed.), *Proceedings of the 2nd Conference on Generative Approaches to Language Acquisition North America (GALANA)* (pp. 320-327). Somerville, MA: Cascadilla Proceedings Project.

Papafragou, A., & Skordos, D. (to appear). Scalar implicature. In J. Lidz, W. Snyder, & J. Pater (Eds.), *The Oxford Handbook of Developmental Linguistics.* Oxford: Oxford University Press.

Papafragou, A., & Tantalou, N. (2004). Children's computation of implicatures. *Language Acquisition, 12*, 71-82.

Papafragou, A., Cassidy, K., & Gleitman, L. R. (2007). When we think about thinking: The acquisition of belief verbs. *Cognition, 105*, 125-165.

Perner, J., & Ruffman, T. (2005). Infants' insight into the mind: How Deep? *Science, 308*, 214-216.

Perner, J., Sprung, M., Zauner, P., & Haider, H. (2003). 'Want that' is understood well before 'say that', 'think that', and false belief: a test of de Villiers's linguistic determinism on German-speaking children. *Child Development, 74*, 179-188.

Peterson, C., & Slaughter, V. (2003). Opening windows into the mind: mothers' preferences for mental state explanations and children's theory of mind. *Cognitive Development, 18*, 399-429.

Politzer-Ahles, S., & Fiorentino, R. (2013). The realization of scalar inferences: Context sensitivity without processing cost. *Poster at the 26th CUNY Conference on Human Sentence Processing.*

Pouscoulous, N., Noveck, I. A., Politzer, G., & Bastide, A. (2007). A developmental investigation of processing costs in implicature production. *Language Acquisition, 14*(4), 347-375.

Pratt, M. W., Kerig, P. K., Cowan, P. A., & Cowan, C. P. (1992). Family worlds: Couple satisfaction, parenting style, and mothers' and fathers' speech to young children. *Merrill-Palmer Quarterly, 38*(2), 245-262.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin, 124*(3), 372-422.

Rooryck, J. (2001a). Evidentiality, Part I. *Glot International, 5*, 125-133.

Ross, J. R. (1973). Slifting. In M. Gross, M. Halle, & M.-P. Schützenberger (Ed.), *The formal analysis of natural languages. Proceedings of the First International Conference.* (pp. 133-169). The Hague: Mouton.

Ruffman, T., Slade, L., & Crowe, E. (2002). The relation between children's and mothers' mental state language and Theory-of-Mind understanding. *Child Development, 73*(3), 734-751.

Russell, B. (2006). Against grammatical computation of scalar implicatures. *Journal of Semantics, 23*, 361-382. doi:10.1093/jos/ffl008

Sauerland, U. (2004). Scalar implicatures in complex sentences. *Linguistics and Philosophy, 27*, 367-391.

Saul, J. M. (2002). What is said and psychological reality: Grice's project and Relevance Theorists' criticisms. *Linguistics and Philosophy*, *25*(3), 347-372.

Schaffer, H. R., & Crook, C. K. (1980). Child compliance and maternal control techniques. *Developmental Psychology, 16*(1), 54-61.

Schulze, C., Grassmann, S., & Tomasello, M. (2013). 3-year-old children make relevance inferences in indirect verbal communication. *Child Development*. doi:10.1111/cdev.12093

Searle, J. R. (1975). Indirect speech acts. In P. Cole, & J. L. Morgan (Eds.), *Syntax and Semantics, Vol. 3: Speech Acts* (pp. 59-82). New York: Academic Press.

Shatz, M. (1978). Children's comprehension of their mothers' question-directives. *Journal of Child Language, 5*, 39-46.

Shatz, M. (1983). Communication. In J. H. Flavell, & E. M. Markman (Eds.), *Handbook of Child Psychology, Vol. III: Cognitive Development* (4th ed., pp. 841-889). New York: John Wiley & Sons.

Shatz, M., Wellman, H. M., & Silber, S. (1983). The acquisition of mental verbs: A systematic investigation of the first reference to mental state. *Cognition, 14*, 301-321.

Simons, M. (2007). Observations on embedding verbs, evidentiality, and presupposition. *Lingua, 117*, 1034-1056.

