Alternatives in Pragmatic Reasoning

by

Judith Degen

Submitted in Partial Fulfillment

of the

Requirements for the Degree

Doctor of Philosophy

Supervised by

Professor Michael K. Tanenhaus
Professor Christine Gunlogson

Department of Brain and Cognitive Sciences
Department of Linguistics
Arts, Sciences and Engineering
School of Arts and Sciences

University of Rochester
Rochester, New York

2013

Biographical Sketch

The author was born in Zürich, Switzerland on November 14th, 1983. She attended the University of Osnabrück from 2003 to 2008, and graduated with a Bachelor of Science in Cognitive Science in 2006 and a Master of Science in Cognitive Science (with majors in Linguistics and Philosophy) in 2009. In the fall of 2008, she began her doctoral studies in the Department of Brain and Cognitive Sciences and the Department of Linguistics at the University of Rochester. She received a Master of Arts degree in Brain and Cognitive Sciences in March 2012. She pursued her research at the University of Rochester under the direction of Professors Michael K. Tanenhaus (Brain and Cognitive Sciences) and Christine Gunlogson (Linguistics). In 2010 and 2012 she was awarded EURO-XPRAG travel grants that supported collaborations with researchers based in Europe.

Acknowledgments

There are many people who deserve thanks for making graduate school and the writing of this thesis not only possible, but interesting, challenging, and just really fun. I would like to start by expressing my sincerest gratitude to my advisors, Mike Tanenhaus and Christine Gunlogson. Their support, enthusiasm, insight, and guidance over the years have been a great source of inspiration. Working under Mike and Christine's mentorship has been an absolute pleasure. I am also greatly indebted to Florian Jaeger, who invested a tremendous amount of time and effort into developing my scientific self and teaching me how to make 3D pie charts in Excel. Greg Carlson's wisdom and inquiry indispensably helped sharpen my thinking.

The work reported in this dissertation has greatly benefitted from critical discussion with and feedback from Harald Baayen, Chris Barker, Anton Benz, Klinton Bicknell, Lewis Bott, Richard Breheny, Herb Clark, Michael Franke, Noah Goodman, Dan Grodner, Yi Ting Huang, Polly Jacobson, Gerhard Jäger, Christina Kim, Chigusa Kurumada, Dan Lassiter, Ira Noveck, Daniele Panizza, Chris Potts, Craige Roberts, Hannah Rohde, Florian Schwarz, Julie Sedivy, Jesse Snekeder, Bob van Tiel, Jack Tomlinson.

I am very grateful to Mike Tanenhaus and the EURO-XPRAG network for giving me the opportunity to travel in order to establish and foster collaborations on other projects with Richard Breheny, Michael Franke, and Gerhard Jäger, all of whom have influenced my thinking in ways which are reflected in this dissertation.

The BCS and Linguistics communities at Rochester, especially the members of the TanenhausLab, HLPlab, and the Experimental Semantics and Pragmatics reading groups provided an encouraging, supportive, intellectually challenging environment in which to develop.

I am incredibly grateful to have found such an amazing group of smart, funny, generous, weird friends-who-are/were-also-colleagues, both in the Rochester community and elsewhere, many of whom have shaped my thinking in more or less direct ways; all of whom have enriched my grad school experience: Klinton Bicknell, Esteban Buz, Jan Drugowitsch, Benjamin Allan Ellis, Thomas Farmer, Masha Fedzechkina, Alex Fine, Maureen Gillespie, Carlos Gomez Gallo, Peter Graff, Priska Herger, Caitie Hilliard, Alyssa Ibarra, Florian Jaeger, Chris Kim, Hyung-Goo Kim, Dave Kleinschmidt, Maria Klimek-Cieschinger, Chigusa Kurumada, Tal Linzen, Jens Michaelis, Dan Miner, Mario Negrello, Fermín Moscoso del Prado, Bozena Pajak, Dan Pontillo, Ting Qian, Jason Quinley, Anne Pier Salverda, Neal Snider, Ilker Yildirim.

I have been fortunate to have worked with some outstanding research assistants and independent study students, who taught me much about the value of mentoring, especially Caitie Hilliard and Laurel Raymond. Indeed, some of the work Laurel did under my supervision is reported in this dissertation.

This research would not have been possible without the support of the incredible administrative and technical staff and lab managers in the BCS and Linguistics departments: Kathy Corser, Chris Freemesser, Jen Gillis, Carla Gottschalk, Chelsea Marsh, Dana Subik, and Andrew Watts.

I would not have made it across the pond without the support of Peter Bosch and especially Graham Katz, who was not only the first person to teach me about and pique my interest in linguistics at Osnabrück, but who also had me read my very first paper on scalar implicatures and who encouraged me to apply to grad school in the US.

Even further back, without Arne Nagengast I would not have known of the existence of the Cognitive Science program in Osnabrück when I applied to college. I might have become a doctor. (Ha!)

In my second year of grad school I discovered Argentine tango. It was one of the best things that could have happened to me and often helped me through darker times. I thank all my teachers - Barbara Warren, Alden Stevens, Alex & Daniel Carcich, Richard Council, Diana Kelly, Katya Klepikova, Travis Widrick - and the lovely leaders (and followers!) of the Rochester/Western NY tango community for the time shared inside and outside the embrace.

Special thanks goes to the coffee and sandwiches at Java's Cafe and the cocktails and sushi at Banzai for keeping me fed and energized during much of the writing of this thesis; and to the symphonies of Dvorak and Beethoven for providing the soundtrack.

Anna Lührmann has been a continuous source of friendship and inspiration over the past 15 years. Despite the different paths we have taken, we have somehow always managed to find common ground. Our ski and hiking trips proved to be crucial times of reflection and dialog. Her unrelenting discipline helped me keep my eye on the ball; and when times were bad, she was an invaluable friend.

I am grateful to my family - my parents Peter and Yvonne, my brothers Sammy, David, and Simon, and my grandfather Karl - who continue to be the most important anchor point in my life. For always supporting me in my career choices; for consistently challenging my assumptions; and for raising me to value good food, travel, open-mindedness, and direct forms of conflict - these things have all turned out to be pretty useful in academia in various direct and indirect ways.

Finally, this is where people typically thank their partner. Instead, I would like to thank my roommate, reliable friend, lab mate, and fellow cohort member of five years Alex Fine, for teaching me American idioms, eating the other halves of my bananas, and generally being an unfailing source of insight, safe haven, and

pillar of support. Grad school would not have been the same without you.

Abstract

In the face of underspecified utterances, listeners routinely and without much apparent effort make the right kinds of pragmatic inferences about a speaker's intended meaning. This dissertation investigates the processing of scalar implicatures as a way of addressing how listeners perform this remarkable feat. In particular, the role of context in the processing of scalar implicatures from *some* to *not all* is explored. Contrary to the widely held assumption that scalar implicatures are highly regularized, frequent, and relatively context-independent, this dissertation suggests that they are in fact relatively infrequent and highly context-dependent; both the robustness and the speed with which scalar implicatures from *some* to *not all* are computed are modulated by the probabilistic support that the implicature receives from multiple contextual cues. Scalar implicatures are found to be especially sensitive to the naturalness or expectedness of both scalar and non-scalar alternative utterances the speaker could have produced, but didn't.

A novel contextualist account of scalar implicature processing that has roots in both constraint-based and information-theoretic accounts of language processing is proposed that provides a unifying explanation for a) the varying robustness of scalar implicatures across different contexts, b) the varying speed of scalar implicatures across different contexts, and c) the speed and efficiency of communication.

Table of Contents

List of Tables

List of Figures

Contributors and Funding Sources

I am the primary author of the entire text of this dissertation. However, Mike Tanenhaus, Christine Gunlogson, Florian Jaeger, and Laurel Raymond have collaborated with me on various aspects of this work, and have co-authored related manuscripts for publication. Overlapping research has been published or is currently under review as follows:

- Degen, J. (under review). A corpus-based study of *some* (but not *all*) implicatures.
- Degen, J. and Tanenhaus, M.K. (under review). Naturalness of alternatives and the processing of scalar implicatures: a visual world eye-tracking study.
- Degen, J. and Tanenhaus, M.K. (to appear in Cognitive Science). Scalar implicature processing: a Constraint-Based approach.
- Degen, J. and Tanenhaus, M.K. (2011). Making inferences: the case of scalar implicature processing. In L. Carlson, C. Hölscher, and T. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*, pp. 3299 - 3304.

The work reported in this dissertation was supported by NIH grant HD27206 awarded to Mike Tanenhaus and NSF CAREER grant BCS-0845059 awarded to Florian Jaeger.

1 Introduction

A remarkable feature of human language is that listeners typically have little difficulty inferring what a speaker intends to convey even though much of the relevant information is not encoded explicitly in the utterance. The question of how listeners settle on an interpretation, given an underspecified utterance, has been a topic of much debate in linguistics, philosophy, and psychology over the past half-century. The most general answer is that general principles of cooperative communication, when coupled with the context in which the utterance occurs, allow listeners to make pragmatic inferences.

Pragmatic inferences are inferences about additional, non-literal meaning conveyed by a speaker in producing a particular utterance in a particular context, a notion going back to Grice (1975), who made a distinction between *sentence meaning* and *speaker meaning*, which distinguish between what is *said* and what is *meant*. Sentences can be considered abstract objects with certain phonological, syntactic, and semantic properties. The study of these grammatical properties falls into the domain of *syntax* and *semantics*. Utterances, on the other hand, are realizations of sentences: they are concrete objects in time and space. They inherit all grammatical properties pertaining to the uttered sentence and have additional properties in virtue of having been uttered in a specific situation by a specific speaker addressing specific audience who in turn have expectations -

presumably mutually known between speaker and audience - about alternative utterances the speaker could have produced in that situation given a particular conversational goal, i.e. in virtue of occurring in a particular *context*. These properties, constituting speaker meaning, are the main object of study in *pragmatics*.

Theories of pragmatic inference, however, differ in the role that context plays: one group of theories assumes that pragmatic inference is initially a matter of context-independent default reasoning, while another group of theories highlights the important role of context in determining speaker meaning. The testing ground for these theories over the past ten years has been *scalar implicature* as in (1), where the speaker is taken to mean not only that at least one of the camp elders ate canned peaches (the *literal* or *lower-bound* interpretation of the Doc's utterance), but also that not all of them did (the *pragmatic* or *upper-bound* interpretation of his utterance).

(1) Doc Cochran: Some of the camp elders ate canned peaches.
$\rightsquigarrow$ Not all of the camp elders ate canned peaches.

Scalar implicatures have achieved this prominent position in the literature in virtue of the systematicity with which they arise; the basic schema of this type of implicature is that the use of a weaker rather than a stronger expression from a partially ordered informativeness scale is interpreted as the speaker conveying that the stronger alternative does not hold. For example, in (1) the weaker alternative with *some* is entailed by a stronger alternative that results from replacing *some* with *all*; use of the weaker *some* is consequently interpreted as conveying the negation of the stronger alternative with *all*.[1]

[1] A brief note about my use of *inference* vs. *implicature* in this dissertation is in order. As Grice noted, inferring is something listeners do, while implicating is something speakers do (albeit only in virtue of there being a listener capable of making the right kinds of inferences, see Section 1.1). This dissertation, in its focus on how listeners arrive at a presumably intended speaker meaning, is in this sense concerned with scalar *inference*. However, I will use the term interchangeably throughout. See the discussion in Section 5.1.1 for further elaboration.

Theories of scalar implicature can seek to account for either or both of the following aspects of scalar implicature: a) the *outcome* of the inference process and b) the *time course* of the inference process.[2] Questions associated with a) are, e.g.: what are the contextual factors that determine whether or not a scalar implicature will arise (or to what degree it will do so)? What are the constraints on whether an alternative expression can function as a scalemate for an expression? What is the relation between literal and non-literal meaning in the computation of implicatures? Questions associated with b) are, e.g.: is there a special status for semantic information in the processing of scalar implicatures? Is either the upper-bound or the lower-bound interpretation the default interpretation? Does contextual information constrain the interpretation of scalar items from the earliest moments of processing, or does it apply only in a second step?

Linguistic theories of scalar implicature have typically focussed on questions of type a), while processing theories have focussed on questions of type b). Different answers to these questions imply different architectures of the language system (and its interaction with cognitive, perceptual, and motor systems) and provide different views on what it means for language use to be rational and efficient. In this thesis I argue for a highly contextualist view of scalar implicatures and provide supporting empirical evidence using a variety of experimental methodologies and exploring a variety of types of contextual information.

The rest of Chapter 1 lays out the phenomenon of scalar implicature as a particular case of conversational implicature; its most important properties are discussed. In Chapter 2 I lay out in detail the two-stage vs. contextualist processing accounts of scalar implicatures proposed by others; I then propose a novel constraint-based contextualist account of scalar implicatures that heavily draws

[2]Theories of scalar implicature can also seek to explain the *acquisition* of scalar implicature, but this thesis will have nothing to say about this aspect, save for Footnote 11 in Section 5.4. For an overview of the developmental literature, see Barner, Brooks, and Bale (2011); Chierchia et al. (2001); Katsos and Bishop (2011); Noveck (2001); Noveck and Reboul (2008); Papafragou and Musolino (2003).

on the idea that the speed and robustness with which a scalar implicature is drawn depends on listeners' beliefs about alternative utterances the speaker could have produced, given the contextual constraints. Chapter 3 will be devoted to showing that both the naturalness and the availability of alternatives affect the speed and robustness of scalar implicatures from *some* to *not all* as measured in response times. Chapter 4 will do the same for eye movements, which provide a measure of interpretation that is more closely time-locked to the inference process itself, thus circumventing some of the problems of the response time measure. In Chapter 5, the frequency of scalar implicatures from *some* to *not all* in naturally occurring speech is estimated, and cues that the strength of the implicature is correlated with are explored. Finally, Chapter 6 discusses the implications of these findings for theories of scalar implicature: in particular, the evidence points toward scalar implicature as a highly context-dependent phenomenon.

1.1 Scalar implicature

Under the Gricean view, for a speaker U to mean that P is for her to have the intention that the hearer should realize that in producing the utterance, she intended him to think that P. In the words of Grice (1969, p. 151):[3]

> "U meant something by uttering x" is true iff, for some audience A,
> U uttered x intending:
> (i) A to produce a particular response r
> (ii) A to think (recognize) that U intends (i).
> (iii) A to fulfill (i) on the basis of the fulfillment of (ii).

[3]The definition provided here is the first in a long series of similar definitions that Grice develops in this article. However, because a) even his final definition doesn't capture the full range of counter-examples dealt with in the article and b) nothing in this dissertation depends on the nuance added by the substantially added complexity of the final definition, I provided the first definition here, which captures the essence of the proposal.

For example, the Doc's response in (1) can also be taken to mean that not all of the camp elders ate canned peaches. Such additionally conveyed meaning, which arises from the context of utterance, is *implicated*; what carries this meaning is an *implicature*; what is implicated is the *implicatum* (Grice, 1975).

How do hearers infer implicated content, given an utterance and a context? According to Grice, underlying the inference process is interlocutors' assumption of mutual rationality. Assuming that discourse is a cooperative activity, speaker and hearer expect each other to follow certain standards. These standards are summarized in Grice (1975)'s Cooperative Principle:

> 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. 41)

Nine more specific conversational maxims further spell out the Cooperative Principle:

Maxims of Quantity

Quantity-1. Make your contribution as informative as is required (for the current purpose of the exchange).

Quantity-2. Do not make your contribution more informative than is required.

Maxims of Quality

Supermaxim. Try to make your contribution one that is true.

Truthfulness. Do not say what you believe to be false.

Evidencedness. Do not say that for which you lack evidence.

Maxim of Relation

Be relevant.

Maxims of Manner

Supermaxim. Be perspicuous.

Obscurity Avoidance. Avoid obscure expressions.

Ambiguity Avoidance. Avoid ambiguity.

Brevity. Be brief (avoid unnecessary prolixity).

Orderliness. Be orderly.

These maxims guide the process of implicature computation. Consider example (1) again. The standard Gricean derivation of the implicature that not all of the camp elders ate canned peaches is the following: Quantity-1 requires that the speaker make the most informative statement possible, given the level of informativeness set by the discourse. The more informative, and equally relevant, statement the Doc could have made is *All of the camp elders ate canned peaches.* In choosing the weaker alternative, he was flouting Quantity-1. Assuming that he is still generally following the Cooperative Principle, he must not believe the stronger alternative to hold (Truthfulness trumps Quantity-1). Thus, Jane is justified in inferring that the Doc indeed intended her to believe that only some, but not all, of the camp elders ate canned peaches.

These implicatures, derived as the result of a clash between Quantity-1 and the Truthfulness maxim, are called *scalar implicatures.* They are scalar because the implicature trigger - *some* in our example - is the weaker expression on a partially ordered scale consisting of a stronger and a weaker element (Horn, 1972; Gazdar, 1979; Hirschberg, 1985; Matsumoto, 1995). In our example, the scale involved is $\langle \text{all}, \text{some} \rangle$, where *all* is the stronger element: the sentence *All of the camp elders ate canned peaches* entails the sentence *Some of the camp elders ate canned peaches.* The ordering relations that may hold between scale elements are manifold: entailment, as in example (1); ranking of entities, states, and attributes; whole/part relationships; type/subtype, instance-of, and generalization/specialization relations (Carston, 1998; Hirschberg, 1985), among many others. The question of what kinds of expressions can serve as alternatives is discussed further in Section 1.3 and plays an important part in one of the main claims of this thesis: that scalar implicatures are highly context-dependent.

A class of implicatures that is intimately related to scalar implicatures is based on a clash between Quantity-1 and the maxim of Evidencedness. In this case, the implicature that arises is not that the stronger statement does not hold, but rather only the weaker inference is licensed that the speaker is not in a position to assert truth or falsity of the stronger alternative. In our example, such an implicature would arise if, for example, the Doc didn't know for each camp elder whether or not they ate canned peaches, and Jane knows this. In this case, the Doc's use of the weaker alternative can be attributed to his lack of evidence for the stronger alternative rather than him knowing that the stronger alternative is not the case, leading to an inference on Jane's end that the Doc does not know whether or not all the camp elders in fact ate canned peaches. Evidencedness-based implicatures of this sort have been called *ignorance implicatures*. The assumption that the speaker is trying to avoid violating Truthfulness rather than Evidencedness amounts to considering the speaker well-informed as to the truth of the stronger alternative, and thus has been termed the *epistemic step* or the *opinionatedness assumption* (Breheny, Ferguson, & Katsos, 2013; Franke, 2009; Grice, 1975; Horn, 1972; Russell, 2006; Sauerland, 2004; van Rooij & Schulz, 2004). For the most part, ignorance implicatures will not be discussed in this thesis, but I will return to them briefly in Chapter 5.

Scalar implicatures exhibit a number of distinctive properties: *calculability, nonconventionality, cancelability, nondetachability, reinforceability*, and *universality* (Grice, 1975; Levinson, 2000). I will briefly review each of these in turn and point out where they are relevant to the theoretical and empirical contribution of the thesis.

1.1.1 Calculability

Conversational implicatures are calculable via the conversational maxims. In particular, scalar implicatures are calculable via the conversational maxims of

Quantity-1, Relevance, and Truthfulness (or Evidencedness). What is required is that the implicature trigger be the weaker element of a partially ordered scale. Above, I provided an example of the way scalar implicature calculation proceeds in a Gricean framework.

A question pertaining to the calculability property is the cognitive status that this property is assumed to have. Grice himself did not intend his theory of conversational implicature to be a processing theory. That is, the application of the maxims in a particular sequence should not be taken to describe a (conscious or unconscious) cognitive process. Rather, Grice's theory is best thought of as inhabiting Marr (1982)'s computational level. Marr distinguishes three levels at which a cognitive phenomenon can be analyzed. The highest is the *computational* level, which specifies the task the individual needs to perform, e.g. pragmatic inference. The lowest is the *implementational* level, at which an implementation of the function in the system's hardware is described. In between lies the *algorithmic* level, where suitable algorithms for implementing the function described at the computational level are explored that are simultaneously compatible with the hardware constraints at the implementational level. Most processing theories function at the algorithmic level (see Chapter 2).

1.1.2 Nonconventionality

Grice made a distinction between conventional and conversational implicatures. In contrast to conversational implicatures, conventional implicatures arise as a word's or expression's agreed-upon, conventional meaning. An example is the word *but*, which does more than simply contribute to sentence meaning. Even though for (2a) to be true nothing else is required than that (2b) be true, there is a difference between the two.

(2) a. She is poor *but* honest.

b. She is poor *and* honest.

The difference between (2a) and (2b) is that *but* indicates a contrast between being poor and being honest. This contrast is not part of what is said (i.e. the truth conditions are the same for both sentences), nor is it entailed by what is said. It is not a conversational implicature either, because it does not have to be worked out on the basis of conversational maxims. Rather, it depends solely on the conventional meaning of the word *but* - it is a conventional implicature. The boundary between conventional implicature, semantic presupposition, and other related phenomena has proven difficult to delineate clearly and different authors have slightly different notions of what constitutes a conventional implicature (Bach, 1999; Grice, 1975; Karttunen & Peters, 1979; Potts, 2005).

Conversational implicatures on the other hand arise from "general features of discourse" (Grice, 1975, p. 45): the interaction of a) what is said (i.e. the conventional meaning of the utterance), b) the conversational maxims, c) the linguistic and extra-linguistic context, d) world knowledge, and e) the interlocutors' assumption that a) - d) are in the interlocutors' common ground.

1.1.3 Cancelability

A key property of implicatures, one that distinguishes them from entailments, is that they are *cancelable*. This is not true of entailments. Sentence (3) entails that there is a patient in the tent.

(3) There is a smallpox patient in the tent.
→ There is a patient in the tent.

To say *There is a smallpox patient in the tent, but there is no patient in the tent* is a contradiction. That is, semantic entailments cannot be canceled. The situation is different for implicatures: although the sentence *Some of the camp*

elders ate canned peaches conversationally implicates that not all camp elders ate canned peaches, it is possible to explicitly cancel this implicature by adding an extra clause, as in *Some of the camp elders ate canned peaches, in fact, all of them did.*

Besides being *explicitly* cancelable, scalar implicatures are also *implicitly* cancelable by not being licensed by the current context.[4] An example of implicit cancellation is given in (4) (taken from Levinson, 2000, p. 51).

(4) John: Is there any evidence against them?
Peter: Some of their identity documents are forgeries.
Not all of their identity documents are forgeries.[5]

In this context, John's question is taken to have explicitly fixed the level of expected informativeness. It suffices for him to know that *at least some* of their identity documents are forgeries - whether all of them are is deemed irrelevant. Thus, according to Levinson the *not all* implicature is implicitly cancelled in compliance with Quantity-2, which demands that the speaker should not make her contribution more informative than required.

Contexts in which a scalar implicature is implicitly canceled are called *lower-bound* contexts. Conversely, *upper-bound* contexts are those that license the implicature. Given the same sentence in two different contexts, whether the sentence gives rise to a scalar implicature depends (among other things) on whether it occurs in an upper-bound or lower-bound context. There are two factors that influence a context's boundedness: the implicature trigger's structural context (syntactic) and the extra-sentential pragmatic context (Katsos et al., 2005; Zon-

[4]I will speak here of implicit cancellation of the implicature; this is the terminology of, e.g. Grice (1975) and Levinson (2000). However, for the purposes of this section this is purely a framing issue; other researchers speak of the implicature not arising rather than it being canceled (Carston, 1998; Grodner, Klein, Carbary, & Tanenhaus, 2010; Huang & Snedeker, 2009; Katsos, Breheny, & Williams, 2005; Wilson & Sperber, 1995).

[5]I use # to mean "does not implicate".

dervan, 2010). In example (4) it is the extra-sentential context, i.e. the interlocutors' conversational expectations and compliance with the conversational maxims, that prevents the implicature from arising. The following is an example of a sentence that, independently of being embedded in a conversational context, does not give rise to scalar implicatures due to syntactic constraints:

(5) It is not the case that some of the camp elders ate canned peaches.
All of the camp elders ate canned peaches.

To see why, if the implicature went through, the result would be that all of the camp elders ate canned peaches, note that the alternative to (5) obtained by substituting the stronger scalemate *all* for *some* is (6).

(6) It is not the case that all of the camp elders ate canned peaches.

The next step in scalar implicature calculation, applying negation to the sentence obtained by the substitution, cancels out the negation in (6), leaving us with the supposed implicature that all of the camp elders ate canned peaches. This clearly does not conform with intuition.

Negation, the antecedent of conditionals, embedding under negative propositional attitude verbs, polar questions, 'before'- and 'without'-clauses are only some examples of downward-entailing contexts, in which scalar implicatures are systematically suspended (Chierchia, 2004; Horn, 1972; Ladusaw, 1979; Levinson, 2000).[6]

There are thus some relatively clear structural constraints on when scalar implicatures are expected to not arise. The pragmatic constraints contributing to whether a context is upper- or lower-bound are much less clear and certainly much less identifiable based on surface-level cues like lexical items. As (4) shows,

[6]Though note that the experimental literature does not paint an unequivocal picture: cancellation seems to not be obligatory, but only a preference (e.g., Chemla & Spector, 2011).

it seems the stronger alternative must be contextually relevant in order for an implicature to arise. What it means for an alternative to be contextually relevant is a matter of debate, and one that I will return to in Section 1.3.

The important points to note here are the following: scalar implicatures differ from entailments in that they are cancelable in various ways. Experimental work on scalar implicatures, which will be reviewed in Chapter 2, has crucially exploited the cancelability property to investigate the relative processing effort associated with upper- and lower-bound interpretations of utterances containing scalar terms like *some* and *or*.

1.1.4 Reinforceability

In addition to being cancelable, scalar implicatures are *reinforceable*. It is often possible to add explicitly what is already implicated without the sense of redundancy that arises when repeating an expression's coded content (Levinson, 2000). Consider again our example of the smallpox patient above, repeated here as (7).

(7) There is a smallpox patient in the tent.

(8) There is a smallpox patient in the tent and there is a patient in the tent.

Under an interpretation of (8) where *a smallpox patient* and *a patient* are coreferential the sentence is clearly strange. Adding the extra clause to the original sentence that makes the entailment of there being a patient in the tent explicit is redundant - upon hearing the first conjunct, we already know that the proposition denoted by the second conjunct is true. Repeating it is a violation of the maxim of Relation, which requires that one not make irrelevant contributions.

The same is not true of scalar implicatures. Consider the following examples:

(9) Some of the camp elders ate canned peaches, *but not all of them did.*
Scale: ⟨all, some⟩

(10) Dan or Johnny killed the Irishman, *but they didn't both do it.*
Scale: ⟨and, or⟩

(11) Brom tried to reconnoiter the rim, *but he didn't succeed.*
Scale: ⟨succeed, try⟩

(12) Al believes that Alma is a dope fiend, *but he doesn't know for sure.*
Scale: ⟨know, believe⟩

Italics indicate the reinforced implicatures. Note that none of these cases give rise to the sense of redundancy illustrated in (8). Reinforceability is thus a further difference between implicatures and entailments. Both cancelability and reinforceability are important features of implicatures because they point to a key feature of implicated content: listeners maintain much more uncertainty about what is implicated than about what is said.[7] This is one of the basic assumptions of probabilistic accounts of scalar implicature, which will be discussed in Section 2.2.3. The gist of these accounts is that scalar implicatures should be treated as probabilistic, rather than as categorical (M. C. Frank & Goodman, 2012; Franke, 2009; N. D. Goodman & Stuhlmüller, 2013; Jäger, 2013; Russell, 2012). Listeners are taken to believe that the speaker intended to convey the implicated content to some *degree*, rather than that the implicated content was either intended or not.

1.1.5 Nondetachability

Scalar implicatures are *nondetachable*, which is to say that any expression that carries the same coded content will carry the same scalar implicatures. This means that it is not possible to say the same thing in a different way without also giving rise to the same implicature. This is because scalar implicatures arise in virtue of what is said, not because of the manner of expression. Grice's example is *try*,

[7]Note that this is not the way Grice talked about this - he talked about the indeterminacy of meaning, but irrespective of speakers or listeners.

which carries some notion of failure, or the potential of failure, as in *Brom tried to reconnoiter the rim* vs. *Brom reconnoitered the rim*; "this implicature would also be carried if one said *[Brom] attempted to do x*, *[Brom] endeavored to do x*, or *[Brom] set himself to do x*" (p. 185 Grice, 1978).

That scalar implicatures are nondetachable is directly related to the nature of the maxims involved in their calculation: Quantity-1 and the Quality maxims are so-called *information-selecting* maxims (Matsumoto, 1995). Information-selecting maxims determine the choice between expressions that differ in meaning, e.g. that influence whether a stronger or weaker element on a scale is used. They are distinguished from those maxims that govern linguistic form rather than semantic content, such as the maxim of Brevity. For instance, the utterances *It is possible that Brom will fall* and *It is not impossible that Brom will fall* have the same semantic content, namely that it is possible that Brom will fall, and only differ in their linguistic form. The scalar implicature that it is not the case that he will certainly fall will arise in either case. However, the maxim of Brevity requires uttering the former, as it is briefer than the latter. Compare this to the maxim of Quantity-1, clearly an information-selecting maxim, which requires making the most informative statement possible. It governs the choice of the stronger *All men are mortal* over the weaker *Some men are mortal*, which carry different semantic content.[8]

1.1.6 Universality

Finally, Grice believed the Cooperative Principle should hold universally, i.e. regardless of cultural background, because the Cooperative Principle and the conversational maxims are ultimately fundamental presumptions of mutual rationality, without which intelligible dialogue seems difficult to imagine. However, the extent

[8]To what extent scalar implicatures are truly nondetachable is unclear (see, e.g., Recanati, 2004, for discussion).

to which the universality property holds is a matter of debate (e.g. Carston, 1998; Keenan, 1976). Different societies have different thresholds for what constitutes an appropriate level of informativeness; relatedly, politeness conventions across cultures have different effects on the packaging of information. These questions are an interesting area of study in sociolinguistics but will not concern us further here.

1.2 The GCI-PCI distinction

A controversial issue in both the linguistic and psycholinguistic literature concerns the relative context-independence of scalar implicatures compared to other types of implicatures. Grice (1975) made a distinction between two different kinds of conversational implicatures, Generalized and Particularized Conversational Implicatures (GCIs vs. PCIs). In the rest of this section I present some of the classic arguments for the distinction. However, this thesis as a whole takes issue a) with the distinction itself, and b) with the still commonly held view that scalar implicature is an instance of GCI.[9] Arguments against both a) and b) will be presented in Section 1.3. Empirical evidence against b) will be presented in Chapters 3, 4, and 5. However, the rest of this section will be concerned with the observations and arguments that gave rise to the distinction in the first place.

The GCI-PCI distinction is based on the observation that some kinds of implicatures seem to arise with much more regularity than others. Grice formulated the distinction in terms of the importance that context plays for the implicature. PCIs are carried by "saying that p on a particular occasion in virtue of special features of the context, cases in which there is no room for the idea that an implicature of this sort is NORMALLY carried by saying that p" (Grice, 1975, p. 56,

[9]Note that b) presupposes a), so giving up the distinction entails giving up the view that scalar implicature is one or the other, but the reverse is not the case.

emphasis in the original). In contrast, of GCIs he says "the use of a certain form of words in an utterance would normally (in the ABSENCE of special circumstances) carry such-and-such an implicature or type of implicature." (Grice, 1975, p. 56, emphasis in the original). Thus, GCIs arise virtually independently of context, while PCIs rely heavily on context, though both kinds of implicatures are deemed cancelable.

Consider (14) (repeated from (1)) as an answer to (13a). The Doc can be taken to mean that not all of the camp elders ate canned peaches and in addition, that there was something wrong with the peaches, as in (15a). In contrast, consider his utterance as a response to (13b): In this case, the scalar implicature that not all of the camp elders ate canned peaches still goes through, but the Doc can no longer be taken to implicate (15a). However, now he can be taken to implicate that the rules were broken by some of the camp elders (because e.g. eating snacks between meals is not allowed), which was not an eligible implicature when the question was (13a). These kinds of observations, where scalar implicatures seem to arise independently of context, have contributed to their status as GCIs in contrast to the more context-dependent PCIs in (15).

(13) a. Jane: Why are so many people sick?

b. Jane: Did anyone break the rules?

(14) Doc Cochran: Some of the camp elders ate canned peaches.
$\rightsquigarrow$ Some, but not all, of the camp elders ate canned peaches. (GCI)

(15) a. $\rightsquigarrow$ There was something wrong with the canned peaches. (PCI)

b. $\rightsquigarrow$ The rules were broken by the people who ate canned peaches. (PCI)

Thus the difference between GCIs and PCIs is in the role that context plays; GCIs arise by default and can be *canceled by context*, while PCIs *arise in virtue of context*. It is these kinds of observations that have contributed to the catego-

rization of scalar implicature as the prime example of GCI (Gazdar, 1979; Grice, 1975; Levinson, 2000).

Some have gone further than simply making the empirical observation that different kinds of inferences arise with different degrees of regularity. Levinson (2000) proposed the distinction as a solution to the *articulatory bottleneck problem*. I present Levinson's framing of the problem and his solution here in brief:

There is a significant articulatory bottleneck in the rate of information that can be transmitted via human speech (estimated by Levinson (2000) as out-of-context phoneme information). In addition, integrating contextual information to derive complex pragmatic inferences is hard and effortful.[10] Nevertheless, linguistic communication proceeds at a miraculous speed. Thus, the communicative system must have evolved a solution to these problems. Levinson's solution is to reduce the overall cost of generating inferences for listeners by taking the GCI-PCI distinction seriously as a distinction between inferences that require different amounts of processing effort. Frequently and systematically arising inferences (GCIs) should be deriveable at no cost, thus balancing out the cost of deriving difficult contextual inferences (PCIs).

The *regularity* with which scalar implicatures arise has thus led to an often implicit, sometimes explicit assumption about the *frequency* of scalar implicatures:

(16) The Frequency Assumption (scale version)
Scalar implicatures arise more often than not when the weaker of two scalemates is used.

[10]Note that there is much evidence from the psycholinguistic literature that suggests that listeners can actually very rapidly integrate information from many contextual cues online (Altmann & Kamide, 1999; Chambers, Tanenhaus, & Magnuson, 2004; Grodner & Sedivy, 2011; Heller, Grodner, & Tanenhaus, 2008; Sedivy, Tanenhaus, Chambers, & Carlson, 1999; Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995). Thus, processing contextual information may not in fact be costly, and so neither may processing PCIs which crucially depend on processing of contextual information. This assumption of Levinson's is thus questionable and I will not discuss it further here, but see Section 2.2.2 for a discussion of the consequences of relaxing it.

For example, Bott, Bailey, and Grodner (2012) cautiously say that "using a weak expression from a set of stronger alternatives often implies that the stronger alternatives are not applicable". A bolder claim is made by Huang and Snedeker (2009): "the lower-bounded interpretation may be vanishingly rare in real-world communication". Horn (1984) restricts the claim to unmarked contexts: "as a generalized implicatum, the aforementioned [scalar] inference goes through in unmarked contexts, but it may be cancelled" (Horn, 1984). And finally, Breheny, Katsos, and Williams (2006) note an almost conventionalized status for scalar implicatures in saying that they "show a degree of regularity and have the intuitive feel of components of conventional meaning".

Thus there is a strong intuition that, given a scale, it is more common than not for a scalar implicature to arise. But: how do listeners know that they are observing the weaker of two elements on a scale?

One answer is that some scales are lexicalized. That is, certain lexical items participate in scales more often than others. For example, the ⟨all, some⟩ scale has been claimed to be lexicalized; *some* is assumed to very frequently occur in utterances that compete with a stronger utterance containing *all* (Levinson, 2000). In contrast, consider the following example.

(17) Context: Two friends (Alex and Florian) are meeting on campus. Florian is late. Both of them know that Florian is biking to campus from home and that on his path he will first have to bike down Averill Ave and then down Mount Hope. Alex calls Florian to figure out where he is, i.e. how much longer he must wait.
Alex: Where are you?
Florian: I'm biking down Averill.
$\rightsquigarrow$ Florian has not biked down Mount Hope yet.

In this case, an ad hoc scale ⟨biked down Mount Hope, biked down Averill⟩ is

exploited, leading to a scalar implicature that is entirely dependent on context, in particular on the involved parties' mutual knowledge about the ordering of biking segments from Florian's home to campus. However, the ⟨biked down Mount Hope, biked down Averill⟩ scale has never been claimed to be lexicalized (and rightly so!), since it is safe to assume that the string *biked down Averill* most likely occurs in many contexts where *biked down Mount Hope* is not a salient alternative.[11]

There are many cases that lie between the lexicalized ⟨all, some⟩ and the ad hoc ⟨*biked down Mount Hope*, *biked down Averill* ⟩ scales (e.g., ⟨beautiful, pretty⟩, ⟨hound, dog⟩). What is common to all of these cases is the Gricean reasoning that is assumed to take place: the speaker could have uttered a stronger alternative; they didn't; thus the stronger alternative must not hold.

Observing a lexical item like *some* that is taken to be a member of a lexicalized scale thus can be thought of as providing a good cue to a scalar implicature, compared to the observation of an item that is less obviously on a lexicalized scale. Some authors make this explicit by tying the Frequency Assumption to particular lexical items rather than propositional alternatives: "*some* typically implicates *not all*" (Gundel, Ntelitheos, & Kowalsky, 2007, p. 4); "*some* generally implies *not all*" (Huang & Snedeker, 2009, p. 407).

This invites an alternative formulation of the Frequency Assumption:

(18) The Frequency Assumption (lexical version)
Scalar implicatures arise more often than not when a scalar item is used.

where I take a scalar item to be a lexical item or expression, e.g., *some*, that typically occurs in sentences that are the weaker of two alternatives.

[11] It is likely that both a) the *absolute* frequency of talking about biking segments in Rochester compared to using quantifiers, as well as b) the *relative* frequency of sentences about biking segments being stronger alternatives to other sentences about biking segments compared to the relative frequency of *all* (and other quantifiers) being stronger alternatives to *some* play a role in whether or not a scale is considered lexicalized, or in the degree of a scale's lexicalization, if one prefers a gradient view. I focus here on the relative frequency.

For Levinson, the Frequency Assumption is important because it provides important justification for the distinction between GCI and PCI as one with processing ramifications. However, the Frequency Assumption remains untested to date. In Chapter 5, it is tested for the case of implicatures from *some* to *not all*, and its importance for the relevant accounts introduced in Chapter 2 is discussed.

Regardless of the frequency of scalar implicatures, the GCI-PCI distinction itself remains controversial. Many researchers have argued that the distinction is of no theoretical use or that it is one of degree rather than type (e.g. Carston, 1998; Hirschberg, 1985; Rooy, 2003; Wilson & Sperber, 1995). This thesis sides with the latter researchers; the empirical evidence presented in Chapters 3, 4, and 5 will hopefully convince the reader that, intuition notwithstanding, scalar implicature is a highly context-dependent phenomenon. Further arguments against the distinction will be presented throughout the thesis. In the next section I discuss one of the main motivations for the claim that scalar implicature is actually a highly context-dependent phenomenon: the inferences drawn about the use of a scalar item like *some* are greatly affected by alternative utterances the speaker could have made in context, but didn't. This lays the foundation for the experiments presented in Chapters 3 and 4.

1.3 Alternatives

One of the clearest ways in which scalar implicatures are context-dependent is that they depend on contextually salient alternatives the speaker could have produced. This includes both *scalar* and *non-scalar* alternatives. While the focus in the literature has been on understanding the constraints on scalar alternatives, I will argue that from a processing perspective, non-scalar alternatives are just as important to the generation of scalar implicatures as scalar alternatives; and indeed, that from a processing perspective the distinction is not particularly help-

ful. In this section, I discuss the constraints that have been proposed on scales; what I mean by non-scalar alternatives affecting scalar implicatures; and how the influence of both kinds of alternatives emphasizes the highly context-dependent nature of scalar implicature.

1.3.1 Scalar alternatives

While some scales are typically discussed as being lexicalized, many scalar implicatures depend on ad hoc scales that are determined by context. In Horn (1972)'s original definition of scales, the only constraint on scales he proposed was asymmetric informational entailment: the stronger alternative must entail the weaker one, but not vice versa. In the interim, it has been recognized that informational entailment is too strong a constraint. Hirschberg (1985) observed that items not ordered by informational entailment can also form scales, like rankable entities (e.g., ⟨lover, friend⟩, ⟨corporal, private⟩), rankable attributes (⟨hot, warm⟩, ⟨terrible, bad⟩), rankable activities (⟨love, like⟩, ⟨need, want⟩), spatial orderings, sets and whole/part relationships, and process stages (⟨finish mowing the lawn, start mowing the lawn⟩); any items that constitute a partially ordered set in which one item can be determined to be higher than another one can function as a scale. Others have suggested further constraints. For example, Gazdar (1979) has suggested that alternatives must share selectional restrictions and item-induced presuppositions; Atlas and Levinson (1981) have suggested that items must be equally lexicalized, belong to the same semantic field, and be of similar length.

Importantly, scales only become *functional* in context. Some scales, like those listed in the previous paragraph, can be identified independently of context. Matsumoto calls these *potential* scales. However, context is necessary for potential scales to become functional, i.e. to give rise to a scalar implicature. Matsumoto (1995) captures this in the *Conversational Condition on Horn Scales*: "The choice of [a weaker] instead of [a stronger alternative] must not be attributed to the obser-

vance of any information-selecting Maxim of Conversation other than the Quality Maxims and the Quantity-1 Maxim (i.e., the Maxims of Quantity-2, Relation, and Obscurity Avoidance, etc.)" (Matsumoto, 1995, p. 25). What the Conversational Condition amounts to is testing for each use of a weak expression from a potential scale, whether the stronger alternative is being avoided because it is irrelevant, too informative, would result in an obscure expression for the hearer, etc., rather than because the speaker does not believe it to be true. Only in the latter case are scalar implicatures expected to arise.

Matsumoto notes that the satisfaction of the Conversational Condition is context-dependent. The difference between GCI and PCI then is reduced to the question of whether or not the Conversational Condition is usually satisfied when the weaker of two expressions on a particular potential scale is uttered.

However, the problem of maintaining the distinction between GCI and PCI in terms of the importance of context becomes worse once one recognizes that there are in fact infinitely many ad hoc potential scales (like the *biked down Mount Hope* example from Section 1.2) that cannot be specified a priori, independently of context. It seems impossible to get an estimate of how many of these types of scales there are, because possibly every expression might be the weaker element on a contextual scale in the right circumstances. But this would mean that every expression is the weaker element of a potential scale. Given that there are infinitely many potential expressions in the language, the number of potential ad hoc scales is infinite. But of course not every expression leads to a scalar implicature - only a very small subset regularly does. Researchers with the courage to draw a line between lexicalized and non-lexicalized scales may muster further courage and attempt to generate an estimate of the number of lexicalized scales. This number, while probably high, will nevertheless be countable and therefore be lower than the number of potential non-lexicalized ad hoc scales, which we have already seen is likely infinite. That is, most potential scales are such that when

the weaker alternative is uttered, the Conversational Condition does not apply (probably most often because the stronger alternative was not used because it was not contextually relevant). But this suggests that most scalar implicatures are, in fact, PCIs.

In summary, the problems involved in determining whether an expression has a stronger scalar alternative with the right kinds of properties highlight the importance of context in the calculation of scalar implicatures. I next discuss how non-scalar alternatives further contribute to the context-dependence of scalar implicatures.

1.3.2 Non-scalar alternatives

Listeners have expectations about the many alternative ways in which particular states of the world can be described. Based on their previous interactions with other speakers in other contexts they make predictions about upcoming words, syntactic structures, in the current context (linguistic and otherwise). Evidence that listeners do this comes, e.g., from reading times. The time it takes for comprehenders to read a word or phrase w_i stands in a monotonic relationship with that word's contextual surprisal (Hale, 2001; Levy, 2008), which is defined as the negative log-probability of w_i in its sentential context $w_1 \dots w_{i-1}$ and extra-sentential context c: $-\log P(w_i|w_1 \dots w_{i-1}, c)$. Surprisal goes to 0 when a word must occur in the given context and approaches infinity as the word becomes less and less likely. The fact that reading times are predicted by surprisal suggests that comprehenders have (implicit) knowledge of likely lexical alternatives at each point in the unfolding utterance.

Further evidence supporting the view that listeners have expectations about likely alternatives comes from the visual world paradigm. For example, listeners have expectations about when a noun should receive additional modification in

order to disambiguate potential referents in a visual scene. Tanenhaus et al. (1995) showed that a prepositional phrase (*on the towel*) that was temporarily ambiguous between modifying either the noun (*the apple*) or the verb (*put*) in sentences that began *Put the apple on the towel*, was rapidly interpreted as modifying the noun when there were multiple apples in the scene, and rapidly interpreted as modifying the verb when there was only one apple in the scene. This suggests that listeners are aware that the modifying material is superfluous in the case where the noun itself completely disambiguates the intended referent. But what this means is precisely that listeners have contextual expectations about alternative ways in which to describe an object.

Similarly, Sedivy et al. (1999) played participants instructions such as *Pick up the tall glass*, in which the noun was pre-nominally modified by a scalar adjective. The target was distinguished more quickly from a competitor object - another object that could also be described as *tall*, such as a pitcher - when a contrasting object of the same type (i.e., a shorter glass) was present in the display than when no contrasting object was present. This suggests that listeners rapidly draw contrastive inferences. Again, drawing this kind of inference requires that listeners have expectations about likely utterances - the speaker could have just said *Pick up the pitcher* if he had meant the pitcher, because the noun is sufficient to disambiguate the target.

Using the same paradigm as Sedivy et al. (1999), Heller et al. (2008) manipulated the knowledge state of the speaker and showed that listeners made the contrastive inference only when the objects required to make the contrastive inference (short glass, tall glass, tall pitcher) were visually accessible to both speaker and listener. This suggests further that listeners rapidly constrain likely alternatives to the knowledge state of the speaker.

Note that this notion of alternatives is different from the one we have been using so far - these are not scalar alternatives. Rather, an alternative in the current

sense is any lexical alternative to an expression that was actually used. Alternatives in this sense needn't even belong to the same syntactic class. Consider, for example, one of the most prominent psycholinguistic examples, in (19).

(19) The horse raced past the barn fell.

The difficulty experienced on the final disambiguating verb is not due to a salient *scalar* alternative to the verb that listeners expect speakers to use. Instead, some alternatives to *fell* are, e.g., silence, an adverb like *quickly*, or a relativizer (with a continuation like *that was rapidly burning*). All of these result in an interpretation of the verb *raced* as a main verb, rather than as the head of a reduced relative clause as in the original example. One of the reasons comprehenders experience such difficulty with (19) is that to express that particular meaning, the speaker had better, less misleading alternatives at their disposal that they could have used, e.g. *The horse that was raced past the barn fell.* Indeed, when disambiguated in this way, listeners typically have much less difficulty processing the final verb.

Further evidence that listeners generate expectations about upcoming words comes from the ERP literature. A commonly used ERP index of semantic processing is the N400, a centroparietally distributed, negative ERP deflection that peaks at approximately 400 msec after word onset (Kutas & Hillyard, 1980). The N400 is elicited by every content word of an unfolding utterance, and its amplitude is inversely related to the ease with which the word at hand is related to its semantic context (e.g., Brown, Hagoort, & Kutas, 2000). For example, out of context, sentences like (21), which violate the selectional restrictions of the noun, elicit a much greater N400 response after the final word than sentences like (20).

(20) The peanut was salted.

(21) The peanut was in love.

However, this pattern can be reversed if the context is set up to generate an expectation for peanuts being in love. For example, Nieuwland and Van Berkum (2006) had participants listen to short stories in which a peanut was dancing happily because he had just met a girl. In this context, the N400 was much larger in response to (20) than to (21). These results provide further support for the hypothesis that listeners update their expectations about upcoming lexical material based on the discourse context, making some alternatives more salient than others and leading to increased processing difficulty upon encountering a less expected alternative.

In the next section I discuss the role of scalar and non-scalar alternatives in scalar implicature computation.