Skordos, D., & Papafragou, A. (2012). Lexical alternatives improve 5-year-olds' ability to compute scalar implicatures. In A. Biller, E. Chung, & A. Kimball (Ed.), *BUCLD 36 Online Proceedings Supplement.*

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

Snedeker, J., & Trueswell, J. C. (2004). The developing constraints on parsing decisions: The role of lexical-biases and referential scenes in child and adult sentence processing. *Cognitive Psychology, 49*, 238-299.

Song, H.-j., & Bailargeon, R. (2008). Infants' reasoning about others' false perceptions. *Developmental Psychology, 44*, 1789-1795.

Song, H.-j., Onishi, K. H., Baillargeon, R., & Fisher, C. (2008). Can an agent's false belief be corrected by an appropriate communication? Psychological reasoning in 18-month-old infants. *Cognition, 109*, 295-315.

Southgate, V., Senju, A., & Csibra, G. (2007). Action anticipation through attribution of false belief by 2-year-olds. *Psychological Science, 10*, 587-592.

Sowalsky, E., Hacquard, V., & Roeper, T. (2009). Is PP opacity on the path to false belief? In J. Crawford, K. Otaki, & M. Takahashi (Eds.), *Generative Approaches to Language Acquisition North America (GALANA) 3* (pp. 263-261). Somerville, MA: Cascadilla Proceedings Project.

Spekman, N. J., & Roth, F. P. (1985). Preschool children's comprehension and production of directive forms. *Journal of Psycholinguistic Research, 14*(3), 331-349.

Sperber, D., & Wilson, D. (1986). *Relevance: Communication and cognition* (2 ed.). Oxford: Blackwell.

Strohner, H., & Nelson, K. E. (1974). The young child's development of sentence comprehension: Influence of event probability, nonverbal context, syntactic form, and strategies. *Child Development, 45*(3), 567-576.

Surian, L., Caldi, S., & Sperber, D. (2007). Attribution of beliefs by 13-month-old infants. *Psychological Science, 18*, 580-586.

Tager-Flusberg, H., & Joseph, R. M. (2005). How language facilitates the acquisition of false belief in children with autism. In J. W. Astington, & J. A. Baird (Eds.), *Why Language Matters for Theory of Mind* (pp. 298-318). New York: Oxford University Press.

Tardif, T., & Wellman, H. M. (2000). Acquisition of mental state language in Mandarin- and Cantonese-speaking children. *Developmental Psychology, 36*, 25-43.

Tomlinson, J. M., Bailey, T. M., & Bott, L. (2013). Possibly all of that and then some: Scalar implicatures are understood in two steps. *Journal of Memory and Language, 69*(1), 18-35.

Traugott, E. C. (2004). A critique of Levinson's view of Q- and M-inferences in historical pragmatics. *Journal of Historical Pragmatics, 5*(1), 1-25.

Trueswell, J. C., Sekerina, I., Hill, N. M., & Logrip, M. L. (1999). The kindergarten-path effect: studying on-line sentence processing in young children. *Cognition, 73*, 89-134.

Urmson, J. O. (1952). Parenthetical verbs. *Mind, 61*, 480-490.

Verbuk, A., & Shultz, T. (2010). Acquisition of Relevance implicatures: A case against a Rationality-baed account of conversational implicatures. *Journal of Pragmatics, 42*, 2297-2313.

Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: the truth about false belief. *Child Development, 72*, 655-684.

White, A. S., Dudley, R., Hacquard, V., & Lidz, J. (2012). Discovering classes of attitude verbs using subcategorization frame distributions. *Talk at the 43rd annual meeting of the North East Linguistic Society.* CUNY.

Wilson, D., & Sperber, D. (1986). Inference and implicature. In C. Travis (Ed.), *Meaning and Interpretation* (pp. 45-75). Oxford: Basil Blackwell.

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition, 13*, 103-128.

Zondervan, A. (2009). Experiments on QUD and focus as a contextual constraint on scalar implicature calculation. In U. Sauerland, & K. Yatsushiro (Eds.), *Semantics and Pragmatics: From Experiment to Theory* (pp. 94-110). Palgrave Macmillan.

Zondervan, A. (2011). The role of QUD and focus on the scalar implicature of most. In J. Meibauer, & M. Steinbach (Eds.), *Experimental Pragmatics/Semantics* (pp. 221-238). Amsterdam: John Benjamins Publishing Company.