1.3.3 Interaction of scalar and non-scalar alternatives

What I have tried to suggest in the previous section is that every utterance, indeed every word of every utterance, is interpreted against a rich backdrop of alternative utterances a speaker could have plausibly produced in context. Having to interpret unexpected (i.e., high-surprisal or unpredictable) utterances leads to increased processing effort, regardless of the scalarity of the uttered expression.

The linguist interested only in the outcome of the interpretation process may consider the foregoing discussion irrelevant: the amount of processing effort expended on interpreting an utterance containing, e.g., *some*, is surely irrelevant to the question of whether or not the interpretation process yields a scalar implicature. I would like to argue that considerations of processing effort in fact *are* important to the outcome of the process. Consider the following scenario:

Charlie is at a candy store with his children. At the store there is a gumball machine that dispenses a random number of gumballs dispensed. The dispensed gumballs need to be distributed as evenly as possible among the children, so it

is important that he know the exact number. Let's assume further that Charlie knows the initial number of gumballs in the machine (e.g., 20). The machine dispenses some number of gumballs, which he doesn't see, and the candy store owner, who knows about Charlie's distribution goal, utters one of the following:

(22) You got four of the gumballs.

(23) You got all of the gumballs.

(24) You got some of the gumballs.

In this situation, both (22) and (23) will provide Charlie with information that is sufficient to achieving his goal: in the former case he knows he needs to distribute exactly four gumballs between his children,[12] in the latter case 20. What about (24)? If the store owner uttered (24), Charlie would have every right to be confused and even annoyed: the utterance conveys only that at least one of Charlie's children will receive at least one gumball. While this is more informative than receiving no information about the number of gumballs dispensed, it is a lot less informative than necessary, given Charlie's goals. Charlie might well eventually calculate the scalar implicature that he did not get all of the gumballs, but his confusion over why the store owner was not sufficiently informative may also lead him to conclude that the store owner is not being a cooperative Gricean speaker and should not be assumed to be following the Cooperative Principle. The extra processing effort incurred by the store owner's vastly underinformative utterance may even lead to the suspension of implicature calculation.

Thus a case where the unexpectedness of the utterance compared to its alternatives (hypothetically) leads to a suspension or at least a delayed implicature.

[12]This is assuming an exact number semantics, i.e., a semantics under which the interpretation of *four* is *exactly four* rather than *at least four* (e.g., Breheny, 2008). Otherwise, upon observing (22) he learns only that he has at least four gumballs to distribute - this is more information than observing no utterance at all, but less than knowing that he has exactly four gumballs to distribute.

Although this case may strike some as contrived, it serves to make the point that one cannot ignore non-scalar alternatives in the study of scalar implicatures. Scalar alternatives are just one kind of alternative to a weak expression. Every weak expression also competes with a great number of non-scalar alternatives. If some of these non-scalar alternatives are much more salient, natural, or expected contextually, then the computation of scalar implicatures may be negatively affected.[13]

Thus, the fact that every utterance is interpreted with respect to a plethora of contextually determined alternative utterances, only some of which are scalar, further emphasizes the highly context-dependent nature of utterance interpretation, and scalar implicature calculation in particular.

1.4 Summary

In this chapter I introduced the phenomenon of scalar implicature and its use as an object of study in linguistics and psychology: scalar implicature is arguably one of the most regularly arising and most studied inferences. As such, it provides a fruitful testing ground for theories of language use, in particular the interaction of semantic and pragmatic information in listeners' calculation of speaker meaning. At the beginning of this chapter I mentioned two types of data that theories of scalar implicature should be able to account for: a) outcome and b) time course data. In addition, any good theory of language use should c) provide a solution for the articulatory bottleneck problem, introduced in Section 1.2. In the next chapter, I review the processing accounts of scalar implicature proposed by others with respect to these three issues. I will focus on the degree to which contextual factors are assumed to play a role in scalar implicature processing, in particular the

[13]This assumes that alternatives are involved in early moments of utterance interpretation. See Footnote 10, Section 1.3.2 and Section 2.2.2 for references that support this claim.

salient contextual, possibly non-scalar, alternatives a speaker could have uttered but didn't.

2 Processing accounts of scalar implicature

Accounts of scalar implicature typically focus either on the *time course* or the *outcome* of the inference process.[1] The main focus of this dissertation is on time course (processing) accounts of scalar implicature, but where appropriate I will address predictions made by accounts that focus on outcome data.

The role that context is assumed to play is a fruitful way of categorizing processing accounts of scalar implicature. On the one hand, *two-stage* accounts propose an initial context-insensitive stage of utterance interpretation, with context entering the process only in a second step. Both the Default model (Levinson, 2000) and the Literal-First hypothesis (Huang & Snedeker, 2009) are examples of two-stage accounts. They differ in that the Default model assumes the pragmatic upper-bound interpretation is more basic, while the Literal-First hypothesis assumes the semantic lower-bound interpretation is more basic; in both cases, *more basic* means that that interpretation is computed before the less basic interpretation is computed (if the less basic one need even be computed; this is decided by context).

[1]This is true especially of psycholinguistic accounts of scalar implicature. In linguistics proper, the focus has typically been on how speakers convey meaning by strategically exploiting the interaction between literal and non-literal meaning, rather than on the listener's inference process. However, recent linguistic accounts of scalar implicature, some of which are reviewed in Section 2.2.3, *have* been concerned with the inference process.

In contrast, *contextualist* accounts like Relevance Theory (Sperber & Wilson, 1995; Carston, 1998) and the Constraint-Based account I propose in Section 2.2.2 do not assume discrete stages of utterance interpretation, but rather assume that contextual information affects the interpretation from the earliest moments of processing. I also discuss a variety of recent probabilistic accounts under this heading that are inspired by information theory, Bayesian decision theory, and game theory. While these accounts have thus far only been developed to make predictions about the outcome, but not the time course of scalar implicature processing, I believe these accounts can be fruitfully extended to include a notion of incrementality precisely because of the high degree of context-dependence that they assume is involved in the computation of scalar implicatures.

In the remainder of this chapter, I review the processing accounts proposed in the literature and provide an overview of the relevant empirical literature in the process. Table 2.1 at the end of this chapter summarizes the main differences between the accounts, including time course predictions.

2.1 Two-stage accounts

2.1.1 Levinson's Default account

In Section 1.2 I briefly discussed Levinson (2000)'s Default theory of scalar implicatures, which is built around the distinction between Generalized and Particularized Conversational Implicatures. Levinson takes the distinction to be a solution to the articulatory bottleneck problem: by making the highly frequent, highly regular, context-independent GCIs cost-free, the assumed high cost of integrating contextual information required for computing PCIs should be counteracted, thereby increasing the speed and efficiency of communication. Context is assumed to enter only in a second step, and if the context does not license the implicature,

it is canceled.

Scalar implicature for Levinson is the prime example of a GCI; an implicature that arises seemingly independent of context, and which only in rare cases needs to be canceled. Scalar implicatures should thus arise immediately and by default when a listener encounters a scalar item, and incur no processing cost. Only if not licensed by context (e.g. because the stronger alternative is not relevant to a contextual Question Under Discussion - (QUD, Roberts, 1996, 2004) - will the implicature be canceled in a second step. Thus, the processing predictions are clear: upper-bound interpretations of utterances with scalar items should be computed more quickly - i.e. incur a lower processing cost - than lower-bound interpretations, which require a canceling cost.

Unfortunately for Levinson's proposal, the bulk of the experimental literature points towards implicatures being costly (Bott & Noveck, 2004; Bott et al., 2012; Breheny et al., 2006; De Neys & Schaeken, 2007; Huang & Snedeker, 2009, 2011; Noveck & Posada, 2003; Tomlinson, Bailey, & Bott, 2013). For example, Bott and Noveck (2004) had participants read underinformative sentences as in (25).

(25) Some elephants are mammals.

a. At least one elephant is a mammal

b. At least one but not all elephants are mammals.

Participants' task was to respond as quickly as possible whether they thought the sentence was TRUE or FALSE by pressing a button. Notice that under the semantic lower-bound interpretation of the sentence given in (25a), the sentence is true. The pragmatic upper-bound interpretation in (25b) is false. In one experiment, participants were trained to respond either semantically or pragmatically to these sentences. In another, they were allowed to respond freely. In both cases, the results were the same: semantic TRUE responses were faster than pragmatic FALSE responses. Thus, responses reflecting that a scalar implicature had been

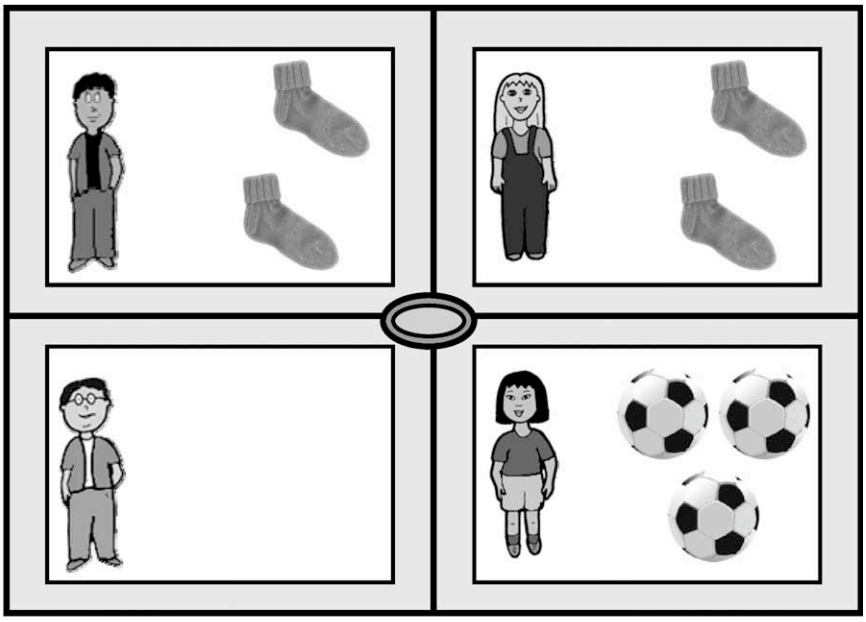

Figure 2.1: Sample display on a critical trial in Huang and Snedeker (2009), where the target utterance was *Point to the girl who has some of the socks.*

drawn were slower than responses that did not.

Similar results were obtained by Huang and Snedeker (2009) in a visual world eye-tracking paradigm. Participants viewed a display with four quadrants, with the two left and the two right quadrants containing pictures of children of the same gender, with each child paired with objects. For example, on a sample trial, the two left quadrants might each contain a boy: one with two socks and one with no objects (see Figure 2.1). The two right quadrants might each contain a girl: one with two socks (pragmatic target) and one with three soccer balls (literal target). A preamble established a context for the characters in the display. In the example, the preamble might state that a coach gave two socks to one of the boys and two socks to one of the girls, three soccer balls to the other girl, who needed the most practice, and nothing to the other boy.

Participants were asked to follow instructions such as *Point to the girl who has some of the socks.* Huang and Snedeker (2009) reasoned that if the literal

interpretation is computed prior to the inference, then, upon hearing *some*, participants should initially fixate both the semantic and pragmatic targets equally because both are consistent with the literal interpretation. If, however, the pragmatic inference is immediate, then the literal target should be rejected as soon as *some* is recognized, resulting in rapid fixation of the pragmatic target. The results strongly indicated that the literal interpretation was computed first. For commands with *all* (e.g., *Point to the girl who has all of the soccer balls*) and commands using number (e.g., *Point to the girl who has two/three of the soccer balls*), participants converged on the correct referent 200-400 ms after the quantifier. In contrast, for commands with *some*, target identification did not occur until 1000-1200 ms after the quantifier onset. Moreover, participants did not favor the pragmatic target prior to the noun's phonetic point of disambiguation (POD; e.g., *-ks* of *socks*). Huang and Snedeker concluded that "even the most robust pragmatic inferences take additional time to compute" (Huang & Snedeker, 2009, p. 408).

There is thus a growing body of evidence that scalar implicatures are not processing defaults;[2] they seem to incur a processing cost. An alternative account, which makes the inverse predictions of the Default model, is the Literal-First hypothesis.

2.1.2 Huang & Snedeker's Literal-First account

The bulk of the empirical literature is better predicted by the Literal-First hypothesis (Huang & Snedeker, 2009). Under this model, the lower-bound semantic

[2]There is a separate body of work devoted to the question of whether, and if so, precisely where, *embedded implicatures* can occur (Chemla & Spector, 2011; Chierchia, Fox, & Spector, 2008; Geurts & Pouscoulous, 2009; Russell, 2012; Sharvit & Gajewski, 2008). One prominent account posits that scalar implicatures arise by default at arbitrary embedded locations (Chierchia et al., 2008). Note that this is a different sense of *default* - here, *default* means that scalar implicatures are generated by the grammar. While making predictions about interpretation outcome patterns, this account does not make time course predictions. I will have nothing further to say about this debate in this dissertation; the interested reader is referred to the above references.

interpretation is computed rapidly and automatically as a by-product of basic sentence processing. All inferences, including scalar implicatures, require extra time and resources. For proponents of the Literal-First hypothesis this follows from the traditional observation in the linguistic literature (e.g. Horn, 2004) that the semantic interpretation of simple declaratives containing *some* are in an important sense more basic than the pragmatic interpretation: the upper-bound interpretation *some but not all* always entails the lower-bound interpretation *at least one*. Huang and Snedeker (2009) translate this into a two-stage processing sequence: upon encountering a scalar item, the semantic interpretation is necessarily constructed before the pragmatic one. To the extent that there is a processing distinction between PCIs and GCIs, it is that the relevant dimensions of the context and the interpretations that drive the inference are more circumscribed and thus perhaps more accessible for GCIs.

Thus, the Literal-First hypothesis predicts the computation of upper-bound interpretations to take more processing effort than that of lower-bound interpretations and the processing of literal content more generally. In the previous section we have seen that much of the literature supports this hypothesis. However, some studies find that scalar implicatures are computed just as quickly as literal controls (Grodner et al., 2010; Breheny et al., 2013).

For example, using a paradigm very similar to the one used by Huang and Snedeker (2009), Grodner et al. (2010) found evidence for rapid interpretation of pragmatic *some*, using displays and a logic similar to that used by Huang and Snedeker (2009). Each trial began with three boys and three girls on opposite sides of the display and three groups of objects in the center. A prerecorded statement described the total number and type of objects in the display. Objects were then distributed among the participants (see Figure 2.2). Participants then followed a pre-recorded instruction of the form *Click on the girl who has some/all/none of the balloons.*

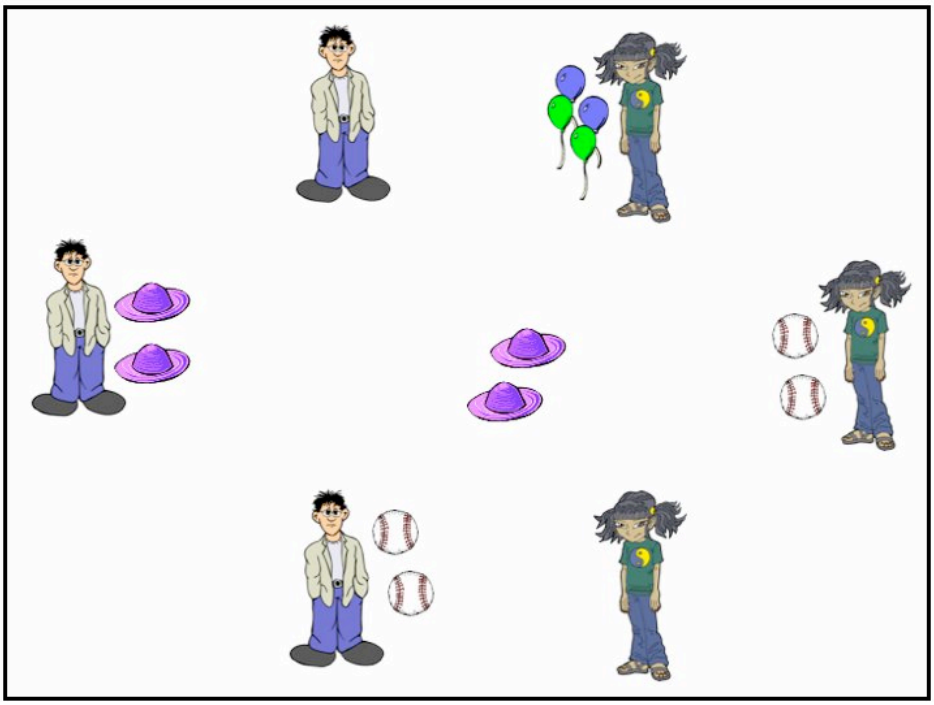

Figure 2.2: Sample display on a critical trial in Grodner et al. (2010), where the target utterance was *Click on the girl who has some of the soccer balls.*

Convergence on the target for utterances with *some* was just as fast as for utterances with *all*. Moreover, for trials on which participants were looking at the character with all of the objects at the onset of the quantifier *some*, participants began to shift fixations away from that character and to the character(s) with only some of the objects (e.g., the girl with the balls or the girl with the balloons) within 200-300ms after the onset of the quantifier. Thus participants were immediately rejecting the literal interpretation as soon as they heard *some*. If we compare the results for *all* and *some* in the Huang and Snedeker and Grodner et al. experiments, the time course of *all* is similar but the upper-bound interpretation of *some* is computed 600-800ms later in Huang and Snedeker's study. The reasons for why two studies as superficially similar as Huang and Snedeker (2009) and Grodner et al. (2010) might find such dramatically different results are discussed in Chapter 3. The important point here is that there are conditions

under which scalar implicatures *are* computed rapidly, with no delay compared to the lower-bound interpretation or literal controls.

Crucially, fast implicature computation has been observed not only with arguably lexicalized scales like the ⟨all, some⟩ scale, but also with ad hoc scales. For example, Breheny et al. (2013) found that participants rapidly computed exhaustivity implicatures of the form shown in (26) that required taking into account knowledge of whether the speaker had full access to the contextual information about which items were placed into two boxes (A and B).

(26) The woman put a spoon into Box B and a spoon and a fork into Box A.
⇝ The woman put nothing else into Box B.

Participants looks to Box B increased significantly in the 200 - 300ms window after the first *into*, suggesting that participants had generated the implicature that nothing else was put into the box in question - there was only one box in the display which had only one spoon placed in it. Crucially, when participants knew that the speaker was not aware of whether a second object had been placed in Box A, no early convergence on the target box was observed. Together, this study provides further evidence that a) listeners can rapidly generate ad hoc Quantity implicatures and b) listeners can rapidly access contextual information about speaker knowledge. These results cannot be explained by the Literal-First account without additional assumptions.

2.1.3 Interim summary of two-stage accounts

Both the Default and the Literal-First account predict that scalar implicature computation is initially a matter of context-independent default interpretation: the basic interpretation is computed first, and contextual information is integrated in a second step. The two accounts differ only in which interpretation is assumed

to be the default interpretation. For the Default model it is the upper-bound interpretation, for the Literal-First hypothesis it is the lower-bound interpretation. The two accounts thus make opposite time course predictions - the Default model predicts rapid implicatures and slow cancellation, while the Literal-First hypothesis predicts slow implicatures and rapid computation of the literal interpretation.

Both of these accounts fail to fully capture the empirical data. While most of the literature reports delayed implicatures, in line with the Literal-First hypothesis, some studies find rapid effects of implicature computation as predicted by the Default model. The divergent evidence suggests that scalar implicature may not be as robust a phenomenon as traditionally assumed, and might instead be very sensitive to contextual constraints such as discourse expectations, alternative expressions a speaker could have used, knowledge about the speaker, and task dependencies. I next review contextualist accounts that assume precisely that scalar implicature computation *is* affected by contextual constraints from the earliest moments of processing and propose a novel contextualist Constraint-Based account.

2.2 Contextualist accounts

Contextualist accounts of scalar implicature have in common that they view scalar implicature as a highly context-dependent phenomenon, but differ in a) whether semantic information is ascribed a special status, b) whether the outcome of the inference process is assumed to be probabilistic or categorical, c) whether one interpretation is assumed to be more basic than the other and if so, which one, and d) resulting differences in the speed with which implicatures are calculated.

2.2.1 Relevance Theory

Relevance Theory (Sperber & Wilson, 1995; Carston, 1998), like the Literal-First hypothesis, assumes that the semantic interpretation is basic. In Relevance Theory the upper-bound interpretation is only computed if required to reach a certain threshold of relevance in context. In contrast to the Literal-First hypothesis, however, Relevance Theory does not necessarily assume a processing cost for the pragmatic inference. If the context provides sufficient support for the upper-bound interpretation, it may well be computed without incurring additional processing cost. However, a processing cost will be incurred if the upper-bound interpretation is relevant but the context provides less support.

Proponents of Relevance Theory share Grice's original intuition that utterances raise expectations of relevance. However, in contrast to the two-stage accounts, they question whether the Cooperative Principle and any set of maxims are necessary or even appropriate to provide a cognitively realistic account of utterance interpretation. They further reject the distinction between Generalized and Particularized Conversational Implicatures - according to Relevance Theory, all pragmatic processes are equally context-dependent. Whether an implicature arises is ultimately determined by the Cognitive Principle of Relevance, which is a meta-principle of cognition meant to replace the Gricean apparatus as a whole.

(27) *Cognitive Principle of Relevance*

Human cognition tends to be geared to the maximization of relevance.

A given utterance is processed by integrating the salient context with the assumption that the utterance itself is maximally relevant to the discourse. This process yields full-blown utterance interpretation. Therefore, the utterance will be interpreted as carrying an implicature if the implicature is necessary to achieve the listener's threshold for utterance relevance.

In Relevance Theory, relevance is defined in terms of cognitive effects and processing effort (Wilson & Sperber, 1995):

(28) Other things being equal, the greater the positive cognitive effects achieved by processing an input, the greater the relevance of the input to the individual at that time.

(29) Other things being equal, the greater the processing effort expended, the lower the relevance of the input to the individual at that time.

A positive cognitive effect is characterized as a worthwhile difference to a person's representation of the world, a true conclusion that matters to that person for example by answering a question, settling a doubt or correcting a mistaken impression. False conclusions are also cognitive effects, but not positive ones.

Processing effort is defined as the "effort of perception, memory, and inference" (Wilson & Sperber, 1995, p. 2) required in the derivation of cognitive effects from an incoming stimulus. The greater the processing effort, it is claimed, the less rewarding the input will be to process, and thus the less deserving of the listener's attention.

In interpreting an utterance, the listener assumes a) that the speaker's utterance is relevant enough to be worth processing and b) that the utterance is the most relevant one compatible with the speaker's abilities and preferences. On the basis of these assumptions, the listener then follows "a path of least effort in computing cognitive effects" (Wilson & Sperber, 1995, p. 8) by testing interpretive hypotheses (implicatures, disambiguations, reference resolutions, etc.) in order of accessibility. This procedure stops when the listener's expectations of relevance are satisfied. The process of utterance interpretation is thus the simultaneous maximization of cognitive effects and minimization of processing effort.

How does this work when applied to the derivation of scalar implicatures? Consider the following exchange:

(30) Ellsworth: Are all of your items on sale?

Sol: Some are.

$\rightsquigarrow$ Not all of the items are on sale.

That Sol has some items on sale is relevant enough to be worth Ellsworth's attention (as indicated by the fact that he asked a question about items on sale). However, it is not sufficient to satisfy his expectations of relevance. Presumably, Sol was able, and not reluctant, to tell Ellsworth whether all of his items are on sale, and that would have been more relevant to Ellsworth (as indicated by the use of *all* in his question). Because Sol did not say that all of them are on sale, Ellsworth is entitled to interpret him as meaning that only some, but not all of the items are on sale.

Lower-bound contexts are assumed to not give rise to scalar implicatures because in those contexts the literal meaning itself is deemed sufficient to satisfy the listener's expectations of relevance. Thus, under the Relevance Theoretic account, the computation of scalar implicatures is reduced to the application of the Principle of Relevance: if a given utterance containing a scalar item like *some* is more relevant (in terms of positive cognitive effects and processing effort) with than without the scalar implicature, then it will arise.

Relevance Theory's main problem is vagueness: it is not clear how to quantify the notion of cognitive effects and processing effort. Further (and as a result of this), it remains unclear which kind and number of cognitive effects are worth which kind and amount of processing effort. This makes it difficult to arrive at any useful empirical (processing) predictions (c.f. Levinson, 2000, for a fully elaborated critique). However, costly implicature effects have been interpreted as evidence for Relevance Theory when it was considered the only option to the Default model (e.g. Bott & Noveck, 2004; Noveck & Posada, 2003), but Relevance Theory is also in principle compatible with the rapid implicature effects found e.g. by Grodner et al. (2010) and Breheny et al. (2013). The one type of time

course pattern that would be incompatible with Relevance Theory is one where the lower-bound interpretation is computed more slowly than the upper-bound one. To my knowledge no study to date has resulted in this pattern.

2.2.2 A Constraint-Based account of scalar implicature

In this section, I outline a Constraint-Based account of scalar implicatures which offers a competing explanation to two-stage accounts for the speed and efficiency of language processing. The account I am proposing is anchored in insights from two traditions which assume that language processing is probabilistic. The first is the information theoretic approach to language, which is based on communication theory, a theory that formalizes communication as the transmission of information through a noisy channel. Information theoretic approaches assume that context plays a central role in maximizing the efficiency of information transmission. The second tradition is in constraint-based approaches to language processing, which propose that perception, cognition and language processing involve the interactive processing of multiple, probabilistic constraints.[3] These approaches emphasize context as a rich source of accessible constraints that are rapidly used in language processing. I first outline the information theoretic approach to language use. I then discuss evidence for rapid use of context that has emerged from constraint-based approaches to language processing, and the principles and mechanisms proposed within the constraint-based tradition. Finally, I spell out how the Constraint-Based account, based on these two traditions, applies to pragmatic inference, and scalar implicature processing in particular.

[3]The terms *cue* and *constraint* are used interchangeably in the literature (e.g., Elman, Hare, & McRae, 2004), and I will do so here, as well.

General assumptions

The Constraint-Based account has its roots in information theoretic approaches to language. The hypothesis is that use of information from context allows language to efficiently use reduced and otherwise underspecified or ambiguous forms (Zipf, 1949). The general idea is that it is cognitively and articulatorily expensive to make information explicit in language, but relatively cheap for listeners to make use of readily available information from context. This idea has recently undergone a renaissance. For example, Jaeger and colleagues (A. Frank & Jaeger, 2008; Jaeger, 2010; Qian & Jaeger, 2012) demonstrate that aspects of language production follow principles that derive from communicative efficiency. Speakers seem to maximize the degree to which information is distributed uniformly across an utterance, following the principle of Uniform Information Density (Levy & Jaeger, 2007; Jaeger, 2010), a provably optimal strategy for language production under the assumption that communication proceeds through a noisy channel. Similarly, language comprehension is sensitive to the contextual probability (formalized as the surprisal or inverse log probability) of words and syntactic structures (Hale, 2001; Levy, 2008).

In seminal work, Shannon (1948) proved that when context is informative, unambiguous utterances are partially redundant with the context and therefore inefficient; an efficient communication system will thus always include individual units that appear ambiguous out of context but are unambiguous in context. Building on Shannon and Zipf's work, Piantadosi, Tily, and Gibson (2012) have recently argued that rather than impeding communication, ambiguity allows for greater ease of processing by allowing efficient (easy to produce) linguistic units to be re-used. From this perspective, ambiguity in language is not a problem that complicates language processing. Instead, ambiguous systems are desirable because they allow for re-use of words and sounds that are easy to produce and understand; that this is indeed the case is supported by the findings that shorter,

more predictable, and less surprising (i.e., easier) words have a greater number of homophones (map onto a greater number of word lemmas) than harder words; shorter, more predictable, and phonotactically less surprising words have a greater number of lexical meanings on average than harder words; and shorter, more predictable, and less phonotactically suprising syllables occur in a greater number of words than harder syllables.

The communicative efficiency approach depends on an assumption that stands in direct conflict with one of the most important basic assumptions made by Levinson (2000) and others who treat the integration of contextual information as a cognitively effortful, strategic, non-automatic process (Neely, 1977; Posner & Snyder, 1975; Shiffrin & Schneider, 1977). The assumption is that listeners in fact make rapid use of information from context.

There is a growing body of evidence in support of this assumption. Indeed, rapid use of context is not limited to syntactic and lexical ambiguity resolution (MacDonald, Pearlmutter, & Seidenberg, 1994; Tanenhaus & Trueswell, 1995; Trueswell, Tanenhaus, & Kello, 1993), but is also found in pragmatic processing: listeners generate expectations about the domain of subsequent reference (Altmann & Kamide, 1999) and rapidly circumscribe referential domains using information that includes the referential context (Spivey, Tanenhaus, Eberhard, & Sedivy, 2002; Tanenhaus & Trueswell, 1995; Trueswell, Sekerina, Hill, & Logrip, 1999), presuppositions and affordances relevant to intended actions (Chambers, 2002; Chambers et al., 2004) and differences between their own perspective and that of their interlocutors (Barr, 2008; Brown-Schmidt, Gunlogson, & Tanenhaus, 2008; Hanna & Tanenhaus, 2004; Hanna, Tanenhaus, & Trueswell, 2003; Heller et al., 2008; Nadig & Sedivy, 2002; but cf. Keysar, Barr, Balin, & Brauner, 2000; Kronmüller & Barr, 2007; Wu & Keysar, 2007).

Rapid use of context, moreover, is not a language-specific property, but rather permeates every sensorimotor activity in which humans are engaged, allowing

behavior to flexibly adapt to novel situations (see, e.g., Elman et al., 2004). But how does this occur? The information listeners have access to is rarely completely transparent - the input merely provides cues to word meaning, syntactic structure, and speaker meaning. These cues are not all equally informative; some of them are very reliable while others are not; and some are very frequent while others are rare. Cues sometimes provide converging evidence for a particular syntactic parse or word meaning; other times they conflict. Successful comprehension requires integrating all of this information and resolving potential conflicts in the right kind of way. As a further complication, due to the temporal nature of unfolding speech, different cues become available at different points in time, so the comprehension process must be amenable to dynamic updating.

Underlying constraint-based accounts of language comprehension is the assumption that listeners solve this complex cue integration problem by relying on probabilistic, distributional knowledge of cues which is acquired by learners over many years of linguistic exposure (e.g., Elman et al., 2004). In order to efficiently deal with the vast amount of incoming information, listeners learn to identify those contextual cues that are reliable indicators of sentence structure or word meaning. That is, listeners acquire sophisticated knowledge of the kinds of utterances speakers are most likely to produce in different contexts to convey different meanings. As a result, listeners have extensive knowledge of the types of alternatives that a speaker is likely to use in different types of situations, as well as knowledge about how alternatives might interact.

Here I propose to apply this constraint-based view to scalar implicature processing, and indeed to pragmatic inference more generally. When applied to scalar implicature processing, I will talk about the Constraint-Based account (in uppercase) as an exemplar of constraint-based approaches to language processing more generally.

Applying the Constraint-Based approach to scalar implicatures

Under the Constraint-Based account, the problem of scalar implicature calculation is treated as a problem of ambiguity resolution analogous to syntactic ambiguity resolution or word sense disambiguation. But for scalar implicature the ambiguity is not an ambiguity between alternative syntactic structures or word senses. Rather the ambiguity is one between different states of the world that are compatible with the linguistic input. For the sake of concreteness, let's assume two relevant states of the gumball world: $s_{\exists\neg\forall}$, in which some, but not all of the gumballs are dispensed and $s_{\forall}$, in which all of the gumballs are dispensed. Assume further that a listener observes the utterance in (31). The lower-bound interpretation of this utterance, shown in (31a), is compatible with both states. The upper-bound interpretation in (31b) is only compatible with $s_{\exists\neg\forall}$.

(31) You got some of the gumballs.

a. You got at least one, and possibly all, of the gumballs.

b. You got some, but not all, of the gumballs.

To understand the way in which the ambiguity between different states of the world that the speaker intends to communicate is resolved under the Constraint-Based account it might be helpful to contrast it with the Default and Literal-First model. Recall that the Default model assumes the upper-bound interpretation to become immediately available upon encountering the implicature trigger (in this case, *some*). If the context does not license the interpretation, it is canceled to yield the lower-bound interpretation of the utterance. The Literal-First model assumes that the lower-bound interpretation is computed first as part of the semantics of the utterance. The upper-bound interpretation is computed in a second step only if licensed by the context.

In contrast to these two models, the Constraint-Based account does not posit two interpretation steps, but rather assumes that both the robustness of the upper-

bound interpretation and the speed with which it is calculated depends on the probabilistic support for the interpretation that is contextually available to listeners. Probabilistic support, in this case, consists of the (potentially very complex) interactions of multiple cues to speaker meaning that differ in reliability and become available at different points during the unfolding utterance. Let's examine this in more detail.

The Constraint-Based account assumes that listeners have prior expectations about which is the actual state of the world, which become updated as more cues to the intended interpretation become available. That is, the listener's belief state prior to observing an utterance with *some* can be modeled as a distribution over states of the world $S = \{s_{\exists\neg\forall}, s_{\forall}\}$. This distribution may be uniform as in the left panel of Figure 2.3, reflecting that the listener has maximal uncertainty about the actual state of the world. It may also be skewed towards either of the two states, reflecting that the listener believes one state of the world to be more likely than the other. For example, the middle panel of Figure 2.3 shows the initial hypothetical subjective beliefs of a listener who believes $s_{\forall}$ is more likely than $s_{\exists\neg\forall}$. In contrast, the listener in the right panel of Figure 2.3 believes $s_{\exists\neg\forall}$ is more likely, and is more certain about this belief.

As the utterance begins to unfold, the listener's initial belief distribution can be updated incrementally based on the observed input. For instance, consider the partitive form *of*. Assume for the time being that speakers, when they use the partitive with *some*, intend the upper-bound interpretation more often than not, and when they don't use the partitive, intend the lower-bound interpretation more often than not. That is, the upper-bound interpretation is more likely with (31) than with (32).

(32) You got some gumballs.

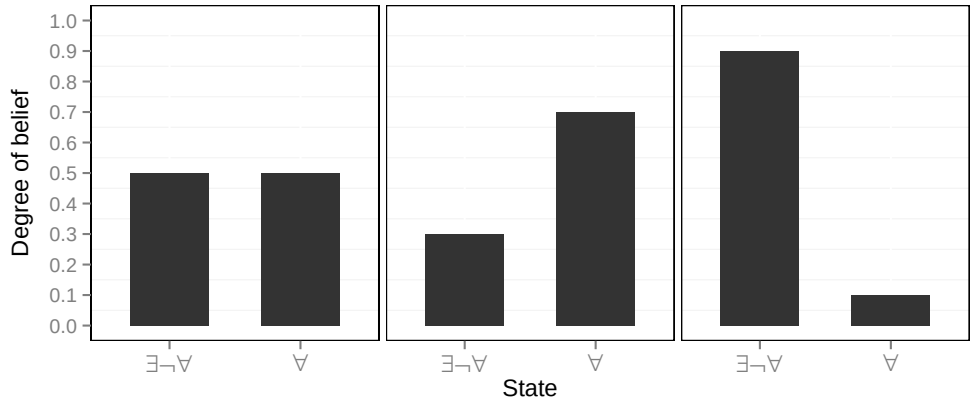

Figure 2.3: Hypothetical listeners' prior belief distributions over states of the world $S = \{s_{\exists\neg\forall}, s_{\forall}\}$. The left panel shows a maximally uncertain listener, the middle panel shows a more certain listener who believes $s_{\forall}$ is more likely, the right panel shows an even more certain listener who believes $s_{\exists\neg\forall}$ is more likely.

Assume further that listeners, over the course of their linguistic experience, learn the statistics of the partitive cue and use this distributional knowledge to make predictions about speaker meaning. Now consider what happens as the utterance unfolds, ignoring all cues except for the partitive. Up until *You got some*, the listener has gained no information, so he should stick with his prior beliefs. But the next piece of input is informative: in (31), the listener observes *of*, while in (32), he observes *gumballs*. That is, he observes the presence or absence of the partitive cue. The listener's beliefs can now be updated; his posterior belief in $s_{\exists\neg\forall}$ should be larger after observing the partitive than after observing the lack of the partitive.

However, the magnitude and direction of the update depends on prior beliefs. Without going into too much detail: if the likelihood function $P(\mathit{Partitive}|S)$, where *Partitive* is a binary variable coding the presence or absence of the partitive form *of*, is more skewed when the partitive is present than than the listener's prior belief distribution, then his posterior beliefs will be more skewed after observing the partitive. If it is less skewed than the prior distribution, the posterior will be

more uniform. That is, the former case *reduces* uncertainty about s, while the latter case *increases* it.

Of course this is a vastly simplified scenario; in practice listeners have much more information available to them, and this information may interact in complex ways. For example, *some* may be prosodically marked in various ways which signals varying degrees of speaker certainty about the upper-bound interpretation; the stronger alternative *all* may be more or less relevant to the QUD; the monotonicity properties of the context may be such that the scale is reversed[4]; the speaker may be estimated by the listener to have various degrees of authority concerning the truth of the stronger alternative; and crucially for the experimental manipulations in Chapters 3 and 4, there may be more or less natural or expected alternatives to *some* the speaker could have produced.[5]

To what extent these cues actually affect scalar implicature processing is currently an open question (but see, e.g., Zondervan, 2010; Geurts & Pouscoulous, 2009; Bergen & Grodner, 2012; Breheny et al., 2013, for investigations of the relevance of the stronger alternative, monotonicity properties of the context, and two studies of speaker knowledge, respectively). Part of the research program under a Constraint-Based view of scalar implicature is precisely to identify those cues that do affect scalar implicature and investigate their interactions. There are two ways in which a particular cue can affect the implicature: a) by modulating the *robustness* with which the implicature is computed and b) by modulating the *speed* with which it is computed.

Chapters 3, 4, and 5 provide evidence of both kinds of modulation by various contextual cues: in Chapter 3 the effect of the partitive and the naturalness and

[4]For example, if the utterance was *I doubt that you got some of the gumballs*, the belief space itself changes: now the relevant interpretations are no longer those listed in (31a) and (31b), but the possibility of not having received any gumballs at all is additionally introduced.

[5]The simplified state space S will be expanded in Chapter 3 to accommodate the more complex gumball situations used in the experiments as well as the way in which the naturalness or expectedness of alternatives to *some* should affect the probability of generating an implicature.

availability of number terms as alternatives to *some* on categorical truth value judgments and response times are investigated. In Chapter 4, the effect of naturalness and availability of number alternatives on scalar implicature processing is investigated using eye movements as a dependent measure. Finally, Chapter 5 departs from categorical implicature judgments and attempts to quantify a) the amount of support generally available for scalar implicatures from *some* to *not all* (this amounts to testing the Frequency Assumption introduced in Section 1.2) and b) the effect of various cues on implicature strength, using corpora of spontaneous speech.

Predictions of the Constraint-Based account

What are the predictions of the Constraint-Based approach for the kinds of situations the Default and Literal-First model have been applied to? Where the Default model predicts rapid implicatures compared to literal controls/the lower-bound interpretation and the Literal-First model predicts slow implicatures, the Constraint-Based model predicts rapid implicatures where support for the implicature is high, and slow implicatures where support is low. That is, where there are one or more cues that support the upper-bound interpretation, a) implicature rates should be higher and b) implicature computation should be faster.

At this point a note on how implicature rate and speed are measured is in order. The theoretical and experimental literature thus far has typically treated scalar implicature as a categorical phenomenon: either an implicature is ultimately computed or it is not. This is reflected in the types of experimental tasks that are used to measure participants' interpretations - truth value judgment and referent identification tasks, that require categorical judgments, are the most prevalent (Bott & Noveck, 2004; Bott et al., 2012; Grodner et al., 2010; Huang & Snedeker, 2009, 2011; Tomlinson et al., 2013). Even researchers who use experimental measures like reading times, that do not require a commitment to an assumption of categor-

ical implicature computation, typically speak of the phenomenon as categorical (Breheny et al., 2006; Bergen & Grodner, 2012). In the experiments reported in Chapters 3 and 4, the tasks employed also force participants to make categorical judgments. However, built into the Constraint-Based account as I have laid it out is the assumption that listeners ultimately arrive at a probabilistic belief in the upper-bound interpretation; that is, rather than computing the implicature or not, listeners have a belief distribution over relevant states of the world. The distribution of categorical judgments obtained in each of the experiments can be seen as reflecting participants' posterior belief distributions. This view is compatible with the probabilistic approaches discussed in the next section (M. C. Frank & Goodman, 2012; Franke, 2009; N. D. Goodman & Stuhlmüller, 2013; Jäger, 2013; Russell, 2012) and in the experiments reported in Chapter 5, gradient implicature judgments are collected that more adequately reflect the assumptions made by the Constraint-Based account.

The experimental tasks that have been used to measure implicature speed also vary: researchers have employed response times in judgment tasks (Bott & Noveck, 2004; Bott et al., 2012), reading times (Bergen & Grodner, 2012; Breheny et al., 2006), or time for eye movements to converge on a target compatible with only the upper-bound interpretation (Breheny et al., 2013; Grodner et al., 2010; Huang & Snedeker, 2009). Each of these measures comes with its own set of measure-internal sources of noise. I will make use of response times and eye movements and discuss the relative merits and problems with the measures where appropriate. However, for each measure the general prediction of the Constraint-Based account is that a) the more support for the upper-bound interpretation that is contextually available, the faster participants' responses reflecting the upper-bound interpretation should be, and b) the more uncertainty that is present in participants' posterior belief distribution, the slower they should respond.

Frequently asked questions

I will spend the last part of this section addressing additional questions that readers might have. In particular

1. How are the competing interpretations selected?
2. What determines a listener's prior belief distribution?
3. What exactly does the Constraint-Based account assume is learned?
4. Is there a place for Gricean reasoning under a Constraint-Based account?

Selection of competing interpretations. The Constraint-Based account assumes that there are some competing states of the world that receive different amounts of probability mass from contextual cues. So far I have assumed the situation relevant to the generation of scalar implicatures: that the states are $s_{\exists\neg\forall}$ and $s_{\forall}$. For the general case, in line with previous work I will assume that the interpretations of interest are selected by virtue of a contextual QUD introducing a partition on the set of possible worlds (Franke, 2009; Russell, 2012). For the example I have been using, the QUD that introduces the right kind of partition is *Did I get all of the gumballs?*[6]. However, other partitions are possible. For example, if the QUD is *How many gumballs did I get?*, the relevant interpretations correspond to each state of the gumball machine (i.e., zero gumballs received, one gumball received, two gumballs, etc.). I spell this out further in Chapter 3. For now, it is sufficient to note that one of the advantages of the Constraint-Based account (as opposed to, e.g., the Default or Literal-First models), is that it can easily be extended to capture such situations.

Prior belief distribution. Listeners' prior beliefs should be a function of the (subjective) contextual probability of each of the states in S being the actual

[6]This is actually not entirely accurate, since the worlds denoted by the sentences in (31) rule out the situation in which the addressee received zero gumballs, but it is of no consequence for our purposes here.

one before observing any input. For example, in a gumball machine situation in which the gumball machine randomly dispenses some number of gumballs, prior beliefs about $s_{\exists\neg\forall}$ will depend on the initial number of gumballs in the machine. For example, if there are initially two gumballs in the machine, there are three states the gumball machine can end up in: dispensing either zero, one, or two gumballs, which I will refer to as s_0, s_1, and s_2. That is, $p(s_{\exists\neg\forall}) = p(s_1)$, the prior probability of the upper-bound interpretation being true is $1/3$. In contrast, if there are 10 gumballs in the machine initially, the prior probability of the upper-bound interpretation being true is $p(s_{\exists\neg\forall}) = p(s_1) + p(s_2) + \ldots + p(s_9) = 9/11$.

Learning. The Constraint-Based approach assumes that listeners learn the joint distribution of a great number of cues and associated inferences. However, listeners are limited in their memory and attentional resources. The extent to which these limitations affect what can be learned, and to what level of detail the actual distribution of cues is approximated in listeners' acquired language models, is a very interesting open question, but one that I will not have anything further to say about here.

Indeed, constraint-based accounts of sentence processing are sometimes criticized for containing too many free parameters. It is argued that these kinds of accounts can be made to fit any behavioral pattern. However, this is not an entirely fair criticism since it is precisely the goal of constraint-based accounts to identify and quantify the different cues to interpretation and their complex interactions that listeners are sensitive to (see also Elman et al., 2004). When the constraints are independently motivated and quantified, then an explicit model makes specific and clearly falsifiable predictions (e.g., McRae, Spivey-Knowlton, & Tanenhaus, 1998; Dell, Schwartz, Martin, Saffran, & Gagnon, 2000).

The role of Gricean reasoning. Some readers may object that there is no role for Gricean reasoning under this kind of account. I will end this section by addressing this objection. In particular, I will argue that the Constraint-

Based account is fully compatible with Gricean principles operating on utterance interpretation. To see this, it is helpful to think about the nature of some of the cues I have been discussing.

Consider, for instance, the naturalness or expectedness of the alternatives to *some*. So far I have argued in Section 1.3 that both scalar and non-scalar alternatives should affect the speed with which scalar implicatures are generated. In doing so, I largely took for granted that there is some set of alternatives that listeners are considering in any instance of utterance interpretation. However, as discussed, exactly *how* this set is generated is still an unresolved question, and not one that I will answer in this thesis. Nevertheless, some constraints on alternatives have been identified in the literature. For example, informativeness relations between sentences affect whether a particular utterance is considered to be an alternative. In the scalar case, this is reflected in that the stronger alternative denotes a proper subset of the worlds that the weaker alternative denotes. But informativeness arguably plays a role in the evaluation of non-scalar alternatives as well. Consider the case of number terms.

Number terms are typically not considered to lie on the same scale as *some*. Nevertheless, many uses of *some* allow for inferences to be drawn about the sentence in which a number term is used instead of *some* (and vice versa). Consider again the sentences in (22) and (24), repeated here as (33) and (34).

(33) You got four of the gumballs.

(34) You got some of the gumballs.

The worlds in which (33) is true form a proper subset of the worlds in which (34) is true, regardless of whether *four* is enriched to mean *exactly four*. In contrast, (34) only rules out that the listener got zero gumballs. This makes the truth of (33) more likely, but does not assign it a truth value. Thus, the same asymmetrical entailment relations hold in this case between the sentence

with *some* and *four* as typically obtains between *some* and *all*. If the speaker is following the Cooperative Principle, he should thus utter (33) if he knows it to be true, since it is more informative than (34).

Thus, considerations of informativeness are clearly involved in the selection of (even non-scalar) alternatives that utterances are interpreted relative to; the contextual naturalness or expectedness of a given utterance and its alternatives is affected by its contextual informativeness. However, that can only be the case if there is some pressure for favoring more informative over less informative alternatives - but this is exactly the maxim of Quantity-1.

Gricean maxims are also involved in other cues to speaker meaning, e.g., the relevance of the stronger alternative to a contextual QUD would not be expected to affect scalar implicatures unless there was a pressure to produce relevant utterances in the first place - but this is exactly what speakers should do according to the maxim of Relevance.

Thus, the Constraint-Based account does not eradicate the need for principles of communication like the Gricean maxims. However, the way in which these maxims come into play is in generating the set of alternatives that a given utterance is interpreted against.

2.2.3 Probabilistic accounts

I next briefly review a class of accounts that, while generally compatible with the Constraint-Based account, make predictions only for the outcome of the inference process and are thus not strictly speaking processing accounts. However, I discuss them briefly because they share much of the spirit of the Constraint-Based account and have the additional advantage that they provide mathematical formalizations of the inference procedure that make clear quantitative predictions, and that, if endowed with a notion of incrementality and cue timing, have the potential for

fruitful cross-pollination with the Constraint-Based account.

Game-theoretic models

The recently emerging branch of pragmatics known as game-theoretic pragmatics has made great advances in implementing (parts of) the Gricean program in a mathematically rigorous framework that makes clear empirical predictions about pragmatic inference behavior. I focus here in particular on the class of Iterated Response models (Franke, 2009, 2011; Jäger, 2013; Degen, Franke, & Jäger, 2013), though there are many different flavors of game-theoretic models of different pragmatic inferences (Parikh, 1991; Ross, 2006; Benz & Rooij, 2007). What these approaches have in common is the assumption that speakers and listeners coordinate their actions so that speakers rationally choose utterances, given a state of the world, and listeners rationally choose interpretations, given an observed utterance.

Iterated Best Response (IBR, e.g. Franke, 2009) models treat communicative settings as signaling games (Lewis, 1969) between two players, a *sender* (speaker) and a *receiver* (listener). The sender knows the actual *state* of the world (what she intends to communicate to the receiver) but the receiver doesn't. The sender chooses a *message* (an utterance) from a set of alternatives, all of which are assumed to have a semantic meaning commonly known between players. The receiver observes the message and chooses an *action* (an interpretation). Choices depend on players' *utilities*. In cooperative situations, utilities are assumed to be symmetric: utility is positive if communicative success is achieved, and 0 otherwise.

Players' behavior is captured in the concept of a *strategy*. For senders, this is a probabilistic function from states to messages; for receivers it is from messages to actions. Players differ in their sophistication, which is captured by the levels of iterated reasoning they perform. Unsophisticated level-0 senders send true

messages at random. Unsophisticated level-0 receivers interpret messages literally. For example, in a situation where all students failed an exam, an unsophisticated sender would choose one of the messages *all of the students failed* or *some of the students failed* at random. An unsophisticated receiver would interpret an ambiguous message like *some of the students failed* literally - she would choose at random between interpretations where not all of the students failed and all of the students failed.

In contrast, more sophisticated players choose the *best response* given their beliefs about the other player's likely choices and beliefs. Best responses are those that maximize expected utility given the other player's strategy. That is, they are a function of utilities of outcomes and beliefs about the other player's likely choices. Receivers' posterior beliefs about the sender's state given the observed message are obtained via Bayesian updating of their prior beliefs over states with the sender's likelihood function of sending that message in any given state. Best responses can be computed iteratively for increasingly sophisticated player types. In our example, a level-1 sender's best response to the unsophisticated receiver is to send the message *all of the students failed* when in fact all students failed, because sending the *some* message does not maximize utility: there is a chance that the literal receiver will interpret it as some, but not all students having failed, in which case the outcome utility is 0. Similarly, a level-1 receiver's best response to the unsophisticated sender would be to interpret the *some* message as conveying a scalar implicature that not all students failed.

These types of models have been successfully employed to account for a variety of pragmatic inferences, though their empirical validity has only started to be investigated in recent years (Degen & Franke, 2012; H. Rohde, Seyfarth, Clark, Jäger, & Kaufmann, 2012). For example, Degen and Franke (2012) tested the IBR model of Franke (2011) on the production and comprehension of referential expressions that required scalar ad hoc reasoning. While inference complexity was

correlated with the number of best responses given (reflecting scalar inferences), participants did not consistently select best responses. This is most likely due to limitations of processing resources; more recent versions of these models currently under development include probabilistic choice rules (Iterated Quantal Response (IQR), Degen et al., 2013).

Rational Speech Act models

Bayesian reasoning also forms the foundation of the *Rational Speech Act* models of M. C. Frank and Goodman (2012) and N. D. Goodman and Stuhlmüller (2013), which are based on pragmatic inference as a rational response to an intuitive theory of language production. As in the IBR models, the listener's task is to infer the state of the world, given an utterance produced by a speaker who is assumed to be informative and approximately rational.[7]

However, in contrast to the game-theoretic IBR model discussed above, speakers and listeners are assumed to make small mistakes in the mapping from expected utilities to choice probabilities, rather than always playing best responses. This allows for capturing deviation from optimality (similar to IQR[8]). To capture the motivation to be informative, utility is modeled as related to the information conveyed by the utterance, specifically the amount of information that a literal listener would not yet know about the state of the world after hearing it described by a particular utterance.[9]

N. D. Goodman and Stuhlmüller (2013) apply this model to standard scalar implicature arising from *some* and numerals, and enrich it to encompass the lis-

[7]These types of Bayesian models have actually found wide applicability, not only to pragmatic inference - they have been used in word learning (M. C. Frank, Goodman, & Tenenbaum, 2009; Xu & Tenenbaum, 2007), causal learning (N. D. Goodman, Ullman, & Tenenbaum, 2011; Tenenbaum, Griffiths, & Kemp, 2006), and social cognition (Baker, Saxe, & Tenenbaum, 2009; Ullman, Goodman, & Tenenbaum, 2012).

[8]In fact, the choice rule used in both types of models is the soft-max function (Sutton & Barto, 1998), also known as the quantal response rule or the logit choice rule.

[9]This is exactly the information-theoretic notion of surprisal.

tener's knowledge about the uncertainty the speaker may have about the actual state of the world. M. C. Frank and Goodman (2012) apply a similar model to the interpretation of referential expressions in contexts similar to the ones employed by Degen and Franke (2012). In both cases, good model fits to experimentally obtained judgment data were observed.

The Rational Speech Act models differ from the Iterated Best Response models in three ways. First, the Iterated Best Response model allows for in principle unbounded iterated reasoning on both the part of the speaker and the listener, whereas the Rational Speech Act model defines the rational listener as a level-1 listener and the rational speaker as a level-1 speaker (though the Rational Speech Act model can in principle be extended to perform recursive reasoning). Second, the Rational Speech Act model assumes that speakers follow a heuristic to be as informative as possible given the listener, while Iterated Response models assume that speakers and hearers both perform full-blown iterated reasoning. Finally, the Rational Speech Act model assumes a probabilistic choice rule, while IBR employs a categorical choice rule. However, IQR is a recent variant of the Iterated Response model that assumes that interlocutors soft-maximize expected utility (as in the Rational Speech Act models). The empirical consequences of these differences are currently being explored.

Russell's theory of utterance interpretation

A related probabilistic perspective is offered by the Bayesian Utterance Interpretation approach (Russell, 2012). In the same spirit as the game-theoretic and Rational Speech Act models considered above, Russell treats scalar implicatures as resulting from listeners' expectations about the kinds of utterances speakers are likely to produce. That is, his system, just like the other probabilistic accounts discussed earlier, is intended to be a general theory of utterance interpretation and pragmatic reasoning which scalar implicature falls out of - it is not a theory

designed to account only for scalar implicatures. This is in line of course with the Gricean aim of providing a theory of rational behavior, whether the behavior be linguistic or not.

Under Russell's theory, listeners take a speaker to implicate the negation of the stronger alternative in proportion to how contextually likely the speaker was to utter the stronger alternative rather than the weaker one if he believed the stronger alternative was true. That is, a speaker can be said to implicate the negation of the stronger alternative if the fact that he uttered the weaker alternative increases the probability that the stronger alternative is not the case. Thus, the problem of implicature computation under Russell's theory is a matter of comparing two conditional probabilities: a) the probability that a speaker will make a weak utterance, given a weak belief and b) the probability that the speaker will make the weak utterance, given a relatively strong belief. These probabilities differ between contexts and are determined by multiple factors: whether the speaker believes the utterance and its alternatives are true, the relevance of the utterance and its alternatives to a contextual QUD (or conversational point being made),[10] and the complexity of the utterance and its alternatives measured in terms of string probability. These probabilities capture the Gricean maxims of Quality, Relevance, and some aspects of Manner.

Crucially, Russell's apparatus requires that there be a set of contextual states of the world and utterance alternatives that interlocutors are mutually aware of (according to the listener). The output of the computation, importantly, is not categorical. Rather than a scalar implicature being computed or not, the speaker's posterior belief in $s_{\exists\neg\forall}$ state is compared to that of $s_{\forall}$. The greater the difference, the stronger the implicature.

[10]Russell employs a probabilistic notion of relevance, *Carnap relevance*, which I will return to in Chapter 3.

Interim summary of probabilistic accounts

The Utterance Expectation account has in common with the Iterated Response and Rational Speech Act models that they all depend crucially on a formalization of the listener's expectations about the speaker's behavior. It is based on estimates of the speaker's likely behavior that listeners make pragmatic inferences, especially with an eye towards the alternative utterances that the speaker could have made in similar and minimally different contexts. Pragmatic inference is assumed to occur probabilistically via rational Bayesian update. The outcome of the inference process is a distribution over states of the world - a scalar inference in these types of systems amounts to increasing the probability of $s_{\exists\neg\forall}$ relative to $s_{\forall}$. These basic assumptions are exactly on par with those made by the Constraint-Based account. However, the Constraint-Based account does not provide a fully spelled out formal theory, and the different probabilistic models make different assumptions about how exactly listeners' prior beliefs are updated with contextual information.

What the probabilistic accounts discussed are currently lacking, and what the Constraint-Based account provides, is the notion of *incremental* belief update. Thus, rather than performing computations on utterances of entire sentences, the Constraint-Based account highlights that cues to speaker meaning become available at different points in the utterance and predicts that processing effort - in particular the processing effort associated with deriving the implicature - should vary depending on how much support for the implicature is provided by contextual cues at different points in the unfolding utterance.

Nothing prevents the probabilistic accounts discussed from being enriched with a notion of incrementality, and the Constraint-Based account would benefit from a fully spelled-out model incorporating cue estimates that can be used to make clear quantitative predictions. Indeed, I believe that combining these kinds of probabilistic accounts with the Constraint-Based account is the most promising

approach towards development of a unifying account of scalar implicature outcome and time course, while at the same time providing an intuitive solution to the articulatory bottleneck problem. This is an interesting and potentially highly rewarding area of future research. However, in this thesis I focus on the first step: showing that both the robustness and the speed of scalar implicature calculation is affected by multiple cues.

2.3 Summary of accounts and predictions

Table 2.1 summarizes the processing speed predictions of the accounts discussed in this chapter: the Default model, the Literal-First hypothesis, Relevance Theory, and the Constraint-Based account. Other important properties of the accounts are listed alongside the time course predictions.

The next three chapters will be devoted to testing the predictions of the different accounts by investigating the context-dependence of scalar implicatures using a wide range of experimental methods. To foreshadow, we will see converging evidence that scalar implicature is a highly context-dependent phenomenon that is very sensitive to the naturalness and availability of both scalar and non-scalar alternatives as well as to many other linguistic and extra-linguistic cues to speaker meaning.

Table 2.1: Overview of time course predictions and general properties of the different accounts of scalar implicatures.

	Default	Literal-First	Relevance Theory	Constraint-Based	Probabilistic
Basic interpretation	upper-bound	lower-bound	lower-bound	neither	neither
Special status for semantic information	no	yes	yes	no	yes
Nature of implicature	categorical	categorical	categorical	probabilistic	probabilistic
Processing predictions for upper-bound interpretation compared to					
... lower-bound	faster	slower	not faster	context-dependent	NA
... literal control	equally fast	slower	not faster	context-dependent	NA

3 The effect of alternatives on response times

3.1 Introduction

In this chapter I investigate the effect of two types of cues on the time course of scalar implicature, both of which are motivated by the idea that speakers could have produced alternative utterances. The first cue is the partitive *of*, which marks the difference between (35) and (36).

(35) Alex ate some cookies.

(36) Alex ate some of the cookies.

Intuitively, (36) leads to a stronger implicature than (35).[1] However, this intuition has not previously been tested. Moreover, researchers have used either one or the other form in their experimental stimuli without considering the effect this might have on processing. For example, some researchers who found delayed implicatures and use some form of the Literal-First hypothesis to explain these findings have consistently used the non-partitive form (Bott et al., 2012; Bott & Noveck, 2004; Noveck & Posada, 2003). Under a Constraint-Based account, the absence of the partitive may provide weaker support for the implicature than use

[1] See Section 5.3.1 for a detailed discussion of the reasons for the propensity of the partitive to increase scalar implicature strength.

of the partitive form would have provided, resulting in increased processing effort to arrive at the implicature.[2]

The second cue is the availability of lexical alternatives to *some* that listeners assume are available to the speaker. Here I focus on the effect of number terms as alternatives, and the effects of intermixing *some* with number terms, e.g., *You got two of the gumballs* vs. *You got some of the gumballs*. The motivation for this second cue comes from two studies (Grodner et al., 2010; Huang & Snedeker, 2009) which used similar methods but found different results.[3]

Both Huang and Snedeker (2009, 2011) and Grodner et al. (2010) used the visual world eye-tracking paradigm (Cooper, 1974; Tanenhaus et al., 1995) to investigate the speed with which scalar implicatures are computed compared to literal controls. However, their results were strikingly different: while Huang and Snedeker (2009) found a large delay for the upper-bound interpretation of *some* compared to the interpretation of literal controls like *all* and number terms, Grodner et al. (2010) found that participants rapidly computed the implicature; there was no delay compared to the literal control sentence with *all*. What caused this difference in studies as similar as these?

The Grodner et al. (2010) design and analysis differed from Huang and Snedeker (2009) in several ways. First, the prerecorded statement in Grodner et al. drew participants' attention to the total cardinality of each set of objects. Second, there were some potentially important differences in the auditory stimuli used by Grodner et al.; participants were asked to click on the target items (rather than point) and *some of* was pronounced *summa* (with the reduced vowel) to avoid potential ambiguity with non-partitive *some*. Grodner et al. also included a *none of* (*nunna*) condition. Third, Grodner et al. corrected for participants' baseline preference to look at the character with more objects, which appeared in all of

[2]The corpus study presented in Chapter 5 provides further evidence from spontaneous speech in corpora that the presence of the partitive form increases support for the scalar implicature.

[3]See Chapter 2 for a more detailed description of the studies.

the experiments reported in Huang and Snedeker and Grodner et al. (though the bias was only reliable in some of the Huang and Snedeker experiments). Although these factors might account for a small difference in timing between the Huang and Snedeker and the Grodner et al. experiments, it seems unlikely that even a combination of these factors could account for a difference as large as 600 to 800 ms.

A more plausible explanation is that Huang and Snedeker included stimuli with numbers, whereas Grodner et al. did not. In Huang and Snedeker (2009), a display in which one of the girls had two socks was equally often paired with the instruction *Point to the girl who has some of the socks* and *Point to the girl who has two of the socks.* In fact, Huang, Hahn, and Snedeker (2010) have shown that eliminating number instructions reduces the delay between *some* and *all* in their paradigm and Grodner (personal communication) reports that including instructions with number results in a substantial delay for *summa* relative to *alla* in the Grodner et al. paradigm.

Why might instructions with exact number delay upper-bound interpretations of partitive *some*? Computation of speaker meaning takes into account what the speaker could have, but did not, say with respect to the context of the utterance. Recall that the upper-bound interpretation of *some* is licensed when the speaker could have, but did not say *all*, because *all* would have been more informative. More generally, it would be odd for a speaker to use *some* when *all* would be the more natural or typical description. In situations like the Huang and Snedeker and Grodner et al. experiments, exact number is arguably more natural than *some*. In fact, intuition suggests that mixing *some* and exact number makes *some* less natural.

Consider a situation where there are two boys, two girls, four socks, three soccer balls and four balloons. One girl is given two of the four socks, one boy the four balloons and the other boy, two of the three soccer balls. The descriptions

in (37) and (38) seem natural, compared to the description in (39) (assuming the speaker knows exactly how many objects of each type each child received):

(37) One of the girls got some of the socks and one of the boys got all of the balloons.

(38) One of the girls got two of the socks and one of the boys got all of the balloons.

(39) One of the girls got two of the socks and one of the boys got some of the soccer balls.

Grodner et al. (2010) provided some empirical support for these intuitions. They collected naturalness ratings for their displays and instructions, both with and without exact number included in the instruction set. Including exact number lowered the naturalness ratings for partitive *some* but not for *all*. However, even without exact number instructions, *some* was rated as less natural than *all*. One reason might be that in most situations, pragmatic *some* is relatively infelicitous when used to describe small sets. Again, intuition suggests that using *some* is especially odd for sets of one and two. Consider a situation where there are three soccer balls and John is given one ball and Bill two. *John got one of the soccer balls* seems a more natural description than *John got some of the soccer balls* and *Bill got two of the soccer balls* seems more natural than *Bill got some of the soccer balls*.

These observations suggest an alternative hypothesis for why responses to pragmatic *some* are delayed when intermixed with exact number for small set sizes: pragmatic *some* is delayed when there is a rapidly available, more natural alternative to *some*. This seems especially likely for small sets because the more natural number alternative is also a number in the subitizing range where number terms become rapidly available and determining the cardinality of a set does not require counting (Atkinson, Campbell, & Francis, 1976; Kaufman, Lord, Reese, &

Volkman, 1949; Mandler, Shebo, & Vol, 1982). In situations where exact number is available as a description, the number term is likely to become automatically available, thus creating a more natural, more available interpretation of the scene.[4] This is especially the case for referential tasks, where the goal is to identify the intended referent as soon as possible; the Huang and Snedeker (2009) and Grodner et al. (2010) tasks were exactly of this sort.

My argument, then, is that delays in response times (to press a button or to fixate a target) that have previously been argued to be due to the costly computation of the Quantity implicature from *some* to *not all* might in fact be due to interference from lexical alternatives to *some* that do not lie on the same scale as *some*. The motivation for this interference, in Gricean terms, comes from the maxim of Manner: if there is a less ambiguous quantifier (e.g., *two*) that could have been chosen to refer to a particular set size, the speaker should have used it. If she didn't, it must be because she meant something else. Finally arriving at the implicature from *some* to *not all* when there are more natural lexical alternatives to *some* that the speaker could have used but didn't, may involve both reasoning involving the Quantity maxim (standard scalar implicature) and the Manner maxim (inference that the speaker must have not meant the partitioned set which could have more easily and naturally been referred to by *two*). If this is indeed the case, it would have serious implications for the interpretation of response time results on scalar implicature processing across the board: delays previously associated with costly scalar implicature computation might be at least partly due to costly reasoning about unnatural, misleading quantifier choices.

[4]See also Section 1.3 for further discussion of the role of scalar and non-scalar alternatives in scalar implicature computation.

3.1.1 The gumball paradigm

The current experiments examine the hypothesis that number selectively interferes with *some* when naturalness of *some* is low and number terms are rapidly available. The hypothesis was tested experimentally in a "gumball paradigm" that was developed for this purpose. To understand the gumball world I begin with an example, which also serves the purpose of illustrating how the Constraint-Based account applies to this situation.

Suppose that there is a gumball machine with 13 gumballs in the upper chamber. Let us assume that Alex knows that this gumball machine has an equal probability of dispensing 0 - 13 gumballs. His friend, Thomas, inserts a quarter and some number of gumballs drops to the lower chamber but Alex cannot see how many have dropped. Thomas, however, can, and he says *You got some of the gumballs.*

Before the start of the utterance Alex will have certain prior expectations about how many gumballs he got. In this case, there are 14 states the gumball machine can be in, which means there is an equal probability of 1/14 that Alex got any number of gumballs. This is shown in the first panel of Figure 3.1. Once Alex hears *You got some*, he has more information about how many gumballs he got. First, the meaning of *some* constrains the set to be larger than 0. However, Alex also has knowledge about how natural it would be for Thomas to utter *some* instead of, for example, an exact number term, to inform him of how many gumballs he got. Figure 3.1 illustrates how this knowledge might shift his subjective probabilities for having received specific numbers of gumballs.

Alex is now more certain that he has received an intermediate set size rather than a small set (where Thomas could have easily said *one* or *two* instead of *some*) or a large set (where Thomas could have said *most*, or even *all*). Finally, once Alex hears the partitive *of*, his expectations about how many gumballs he

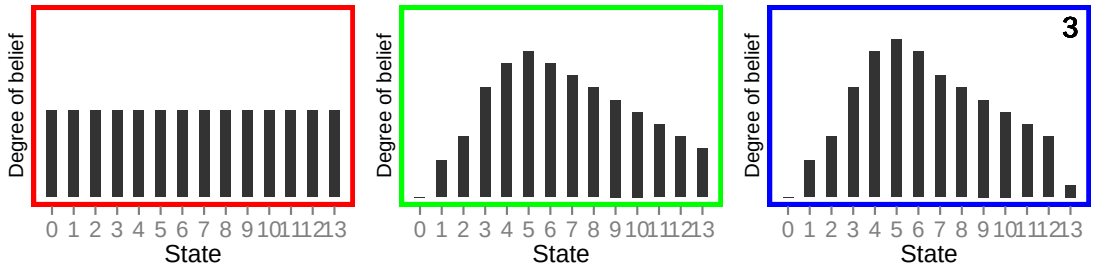

Figure 3.1: Hypothetical constraint-based update of belief distribution over states of the gumballs machine at different incremental time points in the utterance *You got some of the gumballs*. States are represented as size of set of gumballs received. Bars represent the amount of probabilistic support provided for each state, given the (hypothetical) constraints at each point: prior beliefs, semantics of *some*, naturalness of *some* and its alternatives, partitive.

got might shift even more (for example because Alex knows that the partitive is a good cue to the speaker meaning to convey upper-bound *some*), as shown in the third panel of Figure 3.1. Thus, by the end of the utterance Alex will be fairly certain that he did not get all of the gumballs, but he will also expect to have received an intermediate set size, as there would have been more natural alternative utterances available to pick out either end of the range of gumballs.[5]

The gumball paradigm was developed in order to investigate whether listeners are sensitive to the partitive and to the naturalness and availability of lexical alternatives to *some* in scalar implicature processing using a range of different set sizes. On each trial the participant sees a gumball machine with an upper chamber and a lower chamber, as illustrated in Figure 3.2. All of the gumballs begin in the upper chamber. After a brief delay, some number of gumballs drops to the lower chamber. The participant then responds to a statement describing the scene, either by rating the statement's naturalness, by judging whether they agree or disagree with the statement, or by clicking on different regions of the display. This makes it possible to obtain information about participants' judgments while at the same time recording response times (or eye movements in a somewhat modified

[5]See Section 2.2.2 for further constraints that are likely to affect incremental belief update.

version of this paradigm, see Chapter 4) as a measure of their interpretation of different quantifiers with respect to different visual scenes.

Exps. 1a and 1b are rating studies in which the naturalness of *some* is established for different set sizes of interest, in particular small sets (1 - 3), intermediate sets (6 - 8), and for the unpartitioned set (all 13 gumballs). In addition, the relative naturalness of simple *some* vs. partitive *some of* used with the unpartitioned set is investigated. Exp. 1b examines the effect of including exact number descriptions on naturalness ratings for *some* and *some of* used with different set sizes. Exp. 2 tests the extent to which naturalness ratings for *some* and other quantifiers are reflected in response times. In Exps. 4a and 4b the effect of number term inclusion on scalar implicature processing is investigated using eye movements, a more closely time-locked measure of interpretation. Exps. 3a and 3b provide naturalness norms for the stimuli used in Exps. 4a and 4b.

3.2 Exp. 1a: naturalness of *some* in the absence of number terms

Exp. 1a was conducted to determine the naturalness of descriptions with *some*, *some of the* (henceforth *summa*), *all of the* (henceforth *alla*) and *none of the* (henceforth *nunna*) for set sizes ranging from 0 to 13.

3.2.1 Methods

Participants

Using Amazon's Mechanical Turk, 120 workers were paid $0.30 to participate. All were native speakers of English (as per requirement) who were naïve as to the purpose of the experiment.

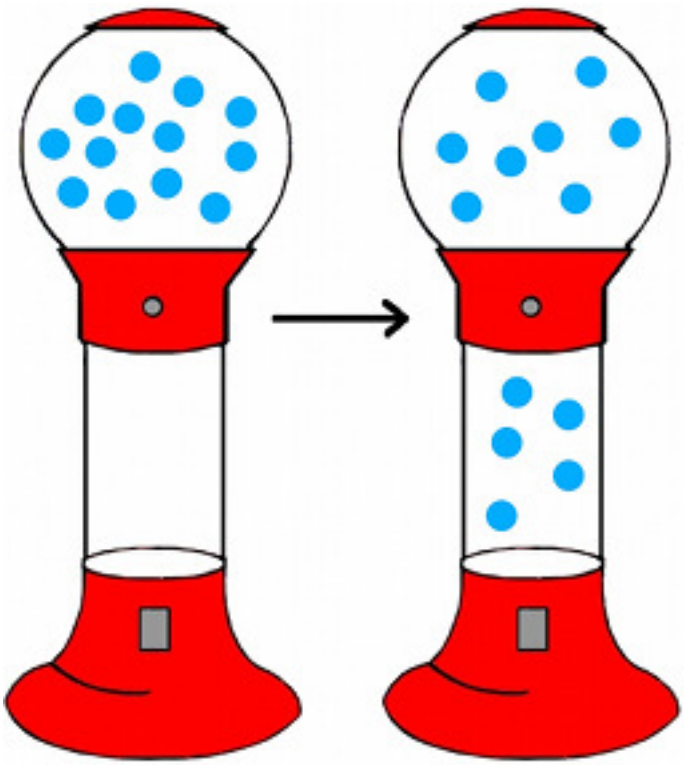

Figure 3.2: Sample displays in the gumball paradigm. Left: initial display. Right: sample second display with dropped gumballs.

Procedure and materials

On each trial, participants saw a display of a gumball machine with an upper chamber filled with 13 gumballs and an empty lower chamber (Figure 3.2). After 1.5 seconds a new display was presented in which a certain number of gumballs had dropped to the lower chamber. Participants heard a pre-recorded statement of the form *You got X gumballs*, where X was a quantifier. They were then asked to rate how naturally the scene was described by the statement on a seven point Likert scale, where seven was *very natural* and one was *very unnatural*. If they thought the statement was false, they were asked to click a FALSE button located beneath the scale. Both the size of the set in the lower chamber (0 to 13 gumballs) and the quantifier in the statement (*some*, *summa*, *alla*, *nunna*) were varied. Some trials contained literally false statements. For example, participants might get none of the gumballs and hear *You got all of the gumballs*. These trials were interspersed in order to have a baseline against which to compare naturalness judgments for *some (of the)* used with the unpartitioned set. If interpreted semantically (as *You got some and possibly all of the gumballs*), the *some* statement is true (however unnatural) for the unpartitioned set. However, if interpreted pragmatically as meaning *You got some but not all of the gumballs*, it is false and should receive a FALSE rating.

Participants were assigned to one of 24 lists. Each list contained 6 *some* trials, 6 *summa* trials, 2 *alla* trials, and 2 *nunna* trials. To avoid an explosion of conditions, each list sampled only a subset of the full range of gumball set sizes in the lower chamber. The quantifiers *some* and *summa* occurred once each with 0 and 13 gumballs. In addition, *nunna* occurred once (correctly) with 0 gumballs and *alla* once (correctly) with 13 gumballs. Each of *alla* and *nunna* also occurred once with an incorrect number of gumballs. The remaining *some* and *summa* trials sampled two data points each per quantifier from the subitizing range (1 - 4 gumballs), one from the mid range (5 - 8 gumballs), and one from the high range

(9 - 12 gumballs). See Appendix A, Table A.1 for the set sizes that were sampled on each list. From each of twelve base lists, one version used a forward order and the other a reverse order. Five participants were assigned to each list.

3.2.2 Results and discussion

Clicks of the FALSE button were coded as 0. Mean ratings for each quantifier for the different set sizes are presented in Figure 3.3. Mean ratings were 5.83 for *nunna* (0 gumballs) and 6.28 for *alla* (13 gumballs) and close to 0 otherwise. Mean ratings for *some*/*summa* were lowest for 0 gumballs (1.4/1.1), increased to 2.7 and 5.2 for 1 and 2 gumballs, respectively, peaked in the mid range (5.81/5.73), decreased again in the high range (4.69/4.74), and decreased further at the unpartitioned set (3.47/2.7).

The data were analyzed using a series of mixed effects linear regression models with by-participant random intercepts to predict ratings. In models run on different subsets of the data that corresponded to different set sizes in the lower chamber, with fixed effects of quantifier (*some*/*summa* vs. *alla*/*nunna*), *some* and *summa* were most natural when used with 6 gumballs, where ratings did not differ from ratings for *nunna* and *alla* used with their correct set size ($\beta = 0.2$, $SE = 0.24$, $p < .4$). Mean ratings for *some* and *summa* did not differ for any set size except at the unpartitioned set, where *some* was more natural than *summa* ($\beta =$ -0.77, $SE = 0.19$, $p < 01$). This naturalness difference between *some* and *summa* used with the unpartitioned set suggests that *summa* is more likely to give rise to a scalar implicature than *some* and is thus dispreferred with the unpartitioned set. Because ratings for *some* and *summa* did not differ anywhere except for the unpartitioned set, I henceforth report collapsed results for *some* and *summa*.

To test the hypothesis that naturalness for *some* varies with set size, a mixed effects linear regression model predicting mean naturalness rating from range

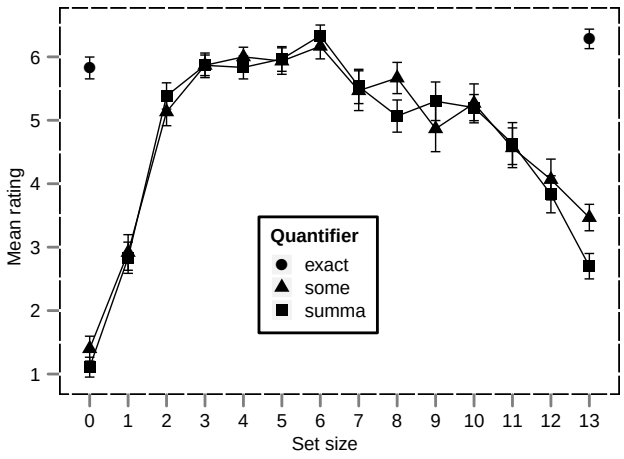

Figure 3.3: Mean ratings for *some*, *summa*, and the exact quantifiers *nunna* and *alla*. Means for the exact quantifiers *nunna* and *alla* are only shown for their correct set size.

(subitizing (1 - 4 gumballs) or mid range (5 - 8 gumballs)) was fit to the subset of the *some* cases in each range. As predicted, naturalness was lower in the subitizing range than in the mid range (β = -0.79, $SE = 0.13$, $p < .001$). Similarly, naturalness ratings for *some* were lower for the unpartitioned set than in the mid range (β = -2.68, $SE = 0.15$, $p < .001$). However, the naturalness ratings differed for set sizes within the subitizing range. (1-4). Performing the analysis on subsets of the data comparing each set size with naturalness of *some*/*summa* used with the preferred set size (6 gumballs) yields the following results. Collapsing over *some* and *summa*, small set versus 6 gumballs were coded as 0 and 1 respectively and subsequently centered. The strongest effect was observed for one gumball (β = 2.95, $SE = 0.17$, $p < .0001$). The effect is somewhat weaker for two ($\beta = 1.0$, $SE = 0.22$, $p < .0001$), even weaker for three ($\beta = 0.36$, $SE = 0.2$, $p < .06$), and only marginally significant for four ($\beta = 0.29$, $SE = 0.19$, $p < .1$). The coefficients for the set size predictor for each subset model are plotted in Figure 3.4. Given these results I will refer to "small set size" effects rather than "subitizing" effects.

In sum, *some* and *summa* were both judged to be quite unnatural for set sizes

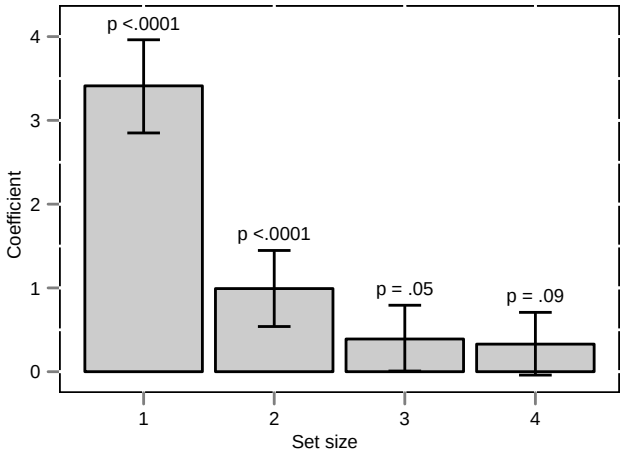

Figure 3.4: Set size model coefficient for each set size in the subitizing range.

of 1 and 2, and more natural but not quite as natural as for the preferred set size (6 gumballs) for a set size of 3. Naturalness also decreased after the mid range (5 - 8 gumballs) and was low at the unpartitioned set. In addition, the partitive, *some of*, was less natural to refer to the unpartitioned set than simple *some*.

Finally, note that naturalness ratings for *some*/*summa* gradually decreased for set sizes above 6. This is probably due to there being other more natural, salient alternatives for that range: *many* and *most*. I would predict that including *many* or *most* among the experimental alternatives would further reduce naturalness for these set sizes.

The naturalness results from this study point to an interesting fact about the meaning of *some*. The linguistic literature standardly treats the semantics of *some* as proposed in Generalized Quantifier Theory (Barwise & Cooper, 1981) as the existential operator (corresponding to *at least one*). Under this view, as long as at least one gumball dropped to the lower chamber, participants should have rated the *some* statements as true (i.e., not clicked the FALSE button). However, this was not the case: *some* received 12% FALSE ratings for one gumball and 9% FALSE ratings for the unpartitioned set; *summa* statements were rated FALSE in 7% of

cases for one gumball and 18% of cases for the unpartitioned set. For comparison, rates of FALSE responses to *some*/*summa* for all other correct set sizes were 0%.

In addition, treating *some* as simply the existential operator does not allow a role for the naturalness of quantifiers. What matters is that a statement with *some* is true or false. Differences in naturalness are not predicted. However, it is clear from this study that language users' internal representations of *some* are more complex. One way of conceptualizing these naturalness results is that we have obtained probability distributions over set sizes for different quantifiers, where the relevant probabilities are participants' subjective probabilities of expecting a particular quantifier to be used with a particular set size. Thus, a vague quantifier like *some*, where naturalness is high for intermediate sets and gradually drops off at both ends of the spectrum, has a very wide distribution, with probability mass distributed over many set sizes. In contrast, for a number term like *two* one would expect naturalness to be very high for a set size of two and close to 0 for all other cardinalities; the distribution would thus be very narrow and peaked around 2.

According to the Constraint-Based account introduced in Section 2.2.2, listeners should be able to process multiple sources of information in parallel. Assume that listeners' expectations about which quantifier will be used for a given set size depend on at least two factors: a) set size and b) awareness of contetxual availability of alternative quantifiers. If no lexical alternative is available, listeners will have some expectations about use of *some* with different set sizes. This expectation distribution is what we have obtained in Exp. 1a. For *some*, naturalness is highest for intermediate set sizes and drops off at both ends of the tested range. That is, listeners' expectation for *some* to be used is highest in the mid range. The Constraint-Based account predicts that listeners expectations about quantifier use are sensitive to alternatives. Including number terms among the experimental items, thus making participants aware that number terms are contextually available alternatives to *some*, should change this distribution. In

particular, the prediction is that due to subitizing processes, which allow number terms to become rapidly available as labels for small sets, the naturalness of *some* should decrease for small sets when number terms are included. In other words, participants' expectations that a small set will be referred to by *some* should decrease. This prediction is tested in Exp. 1b.

3.3 Exp. 1b: naturalness of *some* in the presence of number terms

Exp. 1b tested the hypothesis that when number terms are included as alternatives within the context of an experiment, the naturalness of *some* will be reduced when it is used with small set sizes. Using the same paradigm as in Exp. 1a, number terms were included among the stimuli to test the hypothesis that the naturalness of *some*/*summa* would be reduced when used with small set sizes, where number terms are hypothesized to be most natural.

3.3.1 Methods

Participants

Using Amazon's Mechanical Turk, 240 workers were paid $0.75 to participate. All were native speakers of English (as per requirement) who were naïve as to the purpose of the experiment.

Procedure and materials

The procedure was the same as that described for Exp. 1a with one difference; the number terms *one of the* through *twelve of the* were included among the stimuli. Each participant rated naturalness of statements with quantifiers as descriptions

of gumball machine scenes on 32 trials. Participants were assigned to one of 48 lists. As in Exp. 1a, each list contained 6 *some* trials, 6 *summa* trials, 2 *alla* trials, and 2 *nunna* trials. In addition, 4 number terms were included on each list. Each number term occurred once with its correct set size, once with a wrong set size that differed from the correct set size by one gumball, and once with a wrong set size that differed from the correct set size by at least 3 gumballs. The lists were created from the same base lists used in Exp. 1a. See Appendix A, Table A.1 for the set sizes that were sampled on each list. Four versions of each of the twelve base lists were created. On half of the lists, *some*/*summa* occurred before the correct number term for each set size, on the other half it occurred after the correct number term. Half of the lists sampled incorrect set sizes that were one greater than the correct set size for the number terms employed, and half sampled set sizes that were one smaller.

3.3.2 Results

Clicks of the FALSE button were coded as 0. Mean ratings were 5.71 for *nunna* with 0 gumballs and 6.31 for *alla* with 13 gumballs, and close to 0 otherwise. Mean ratings for *some*/*summa* were lowest for 0 gumballs (1.08/0.96), increased from 2.02/1.99 to 4.91/4.54 in the small set range peaked in the mid range at 6 gumballs (5.82/5.97), decreased again in the high range (4.33/4.47), and decreased further at the unpartitioned set (3.42/2.65), replicating the general shape of the curve obtained in Exp. 1a.

Again, *some* and *summa* were most natural when used with 6 gumballs. As in Exp. 1a, mean ratings for *some* and *summa* did not differ for any set size except at the unpartitioned set, where *some* was more natural than *summa* (β = -0.77, SE = 0.13, $p < .001$).

To test the hypothesis that adding number terms decreases the naturalness for

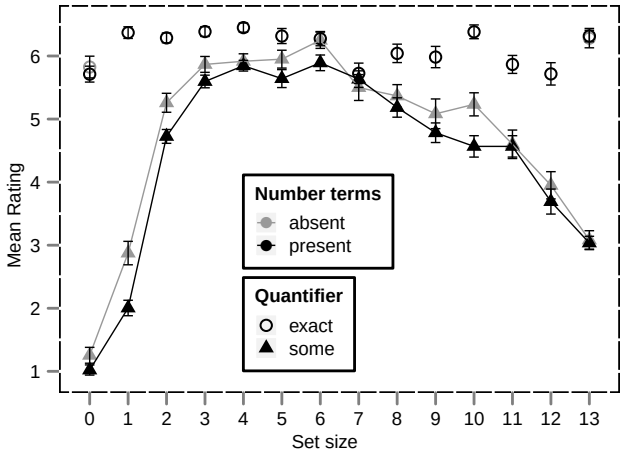

Figure 3.5: Mean ratings for *some* (collapsing over simple and partitive *some*) and exact quantifiers/number terms when number terms are present (Exp. 1b) vs. absent (Exp. 1a). Means for the exact quantifiers are only shown for their correct set size.

some/*summa* in the small set range but nowhere else, I fit a series of mixed effects linear models with random by-participant intercepts, predicting mean naturalness rating from number term presence for each range (no gumballs, small, mid, high, unpartitioned set). The models were fit to the combined datasets from Exp. 1a (numbers absent) and Exp. 1b (numbers present). As predicted, naturalness was lower for both *some* and *summa* when numbers were present in the small set range (β = -0.49, SE = 0.16, $p < .001$), but not for 0 gumballs (β = -0.19, SE = 0.17, $p < .28$), in the mid range (β = -0.14, SE = 0.14, $p < .14$), in the high range (β = -0.38, SE = 0.23, $p < .1$), or with the unpartitioned set (β = -0.11, SE = 0.23, $p < .8$). Figure 3.5 shows the mean naturalness ratings when numbers were present (Exp. 1b) and absent (Exp. 1a).

Performing each analysis individually for each set size in the subitizing range shows that the strength of the number presence effect in the small set range differed for different set sizes. A coefficient plot for the effect of number presence for different set sizes is provided in Figure 3.6. The effect was strongest for one

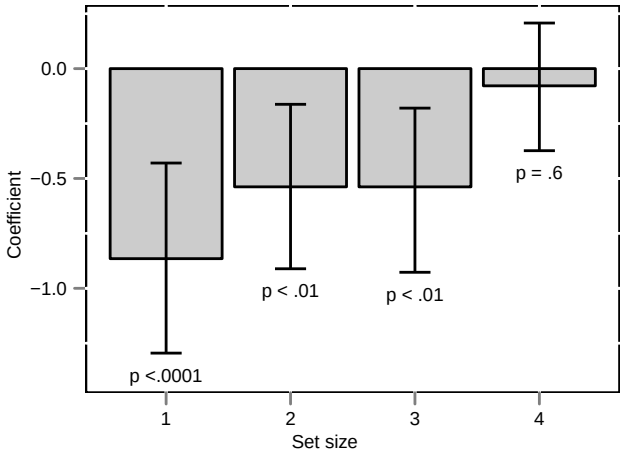

Figure 3.6: Model coefficients for number term presence predictor for each set size in the subitizing range.

gumball (β = -0.83, SE = 0.27, $p < .0001$), less strong for two (β = -0.54, SE = 0.22, $p < .01$) and three gumballs (β = -0.54, SE = 0.22, $p < .01$), and non-significant for four gumballs (β = -0.1, SE = 0.18, $p < .61$). This suggests that naturalness effects are not due to subitizing per se, as initially hypothesized. Rather, subitizing might interact with naturalness to determine the degree to which number alternatives compete with *some*.

Number terms did not reduce naturalness for *nunna* (β = -0.1, SE = 0.22, $p < .61$) and *alla* (β = -0.02, SE = 0.17, $p < .84$) when used with their correct set sizes. Finally, although ratings for number terms were very high throughout when used with their correct set size and close to floor otherwise, number terms are judged as more natural when used with small set sizes than when used with large ones as determined in a model predicting mean naturalness rating from a continuous set size predictor (β = -0.05, SE = 0.01, $p < .001$). The exception is *ten*, which was judged to be slightly more natural than the surrounding terms.

3.3.3 Discussion

The results of the two rating studies suggest that a listener's judgment of an expression's naturalness is directly affected by the availability of lexical alternatives. With the exception of 6 and 7 (around half of the original set size), numbers are always judged to be more natural than *some*/*summa* when they are intermixed. This difference, however, is largest for the smallest set sizes. As predicted, the reduced naturalness of *some*/*summa* used with small sets, established in Exp. 1a, decreased further when number terms were added to the stimuli. Therefore, at least in off-line judgments, listeners take into account non-scalar alternatives to *some* that the speaker could have uttered, but didn't.

3.4 Exp. 2: response time to *some* in the presence of number terms

Exp. 2 was designed to test whether the effect of available natural alternatives is reflected in response times. Using the same paradigm and stimuli, participants' judgments and response times to press one of two buttons (YES or NO) depending on whether they agreed or disagreed with the description were recorded. Based on the naturalness results from Exps. 1a and 1b, the Constraint-Based account predicts that participants' YES responses should be slower for more unnatural statements. Specifically, for *some* and *summa*, response times are predicted to be slower compared to their more natural alternatives when used with a) the unpartitioned set, where *alla* is a more natural alternative and b) in the small set range, where number terms are more natural and more rapidly available. Based on the naturalness data, the largest effect is expected for a set size of one, a somewhat smaller effect for two, and a still smaller effect for three. Response times for *some*/*summa* with these set sizes should be slower than when *some* and

summa are used in the preferred range (4 - 7 gumballs).

Recall that in Exps. 1a and 1b there was also a difference in naturalness between simple and partitive *some* used with the unpartitioned set. This difference should be reflected both in the number of YES responses (more YES responses to *some* than to *summa*) and in response times (faster YES responses for *some* than for *summa*).

The conditions in which *some*/*summa* are used with the unpartitioned set are of additional interest because they can be linked to the literature using sentence-verification tasks. In these conditions, enriching the statement to *You got some but not all (of the) gumballs* via scalar implicature makes it false. However, if no such pragmatic enrichment takes place, it is true. That is, YES responses reflect the semantic, *at least*, interpretation of the quantifier, whereas NO responses reflect the pragmatic, *but not all*, interpretation. Noveck and Posada (2003) and Bott and Noveck (2004) called the former *logical* responses and the latter *pragmatic* responses; I will make the same distinction but use the terms *semantic* and *pragmatic* and in addition take into account participants' response consistency.

In Bott and Noveck's sentence verification paradigm participants were asked to perform a two alternative forced choice task. Participants were asked to respond TRUE or FALSE to clearly true, clearly false, and underinformative items (e.g., *Some elephants have trunks*). Bott and Noveck found that: a) pragmatic responses reflecting the implicature were slower than semantic responses; and b) pragmatic responses were slower than TRUE responses to *alla* for the unpartitioned set. If processing of the scalar item *some* proceeds similarly in the gumball paradigm, Bott and Noveck's result should replicate: YES responses to *some* used with the unpartitioned set should be faster than NO responses, and NO responses to *some* should be slower than YES responses to *alla*.

3.4.1 Methods

Participants

Forty-seven undergraduate students from the University of Rochester were paid $7.50 to participate.

Procedure and materials

The procedure was the same as in Exps. 1a and 1b, except that: a) participants heard a "ka-ching" sound before the gumballs moved from the upper to the lower chamber and; b) participants responded by pressing one of two buttons to indicate that YES, they agreed with, or NO, they disagreed with, the spoken description. Participants were asked to respond as quickly as possible. If they did not respond within four seconds of stimulus onset, the trial timed out and the next trial began. Participants' judgments and response times were recorded.

Participants were presented with the same types of stimuli as in Exps. 1a and 1b. Because this experiment was conducted in a controlled laboratory setting rather than over the web, more data was collected from each participant. However, even in the lab setting, collecting judgments from each participant for every quantifier / set size combination was not possible; that would have required 224 trials to collect a single data point for each quantifier / set size combination (14 set sizes and the 16 quantifiers from Exp. 1b). Instead, a subset of the space of quantifier / set size combinations was sampled for each participant. Each participant received 136 trials. Of those, 80 were the same across participants and represented the quantifier / set size combinations that were of most interest. (see Table 3.1).

The remaining 56 trials were pseudo-randomly sampled combinations of quantifier and set size, with only one trial per sampled combination. Trials were sampled as follows. For *some* and *summa*, four set sizes were randomly sampled from

Table 3.1: Distribution of the 80 trials each participant saw in Exp. 2 over quantifiers and set sizes.

	Set size								
Quantifier	0	1	2	3	4	10	11	12	13
some	10	2	2	1	1	1	1	2	4
summa	10	2	2	1	1	1	1	2	4
none	8	2	2						
one		4							
two			4						
all							2	2	8

the mid range (5 - 8) gumballs. For *alla*, four set sizes were randomly sampled from 0 to 10 gumballs. For *nunna*, four set sizes were randomly sampled from 3 to 13 gumballs. For both *one* and *two*, four additional incorrect set sizes were sampled, one each from the small set range (1 - 4 gumballs, excluding the correct set size), the mid range (5 - 8 gumballs), the high range (9 - 12 gumballs) and one of 0 or 13 gumballs. Finally, four additional number terms were sampled (one each) from the sets *three* or *four*, *five* to *seven*, *eight* or *nine*, and *ten* to *twelve*. This ensured that number terms were not all clustered at one end of the full range. Each number term occurred four times with its correct set size and four times with an incorrect number, one each sampled from the small set range, the mid range, the high range, and one of 0 or 13 gumballs, excluding the correct set size. For example, *three* could occur 4 times with 3 gumballs, once with 4, once with 7, once with 11, and once with 13 gumballs. The reason so many false number trials were included was to provide an approximate balance of YES and NO responses to avoid inducing an overall YES bias that might influence participants' response times.

To summarize, there were 28 *some* and *summa* trials each, 16 *alla* trials, 16 *nunna* trials, 8 *one* trials, 8 *two* trials, and 8 trials each for four additional number terms. Of these, 64 were YES trials, 60 were NO trials, and 12 were

critical trials - cases of *some/summa* used with the unpartitioned set, where they were underinformative, and *some/summa* used with one gumball, where a NO response is expected if the statement triggers a plural implicature (*at least two gumballs,* Zweig, 2009) and a YES response if it does not. Finally, there were three different versions of each image, with slightly different arrangements of gumballs to discourage participants from forming associations between quantifiers and images.

3.4.2 Results

A total of 6392 responses were recorded. Of those, 26 trials were excluded because participants did not respond within the four seconds provided before the trial timed out. These were mostly cases of high number terms occurring with big set sizes that required counting (e.g., 11 *eleven* trials, 8 *twelve* trials, 5 *ten* trials). An additional 33 cases with response times above or below 3 standard deviations from the grand mean of response times were also excluded. Finally, 254 cases of incorrect responses were excluded from the analysis. These were mostly cases of quantifier and set size combinations where counting a large set was necessary and the set size differed only slightly from the correct set size (e.g., *ten* used with a set size of 9). In total, 4.9% of the initial dataset was excluded from the analysis. The results are organized as follows. First, the proportion of YES and NO responses are reported, focusing on the relationship between response choice and the naturalness ratings. I then turn to the relationship between response times and naturalness ratings, testing the predictions I outlined earlier. Finally, I examine judgments and response times for pragmatic and semantic responses, relating the present results to earlier work by Bott and Noveck (2004) and Noveck and Posada (2003).

Judgments

All statistical analyses were obtained from mixed effects logistic regressions predicting the binary response outcome (YES or NO) from the predictors of interest. All models were fitted with the maximal random effects structure with by-participant random slopes for all within-participant factors of interest unless mentioned otherwise, following the guidelines in Barr, Levy, Scheepers, and Tily (2013).

I first examine the unpartitioned *some*/*summa* conditions, which are functionally equivalent to the underinformative conditions from Noveck and Posada (2003) and Bott and Noveck (2004). Recall that under a semantic interpretation of *some* (*You got some, and possibly all of the gumballs*), participants should respond YES when they get all of the gumballs, while a pragmatic interpretation (*You got some, but not all of the gumballs*) yields a NO response. The judgment data qualitatively replicates the findings from the earlier studies: 100% of participants' responses to *all* were YES, compared to 71% YES responses to partitive *some*. Judgments for simple *some* were intermediate between the two, with 82% YES responses. The difference between *some* and *summa* was significant in a mixed effects logistic regression predicting the binary response outcome (YES or NO) from a dummy-coded quantifier predictor. The log odds of a YES response were lower for *summa* than *some* (β = -1.18, SE = 0.34, $p < .001$), reflecting the naturalness results obtained in Exps. 1a and 1b, where *summa* was judged as less natural than *some* when used with the unpartitioned set. Thus both the word *some* and its use in the partitive construction increase the probability and/or strength of generating an implicature.

Response time analysis of naturalness effects on YES responses

Response times ranged from 577 to 3574ms (mean: 1420ms, SD: 444ms). Results of statistical analyses were obtained from mixed effects linear regression models predicting log-transformed response times from the predictors of interest. As with the judgment data, all models were fitted with the maximal random effects structure with by-participant random slopes for all within-participant factors of interest. Reported p-values were obtained by performing likelihood ratio tests, in which the deviance (-2LL) of a model containing the fixed effect is compared to another model without that effect that is otherwise identical. This is one of the procedures recommended by Barr et al. (2013) for models containing random correlation parameters, as MCMC sampling (the approach recommended by Baayen, Davidson, and Bates (2008) is not yet implemented in the R lme4 package.

For YES responses at the unpartitioned set, quantifier was Helmert-coded. Two Helmert contrasts over the three levels of quantifier were included in the model, comparing each of the more natural levels against all less natural levels (*alla* vs. *some*/*summa*, *some* vs. *summa*). YES responses to *alla* were faster than to *some* and *summa* (β = -.26, SE = .02, $p < .0001$) and YES responses to *some* were faster than to *summa* (β = -0.09, SE = .03, $p < .001$). YES responses to *some* and *summa* were slower for the unpartitioned set than in the most natural range determined in Exps. 1a and 1b, 4 - 7 gumballs (β = .09, SE = .03, $p < .01$).

A similar pattern holds in the small set range for the comparison between *some*/*summa* and number terms: responses to both *some* (β = 0.12, SE = .02, $p < .0001$) and *summa* (β = .12, SE = .02, $p < .0001$) were slower than number terms. Response times in the small set range did not differ for *some* and *summa*, as determined in model comparison between a model with and without a quantifier predictor (β = .02, SE = .02, $p < .25$), so I collapse them for further analysis.

There was a main effect of set size in the small set range: responses were

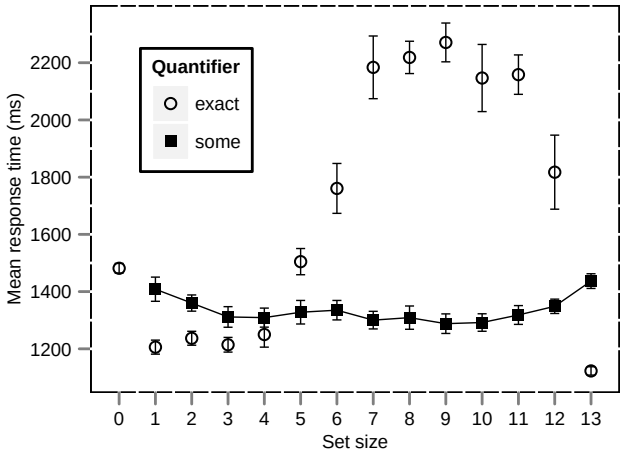

Figure 3.7: Mean response times of YES responses to *some* (collapsed over simple and partitive use) and for exact quantifiers and number terms. For exact quantifiers, only the response time for their correct cardinality is plotted.

faster as set size increased (β = -0.02, SE = .009, $p < .05$). The interaction between set size and quantifier (number term vs. *some*) was also significant (β = -0.08, SE = .02, $p < .0001$), such that there was no difference in response times for number terms used with different set sizes in the subitizing range, but response times decreased for *some/summa* with increasing set size. That is, the difference in response time between *some/summa* and number terms is largest for one gumball, somewhat smaller for two gumballs, and smaller still for three gumballs. This mirrors the naturalness data obtained in Exps. 1a and 1b (see Figure 3.5). Comparing response times for *some/summa* in the small set range to those in the preferred range (4 - 7), the results are similar: responses in the small set range are slower than in the preferred range (β = 0.04, SE = .01, $p < .05$). Mean response times for YES responses are shown in Figure 3.7 (response times for *some* and *summa* are collapsed because they did not differ).

Analyzing the overall effect of naturalness on response times (not restricted to *some/summa*) yields the following results. The Spearman r between log-transformed response times and mean naturalness for a given quantifier and set

size combination was -0.1 for YES responses overall (collapsed over quantifier). This value increased to -0.3 upon exclusion of cases of number terms used outside the subitizing range, where counting is necessary to determine set size. This correlation was significant in a model predicting log-transformed response time from a centered naturalness predictor and a centered control predictor coding cases of number terms used outside the small set range as 1 and all other cases as 0. The main effect of naturalness was significant in the predicted direction, such that more natural combinations of quantifiers and set sizes were responded to more quickly than less natural ones (β = -.04, SE = .01, $p < .0001$). In addition, a main effect of the control predictor revealed that number terms used outside the subitizing range were responded to more slowly than cases that don't require counting (β = .44, SE = .02, $p < .0001$).

Judgments and YES response times analyzed by pragmatic and semantic responders

Table 3.2 shows the distribution of participants over number of semantic responses to *some/summa* at the unpartitioned set. Noveck and Posada (2003) and Bott and Noveck (2004) found that individual participants had a strong tendency to give mostly pragmatic or mostly semantic responses at the unpartitioned set.[6] Therefore they conducted sub-analyses comparing pragmatic and semantic responders. Participants in the gumball paradigm were less consistent. Rather than two groups

[6]Both Noveck and Posada (2003) and Bott and Noveck (2004) found that participants tended to respond YES or NO consistently to underinformative items. For example in Noveck and Posada (2003)'s study there was one group of 7 consistently semantic responders (37%) and another group of 12 consistently pragmatic responders (63%). Similarly, in Bott and Noveck (2004)'s Study 3, 41% of responses to underinformative items were semantic while 59% were pragmatic. However participants within these categories were not always entirely consistent. Noveck and Posada note that their semantic responders were 96% consistent in their answers, while their pragmatic responders were 92% consistent. That is, some responders classified as semantic also gave some pragmatic responses and vice versa. Similarly, Bott and Noveck found that of their 32 participants, only 9 participants gave entirely consistent answers (2 semantic, 7 pragmatic). The remaining 23 participants gave both types of answers.

Table 3.2: Distribution of participants over number of semantic responses given in Exp. 2.

Number of semantic responses	0	1	2	3	4	5	6	7	8
Number of participants	2	2	2	1	2	5	6	6	21

of responders clustered at either end of the full range, there was a continuum in participants' response behavior, with more participants clustered at the semantic end. 42% of the participants gave 100% (8) semantic responses. Dividing participants into two groups, semantic and pragmatic responders, where pragmatic responders are defined as those participants who gave pragmatic responses more than half of the time and semantic responders as those who responded semantically more than half of the time yields a large group of semantic responders (38 participants, 81%) and a smaller group of pragmatic responders (7 participants, 15%). Two participants (4%) gave an equal number of semantic and pragmatic responses.

Given the nature of the distribution, rather than analyzing response times for semantic and pragmatic responders separately, I included a continuous predictor of responder type (degree of "semanticity" as determined by number of semantic responses) as a control variable in the analyses. That is, a participant with 1 semantic response was treated as a more pragmatic responder than a participant with 5 semantic responses. I analyzed the effect of (continuous) responder type on the response time effects reported above, specifically a) the naturalness effect at the unpartitioned set and b) the naturalness effect for small sets. First, centered continuous responder type was included as an interaction term with the Helmert contrasts for quantifier (*alla* vs. *some*/*summa*, *some* vs. *summa*) for the unpartitioned set. In this model both interactions were significant: the difference between *some* and *summa* was more pronounced for more pragmatic responders ($\beta = 0.05$, $SE = .02$, $p < .001$), while the difference between *alla* and *some*/*summa* was sig-

nificantly different for different responder types ($\beta = 0.04$, $SE = .01$, $p < .001$) but seems to be better accounted for by participants' consistency (see below). This suggests that more pragmatic responders are more sensitive to the relative naturalness of simple and partitive *some* used with an unpartitioned set. I return to this finding in the discussion.

An analysis of responder type for the naturalness effects for small set sizes yielded no significant results. That is, responder type did not interact with the quantifier by set size interaction reported above.

There was also an interesting correlation between participants' response behavior to *some/summa* used with the unpartitioned set (upper bound) versus when it was used for one gumball (lower bound). At the lower bound, the sentences *You got some (of the) gumballs* are strictly speaking true but often trigger a multiplicity implicature, whereby the cardinality of the set denoted by *some (of the)* gumballs is expected to be greater than one (Zweig, 2009). Response behavior to *some* at the lower bound was similar to that at the upper bound: 52% of responses were YES responses, while 48% were NO responses. Of these, 38 participants (81%) were completely consistent, 8 participants (17%) gave one response that was different from the rest, and only one participant gave 50% pragmatic and 50% semantic responses at the lower bound. Interestingly, 88% of participants who responded pragmatically at the upper bound also responded pragmatically at the lower bound. In addition, 40% of participants who had responded semantically or inconsistently at the upper bound responded pragmatically at the lower bound. That is, most participants who responded pragmatically at the upper bound also responded pragmatically at the lower bound, and most people who responded semantically at the lower bound also responded semantically at the upper bound (see Figure 3.8). The correlation between the number of semantic responses a participant gave at the upper bound vs. at the lower bound is .45.

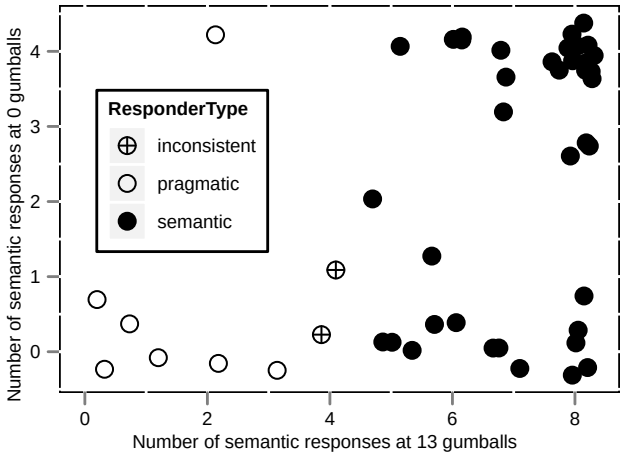

Figure 3.8: Scatterplot of participants' response behavior at the upper and lower bound. Each point represents one participant.

Response time analysis for semantic vs. pragmatic responses to *some*/*summa*

Recall that the Default model predicts pragmatic NO responses to *some*/*summa* with the unpartitioned set to be faster than semantic YES responses, while the reverse is the case for the Literal-First hypothesis. The latter has found support in a similar sentence verification task as the one reported here (Bott & Noveck, 2004). I thus attempted to replicate Bott and Noveck's finding that pragmatic NO responses are slower than semantic YES responses. To this end I conducted four different analyses. All were linear regression models predicting log-transformed response time from response and quantifier predictors and their interaction (the interaction terms were included to test for whether NO and YES responses were arrived at with different speed for *some* and *summa*): in model 1, all YES responses were compared to all NO responses. Model 2 was the Bott and Noveck between-participants analysis comparing only responses from participants who responded entirely consistently to *some*/*summa* (i.e. either 8 or 0 semantic responses in total). In a very similar between-participants analysis, response times from participants who responded entirely consistently to either *some* or *summa*

(i.e. either 4 or 0 semantic responses to either quantifier) were compared in model 3. Finally, again following Bott and Noveck, response times to YES and NO responses were compared within participants, excluding the consistent responders that entered model 2 from model 4. Results are summarized in Table 3.3.

The interaction between quantifier and response was not significant in any of the models, suggesting there was not a difference between *some* and *summa* in the speed with which participants responded YES or NO. The main effect of quantifier reached significance in both model 1 (all responses to *some*/*summa* at unpartitioned set) and model 4 (including only inconsistent responders), such that responses to *summa* were generally slower than those to *some* (see Table 3.3 for coefficients). Finally, the main effect of response was marginally significant in models 2 and 3 (including only consistent responders, either overall or within quantifier condition), such that YES responses were marginally faster than NO responses.

Response times as a function of response inconsistency

I conducted a final response time analysis that was motivated by the overall inconsistency in participants' response behavior at the unpartitioned set. Rather than analyzing only how participants' degree of semanticity impacted their response times, I analyzed the effect of within-participant response inconsistency on response times. Five levels of inconsistency were derived from the number of semantic responses given. Participants with completely inconsistent responses (4 semantic and 4 pragmatic responses) were assigned the highest inconsistency level (5). Participants with a 3:5 or 5:3 distribution were assigned level 4, a 2:6 or 6:2 distribution were assigned level 3, a 1:7/7:1 distribution level 2, and a 0:8/8:0 distribution (participants who gave only semantic or only pragmatic responses) level 1. There is a clear non-linear effect of inconsistency on YES responses for the unpartitioned set (see Figure 3.9).

Table 3.3: Model coefficients (β), standard error (SE), t-value, and p-value for the three predictors (quantifier, response, and their interaction) in each of the four different models testing whether NO responses are slower than YES responses.[a]

		Quantifier (*some*, *summa*)				Response (NO, YES)				Quantifier:Response			
Model	Obs.	β	SE	t	p	β	SE	t	p	β	SE	t	p
1. Overall	375	0.06	0.02	2.69	$<.05$	0.01	0.04	0.25	0.86	-0.05	0.06	-0.88	0.23
2. Consistent	184	0.03	0.03	0.95	.38	-0.19	0.15	-1.25	<0.09	0.23	0.1	0.24	0.83
3. Consistent within quantifier	256	0.01	0.03	0.50	.92	-0.12	0.08	-1.51	<0.06	-0.02	0.09	-0.23	0.52
4. Inconsistent	191	0.1	0.04	2.78	$<.05$	0.03	0.04	0.76	.53	-0.04	0.08	-0.56	.33

[a]**Overall:** model comparing all YES responses to all NO responses. **Consistent:** model comparing YES and NO responses of completely consistent responders. **Consistent within quantifier:** model comparing YES and NO responses of participants who gave completely consistent responses within either the *some* or the *summa* condition at the unpartitioned set. **Inconsistent:** model comparing YES and NO responses of participants who gave at least one inconsistent response.

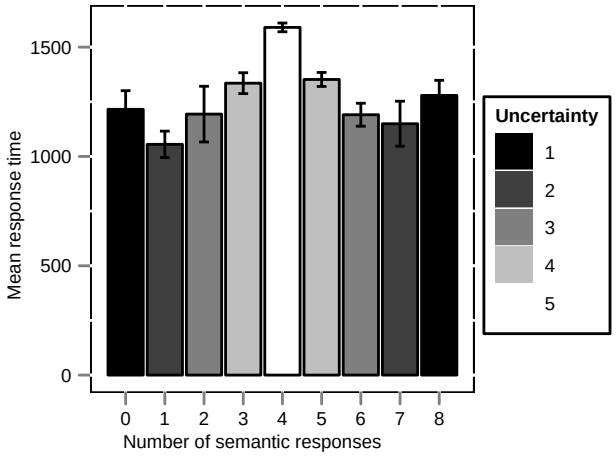

Figure 3.9: Mean response times to YES responses at the unpartitioned set (collapsed across quantifiers) as a function of responder inconsistency.

Model comparison of models with polynomial terms of different orders for inconsistency and their interactions with naturalness reveal that there is a significant main effect of naturalness as observed earlier (β = -0.07, SE = .01, $p <$ .0001), a significant effect of the second-order inconsistency term (β = 0.04, SE = .02, $p <$ .01), and a significant interaction of naturalness and the first-order inconsistency term (β = -0.02, SE = .01, $p <$ .001). That is, those participants who responded most inconsistently also responded most slowly to *some*, *summa*, and *alla* when used with the unpartitioned set. Response times decreased as inconsistency increased, but both completely pragmatic and completely semantic responders showed a slight but significant increase in response times (as revealed in the significant second-order term).

As with for responder type, the effect of inconsistency on the naturalness effect at the unpartitioned set was analyzed by including centered inconsistency level as interaction terms with the Helmert contrasts for quantifier (*alla* vs. *some*/*summa*, *some* vs. *summa*).

This yields a highly significant interaction of inconsistency with the *alla* vs.

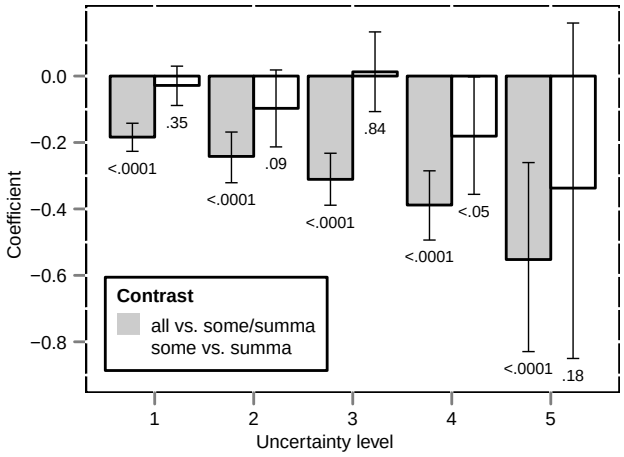

Figure 3.10: Model coefficients obtained from models predicting YES response times from Helmert-coded quantifier contrasts at the unpartitioned set. Coefficients are plotted as a function of responder consistency. Numbers below the bars indicate that coefficient's p-value.

some/summa contrast (β = -0.07, SE = .01, $p < .001$) and a weak trend for the interaction with the *some* vs. *summa* contrast (β = -0.03, SE = .02, $p < .17$) in the same direction. The effect of the difference between the different levels of the quantifier predictor on log response times is larger for more inconsistent responders. This is visualized in Figure 3.10, where the coefficients of the two contrasts are plotted for models run on subsets of the data depending on inconsistency level. As with responder type, letting the inconsistency term interact with quantifier (*some/summa* or number) for small sets replicates the effects reported previously, but there are no significant interactions with inconsistency.

3.4.3 Discussion

In Exp. 2, the mean naturalness of the quantifier *some* for different set sizes predicted response times. In particular, YES responses were processed more slowly with decreasing naturalness. This effect was particularly strong where the more natural alternative was very rapidly available due to subitizing pro-

cesses (e.g., *two*). In addition, responders who gave many pragmatic responses to *some/summa* were more sensitive to the naturalness difference between *some* and *summa* when used with the unpartitioned set, whereas less consistent participants were more sensitive to the naturalness difference between *alla* and *some/summa*. The main result of Bott and Noveck's, namely that pragmatic NO responses to *some/summa* are slower than semantic ones, was replicated marginally only in a subset of the conducted analyses. I discuss the responder type effects and the implications of the partial replication of Bott and Noveck's results in more detail here and defer discussion of the more general theoretical implications to Section 3.5. The effect of response inconsistency on response times will be interpreted as uncertainty about the QUD and will be discussed in detail in Section 3.6.

Varying naturalness effects for semantic/pragmatic responders

An interesting difference in responder type was that participants who gave more pragmatic responses to *some/summa* at the unpartitioned set were more sensitive to the naturalness difference between *some* and *summa*; when compared to semantic responders, they responded YES more slowly to *summa* relative to *some*. Why do participants who respond more pragmatically exhibit a stronger naturalness effect than participants who respond more semantically?

One possibility is that there was a difference in the participant population between listeners who are generally more sensitive to naturalness differences between different utterance alternatives and those less sensitive to naturalness differences. The more sensitive ones should have responded more pragmatically overall because the naturalness of *some/summa* is generally rated lower than the naturalness of all for the unpartitioned set. Therefore, participants who are more sensitive to naturalness are likely to give more pragmatic responses. It is plausible to assume that participants who are more sensitive to the naturalness difference between *alla* and *some/summa* with the unpartitioned set (as reflected in judgments), are

also more sensitive to the difference between *some* and *summa*. If this difference is reflected in response times, it would result exactly in the pattern of results observed.

If this is the right explanation for the observed difference in naturalness effects, it raises a number of interesting questions about individual differences in pragmatic inference. Was this an experiment-specific difference? For example, perhaps some participants were more attentive than others? Or are there population-level individual differences in how attuned to alternative utterances listeners are in general? For example, individuals vary in executive function, which among other things includes ability to hold and evaluate multiple alternatives in memory (e.g., Miller & Cohen, 2001; Novick, Trueswell, & Thompson-Schill, 2005). This would predict a correlation between participants' executive function and their propensity to respond pragmatically in the gumball task. Exploring individual differences in scalar implicature processing is an interesting avenue for future research.

Implications of semantic/pragmatic response time analysis

Recall that Bott and Noveck (2004) found that pragmatic responses to *some* were slower than semantic responses both between and within participants. They interpreted this as evidence that the interpretation of the semantic *some and possibly all* interpretation comes for free and is more basic than the *some but not all* interpretation, which is assumed to be computed only if the semantic interpretation does not suffice to meet expectations of relevance in context and incurs additional processing cost.

I partly replicated this result: in the analysis comparing semantic and pragmatic responses between participants who were entirely consistent in their judgments to *some*/*summa* used with the unpartitioned set, semantic responses were marginally faster than pragmatic responses. The weakness of the effect may have been due to data sparseness as there were relatively few pragmatic responders

overall. However, the result did not hold up within participants, where more data were available. Moreover, for all set sizes, consistently pragmatic responders were marginally slower to respond to *some*/*summa* than consistently semantic responders. Proponents of the Literal-First hypothesis might interpret this as compatible with their claims: by generally taking more time to reach an interpretation of an utterance containing a quantifier, pragmatic responders could be giving themselves the time necessary to generate the scalar inference once they encounter *some* with an unpartitioned set. Some support for this interpretation comes from Bott and Noveck (2004), who showed that participants were more likely to draw a scalar inference when they were given more time to make their judgment. Similarly, De Neys and Schaeken (2007) found that placing participants under increased cognitive load led to fewer scalar implicatures.

However, there is an alternative interpretation. The delay might not be caused by the computation of an interpretation, but rather by the time it takes to verify that interpretation in the relevant context. That is, pragmatic responders might be investing more effort in carefully evaluating whether the observed utterance is true in the visual context. These two different explanations cannot be teased apart in this dataset, but we will see in Exps. 4a and 4b that semantic and pragmatic responders have different verification strategies and that the difference in response time replicates even when participants' eye movements reflect that they have rapidly generated a scalar implicature.

3.5 Discussion of Exps. 1 and 2

These results provide a challenge for both of the most influential approaches to the processing of scalar implicatures. According to the Literal-First hypothesis the upper-bound interpretation of *some* is computed only after the lower-bound interpretation, predicting that NO responses to *some* with the unpartitioned set

should always be slower than YES responses. This was indeed the marginally significant result observed when comparing YES and NO responses between (entirely consistent) participants. Crucially, however, it was not reliable within participants. This cannot be attributed to lack of power. There were three times as many NO judgments in the within-participants analysis compared to the between-participants analysis. Moreover, we cannot rule out the possibility that participants took longer to respond pragmatically than semantically because they are more careful in verifying the utterance in context. The fact that pragmatic responders are generally slower (even in YES responses) than semantic responders is more consistent with a verification hypothesis.

The Literal-First account has trouble explaining the naturalness effects observed for the YES responses, in particular the slower response time for YES responses to *some/summa* at the unpartitioned set compared to the preferred range. The Literal-First hypothesis predicts that as soon as the first stage of semantic processing is complete and the lower bound verified, a YES response can be made. Thus, according to this account, there should be no difference in YES response times to *some/summa* from 1 to 13 gumballs, where the semantic interpretation holds. Yet the results revealed clear response time differences.

Proponents of the Literal-First hypothesis might argue that the slowdown effect at the unpartitioned set is due to participants having initially computed the semantic interpretation and then the pragmatic interpretation, before reverting to the semantic interpretation. That is, rather than the semantic response itself taking longer to compute, it is the integration with context that might result in an intermediate computation and subsequent cancelation of the implicature that led to the delayed YES response. However, this type of explanation cannot account for the slowdown effects observed for the small sets. For small sets, YES responses to *some/summa* were again slower than responses to number terms. Moreover, the difference in response times was a function of their naturalness relative to number

terms. Thus the more parsimonious explanation is that the effects in both cases arise from the same underlying cause: there is a more natural alternative to *some* that the speaker could have used, but didn't.

These results are also incompatible with the Default model, which assumes that scalar implicatures are computed by default upon encountering *some*. Like the Literal-First hypothesis, the Default approach cannot account for effects of naturalness. Under the Default model, contextual factors such as the naturalness of an utterance with a particular quantifier should only come into play when the participant is deciding whether to cancel the already computed implicature. That is, YES responses to the unpartitioned set should be the only set size where naturalness of *some* affects response times. To see this, consider what interpretation is necessary for each set size to arrive at a YES response. For the unpartitioned set, only the semantic *some and possibly all* interpretation yields a YES response. In contrast, for set sizes 1 - 12, both the semantic and the pragmatic interpretation yield YES responses. Thus, under the Default model YES responses to *some* should have been equally long for set sizes 1 - 12, and a slowdown is expected only for the unpartitioned set, where an additional cost for canceling the implicature is incurred.

The Default account correctly predicts the slowdown effect for YES responses to *some* with the unpartitioned set compared to its most preferred range. However, it cannot explain a) the slowdown effect for small sets and b) the fact that pragmatic NO responses to the unpartitioned set are also slower compared to YES responses in the preferred range.

The pattern of results obtained in Exps. 1 - 2 is thus most compatible with a Constraint-Based account in which the speed and robustness of an implicature is determined by the probabilistic support it receives from multiple cues available in the linguistic and discourse context, including the task/goal relevant information. In Exps. 1 - 2 I investigated the effect of alternatives on processing *some* and the

scalar inference from *some* to *not all*. In particular, I investigated a) the syntactic partitive as a cue to the implicature and b) the naturalness and availability of lexical alternatives to *some* as inhibitory cues to arriving at an interpretation for *some*.

Some observations on the naturalness/availability of lexical alternatives are in order. While linguists and logicians treat the meaning of *some*, *all*, and other quantifiers as set-theoretically well-defined as for example in Generalized Quantifier Theory (Barwise & Cooper, 1981), the naturalness results from Experiments 1 and 2 show that not only are quantifiers like *some* more natural for some set sizes than others, but that their naturalness depends crucially on alternative lexical items that the speaker might have used. That is, while there may be ways to unambiguously define quantifier meanings, quantifier interpretation is a matter of degree. Some set sizes (e.g. 6 or 7 out of 13 gumballs) are better fits for *some* than others (e.g., 2 or 13 out of 13 gumballs) and introducing alternatives can change the goodness of fit. This is paralleled in the concepts and categories literature, where typicality influences categorization even for logically well-defined concepts such as parity (Armstrong, Gleitman, & Gleitman, 1983). Some numbers are judged as better instances of an odd or an even number than others. Similarly, in the gumball studies *some* is judged as a more or less natural label for different sizes (and processed more or less quickly accordingly), despite its logical definition unambiguously assigning a TRUE or FALSE value for any given set size.

The effects of naturalness and availability are also compatible with accounts that treat the meaning of quantifiers as distributions over quantities. These distributions reflect listener beliefs about the probability of a particular quantifier being uttered given a particular set size. Listener beliefs are updated by different contextual factors, among them the availability of alternatives. That is, the distributions may shift contextually depending on the available alternatives. Mapping this onto the naturalness results for small sets: the posterior probability of ob-

serving an utterance of *some* to refer to a set of size 2 given that number terms are contextually available is lower than when they are not.

Methodologically, this means that it is important for researchers to be aware of the relative naturalness of the quantifiers under investigation, i.e., which range of the interpretive distribution one is sampling from. For example, in light of the observed intrusion effects, the delays in the processing of *some* observed by Huang and Snedeker (2009) could have an explanation that is unrelated to the processing of implicatures per se: not only did they use a set size that is relatively unnatural for *some* (2), but *two* was also explicitly included as a lexical alternative among the experimental items. Therefore *some* may have been a dispreferred label for referring to the set size it was used for, thus causing the delay. Note that Huang and Snedeker did not find a similar delay for *all*,[7] which was used with a larger set size (3). The naturalness ratings obtained here predict a delay for *all* if it had been paired with a set size of 2 (where it is relatively unnatural) and a smaller delay for *some* if it had been paired with a larger set size (where the difference in naturalness between *some* and number terms is smaller). These predictions are tested in Exps. 4a and 4b.

In addition, Exps. 4a and 4b address the question whether the availability of more natural alternatives to *some* affects the earliest moments of scalar implicature processing using eye movements, a measure of listeners' interpretation that is more closely time-locked to the interpretive process than response times. However, before reporting these studies, Section 3.6 examines in more detail the extent to which uncertainty about the contextual QUD can explain an interesting response time pattern observed in Exp. 2.

[7]Though they do report a delay for *all* relative to *three* in their Exp. 3, which they explain by reference to the increased difficulty involved in appropriate domain restriction in this experiment.

3.6 Exp. 2a: relevance effects

3.6.1 Response inconsistency as uncertainty about the Question Under Discussion

One of the effects observed in Exp. 2 merits its own independent discussion: responders who were more inconsistent in their responses at the upper bound exhibited a stronger sensitivity to the difference between *alla* and *some/summa*, and they were slower to respond than more consistent responders. Why? In the explanation I draw upon the notion of the Question Under Discussion (QUD, Roberts, 1996, 2004), which is becoming increasingly influential within formal pragmatics. The QUD is the question that interlocutors are trying to answer at the point of an utterance. Depending on what the QUD is, an utterance of the same sentence in different contexts may lead to different implicatures. Zondervan (2010) has shown that the QUD can affect how reliably an implicature is drawn: more scalar implicatures are generated when the triggering item (e.g. *or* or *most*) is in a focused constituent that addresses the QUD than when it is in an unfocused constituent that does not.

Many researchers have noted that scalar implicatures arise only when the stronger alternative is contextually relevant (e.g., Carston, 1998; Green, 1995; Levinson, 2000; Matsumoto, 1995). Thus, for example, (41) in response to (40a) is taken to implicate that not all of their documents are forgeries, whereas the same implicature does not arise if (41) is uttered in response to (40b) (Levinson, 2000). The stronger alternative is assumed to be relevant to the QUD in (40a), but not to the QUD in (40b).

(40) a. Are all of their documents forgeries?

b. Is there any evidence against them?

(41) Some of their documents are forgeries.

Similarly, there are many different potential QUDs that the utterance *You got some of the gumballs* might be interpreted relative to, as is illustrated in the examples in (42):

(42) a. Did I get all of the gumballs?

b. How many gumballs did I get?

c. Did I get any of the gumballs?

Intuitively, *You got some of the gumballs* implicates that you did not get all of them if it is an answer to the question in (42a), but not to the question in (42c), and only weakly to the question in (42b).

In a particular context, the QUD is sometimes established explicitly, e.g., by asking a question. However, the QUD is often a matter of negotiation between interlocutors. That is, at a particular point in discourse interlocutors may have uncertainty about the actual QUD. One way to view the inconsistency results above is that participants had different amounts of uncertainty about the QUD when they observed an utterance of *You got some (of the) gumballs* with the unpartitioned set. Under this view, participants who consistently responded either semantically or pragmatically had little uncertainty about the QUD: semantic responders consistently adopted the QUD in (42c) whereas pragmatic responders consistently adopted the QUD in (42a). The distributions of participants' responses who were intermediate between the two extremes reflect their increased uncertainty about the actual QUD and it may have been this uncertainty that resulted in slower response times. This explanation is also consistent with the result that more inconsistent/uncertain responders are more sensitive to the difference between *alla* and *some*/*summa* than to the difference between *some* and *summa*: *You got all of the gumballs* is true for the unpartitioned set regardless of the QUD, thus the difference between more and less certain responders should be small in their response times to *alla*, and it was. Uncertainty about the QUD

should have a much larger effect on responses to *some/summa*, as the interpretation (and consequently the truth or falsity) of *You got some (of the) gumballs* with the unpartitioned set depends on the QUD. Greater uncertainty about the QUD should result in slower responses. This is just the pattern of results obtained in Exp. 2.

This appeal to the uncertainty about the QUD to explain the observed response time patterns is necessarily post-hoc because the QUD was not explicitly manipulated in Exp. 2. Future research should investigate the effect of QUD uncertainty on scalar implicature processing by explicitly manipulating the (implicit or explicit) QUD that statements containing scalar items are interpreted relative to. The first step in this enterprise is to test whether the QUD modulates implicature rates and response times at all in the gumball paradigm.

3.6.2 Manipulating the relevance of the stronger alternative in the gumball paradigm

Exp. 2a[8] manipulated the contextual relevance of the stronger alternative *You got all of the gumballs* by varying the cover story presented to different groups of participants. Each cover story set up a different salient, but implicit, QUD. The two QUDs are shown in (43).

(43) Implicit QUDs in the *relevant* vs. *less relevant* conditions

a. Did I get all of the gumballs?

b. Did I get none of the gumballs?

In the *relevant* condition, where the implicit QUD was (43a), the stronger alternative *You got all of the gumballs* was highly relevant. In the *less relevant* condition, where the implicit QUD was (43b), the stronger alternative was much

[8]This experiment was conducted in collaboration with Laurel Raymond.

less relevant. Under the Constraint-Based account, the relevance of the stronger alternative should affect both the rate and the speed of the scalar implicature: in the *relevant* condition, higher implicature rates and faster implicatures are predicted than in the *less relevant* condition. I first briefly discuss the notion of relevance I assume before presenting the details of the study.

I follow Russell (2012)[9], who treats implicatures as varying in strength, and strength in turn depends (among other things) on the relative relevance of the stronger and weaker alternatives to a contextually salient QUD. Russell's notion of relevance is not a categorical one as assumed at the beginning of this section, but rather a gradient one: the relevance of a proposition E to a proposition H is relevant to the degree that observing the truth of E leads to a change in the probability with which H is believed. This notion of relevance was first formalized in Carnap (1950) and is thus aptly named Carnap relevance:

$$r_H(E) = p(H|E) - p(H) \tag{3.1}$$

In comparison to categorical notions of relevance, Carnap relevance captures the intuition that propositions are relevant to each other to varying degrees, rather than either being relevant or not.[10] Two crucial properties of Carnap relevance are that a) $r_H(E)$ is always a value between -1 and 1 and b) relevance is a symmetric relation (Russell, 2012). One of the consequences of this is that if E is positively relevant to H, it is negatively relevant to the same degree to $\neg H$.

Let us calculate the Carnap relevance of the strong proposition s (that you got all of the gumballs) and weak proposition w (that you got some of the gumballs)

[9]See Section 2.2.3 for a more detailed description of the account.

[10]Categorical notions of relevance take a proposition to be relevant to a QUD if that proposition provides a full or partial answer to the QUD (e.g., Roberts, 1996; van Kuppevelt, 1996; van Rooij & Schulz, 2004), and irrelevant otherwise. Note that under these notions of relevance, *You got some of the gumballs* is equally relevant to all of the QUDs in (42): it provides a partial answer to questions (42a) and (42b), and a complete answer to (42c). This notion of relevance can therefore not capture the intuitive relevance differences important to our situation.

in the gumball paradigm, given the QUDs proposed in (43). I will make one simplifying assumption in the calculation of the relevance values, which is based on Russell's observation that only assertions that are positively relevant to a conversational point being made ("a partition in the denotation of a question under discussion" Russell, 2012, p. 67), are felicitous. Given this observation, I will assume that H corresponds to the proposition that maximizes the Carnap relevance of s and w for each QUD in (43), respectively.[11] These propositions, which I will call *all* and $\neg$*none*, are shown in (44).

(44) Relevance-maximizing propositions in the *relevant* (a) and *less relevant* (b) condition

a. *all*: ⟦You got all of the gumballs⟧

b. $\neg$*none*: ⟦You got at least one of the gumballs⟧

We can now calculate the relevance values for s and w. To calculate each probability, recall that there are 14 states of the gumball machine; it can dispense 0, 1, ..., 13 gumballs. This yields the following relevance values for the QUD in (43b):

$$r_{\neg none}(s) = p(\neg none|s) - p(\neg none) = 1 - 13/14 = 1/14 = 0.07 \tag{3.2}$$

$$r_{\neg none}(w) = p(\neg none|w) - p(\neg none) = 1 - 13/14 = 1/14 = 0.07 \tag{3.3}$$

[11] This move is justified because both of the QUDs in (43) are polar questions and relevance is a symmetric relation. Therefore, s and w are equally positively relevant to one of the answers to the question as they are negatively relevant to the other answer. Because only positively relevant assertions are allowed, we can assume that the proposition that s and w are positively relevant to is the salient conversational point being addressed. I ignore here that it is perhaps dubious to assume that speaking gumball machines are in the habit of addressing conversational points.

What this means is that s is just as relevant as w to the implicit QUD in the *less relevant* condition. Compare this to the relevance values for the QUD in (43a):

$$r_{all}(s) = p(all|s) - p(all) = 1 - 1/14 = 13/14 = 0.93 \tag{3.4}$$

$$r_{all}(w) = p(all|w) - p(all) = 1/13 - 1/14 = 1/182 = 0.005 \tag{3.5}$$

Here, s is highly relevant, while w is almost irrelevant. The relative relevance of s compared to w is thus much higher when the QUD is whether you got all of the gumballs compared to when it is whether you got none of the gumballs. Exp. 2a tests the prediction that scalar implicature rates should be higher and response times reflecting a scalar implicature faster for the implicit QUD in (43a) compared to the implicit QUD in (43b).

3.6.3 Methods

Participants

Fifty-seven undergraduate students from the University of Rochester were paid $10.00 for participation.

Procedure and materials

The procedure was the same as in Exp. 2, but one group of participants was initially presented with a cover story that implicitly established the QUD in (43a) (rendering *s relevant*), while another group was presented with a cover story that implicitly established the QUD in (43b) (rendering *s less relevant*).

Participants in both conditions began by reading a story explaining that they were in a candy store full of gumball machines that verbally reported how many

gumballs had dropped to the lower chamber, though sometimes this statement was correct and sometimes it was incorrect. They were also told that there was a store worker who was monitoring the state of the gumball machines by listening to the machines' statements and taking note of participants' reactions to those statements. Participants were told that their task was to indicate whether they agreed or disagreed with the statements by pressing the YES or NO key. In the *relevant* condition, participants were told the store worker would be fired if he did not refill the machines, while in the *less relevant* condition they were told the store worker would be fired if he did not fix jammed machines.

Participants then went through a scripted demonstration that illustrated the store worker's reaction to various statements and responses. In the *relevant* condition, participants were shown a picture of the store worker refilling the gumball machine when their response indicated that the machine was empty, and in the *less relevant* condition participants were shown a picture of the store worker fixing the machine when their response indicated that the machine was not delivering gumballs. The consequence of the worker's firing was also illustrated; in the *relevant* condition, participants were told that one time someone responded such that the store worker did not know the upper chamber was empty and thus did not refill it. He consequently got fired. In the *less relevant* condition participants were told that the firing occurred when someone responded such that the store worker did not know the machine was not delivering gumballs and thus did not fix it. Crucially, participants never saw the situation in which the machine delivered all of the gumballs and they heard *You got some of the gumballs*, which I will refer to as the critical condition, during this demonstration.

In order to ensure that participants paid attention to the cover story, they were then asked a multiple choice comprehension question: "When will the store worker be fired?" In the *relevant* condition the correct answer was "when the machines are empty" and in the *less relevant* condition the correct answer was

Table 3.4: Distribution of experimental trials over quantifiers and set sizes in Exp. 2a.

	Set size						
Quantifier	0	2	5	8	11	13	**Total**
some	4	1	1	1	1	8	16
alla	2	1	2	1	2	8	16
nunna	4	1	0	1	1	1	8
number	3	7	7	7	5	3	32
Total	13	10	10	10	9	20	72

"when the machines jam". These and two filler answers were the available choices for all participants. If participants answered incorrectly, they were asked to reread the context until they answered the question correctly. Participants then began the experiment. Halfway through the experiment, participants had to answer the comprehension question again. This was done to prevent decay of the QUD over the course of the experiment.

There were 72 trials. On 32 of the trials the expected answer was YES and on 32 of the trials the expected answer was NO to avoid response bias. There were also 8 occurrences of the critical condition, in which all 13 of the gumballs dropped to the lower chamber and the participants were told they got some of the gumballs. As in Exp. 2, a NO response on a critical trial indicated a pragmatic interpretation while a YES response indicated a semantic interpretation. Number of expected YES and NO responses as well as well as number of critical trials was counterbalanced by quarters. Trial order was randomized within each quarter. Four different lists were created in this way; each list occurred once in the *relevant* and once in the *less relevant* condition. Table 3.4 shows how the 72 trials were distributed over quantifier/set size combinations. Note that in this experiment, only the critical condition is of theoretical interest; the remaining trials were fillers.

3.6.4 Results and discussion

One participant was excluded because they answered the second comprehension question incorrectly five times. The remaining participants all correctly answered the second comprehension question the first time. Two lists (seven participants on each list) were excluded because of an experimenter error, leaving 42 participants, 21 in each relevance condition. These exclusions had no effect on the qualitative pattern of judgment and response time results.

I first present the judgment data from the critical condition before presenting the response time results. In both cases, results from Exp. 2a are compared to the results from Exp. 2.

Judgments

In the *relevant* condition there were 92% pragmatic responses in the critical condition, compared to only 50% pragmatic responses in the *less relevant* condition. A mixed effects logistic regression predicting YES over NO responses with random by-participant intercepts revealed a highly significant main effect of relevance such that semantic responses were much less likely when the stronger alternative was *relevant* compared to when it was not (β = -5.15, SE = 1.33, $p < .0001$). The left panel of Figure 3.11 shows the proportion of semantic YES responses in the *relevant* and *less relevant* conditions alongside the proportion of semantic responses in Exp. 2. The graph shows clearly that not only are scalar implicatures more likely when the stronger alternative is relevant, but simply introducing a cover story in which the quantity of gumballs dispensed matters to somebody, also makes the implicature more likely. Including a three-level relevance factor with the *less relevant* condition as reference level confirms this difference: semantic responses are less likely in the *relevant* condition (β = -3.26, SE = 0.47, $p < .0001$) and more likely in Exp. 2 where the QUD was unclear (β = 1.51, SE = 0.27, $p < .0001$).

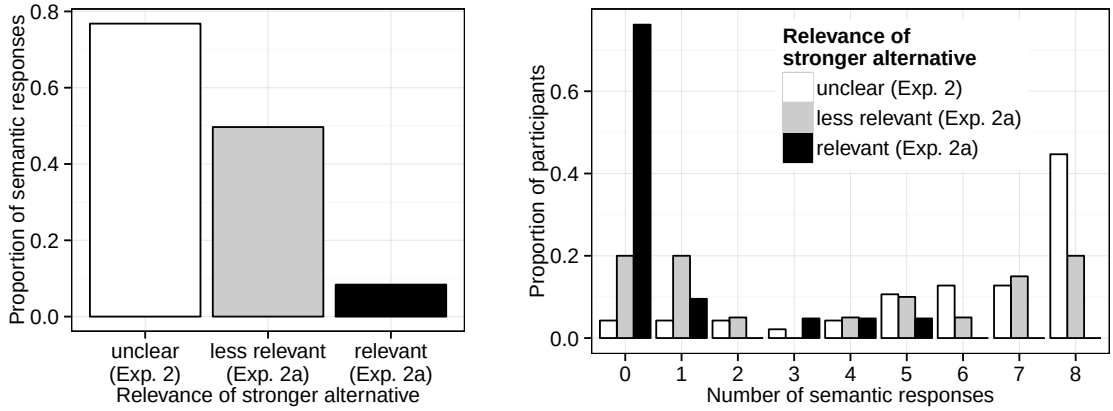

Figure 3.11: Proportion of pragmatic responses at the upper bound in Exp. 2 and both relevance conditions of Exp. 2a (left panel) and distribution of participants over number of semantic responses given at the upper bound in Exp. 2 and Exp. 2a (right panel).

The right panel of Figure 3.11 shows the distribution of participants over number of semantic responses given on critical trials.

Response times

The left panel of Figure 3.12 shows mean response times on critical trials in the *relevant* and *less relevant* condition alongside response times from Exp. 2. The question was whether response times for pragmatic and semantic responses would be differentially affected by the relevance of the stronger alternative. In particular, an interaction between relevance and response would constitute the strongest evidence that the relevance of the stronger alternative makes pragmatic interpretations (reflected in NO responses) easier to process and semantic interpretations harder to process. As can be seen in the left panel of Figure 3.12, this interaction seems indeed to be numerically present. However, it did not reach significance in a mixed effects linear regression model with random by-participant intercepts predicting log-transformed response time from fixed effects of relevance (*relevant* vs. *less relevant*), response (YES vs. NO), and their interaction. Predictors were

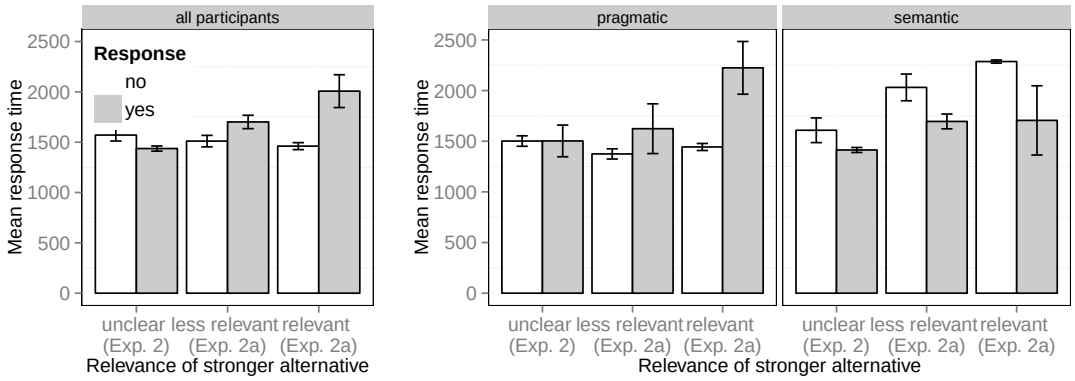

Figure 3.12: Mean response times for YES and NO responses in the critical condition as a function of the relevance of the stronger alternative. Left panel shows response times collapsed across all participants. Right panel shows response times for pragmatic and semantic responders separately (inconsistent responders excluded).

centered before entering the analysis. There was a main effect of response such that YES responses were slower than NO responses ($\beta = 0.08$, $SE = 0.05$, $t = 1.6$, $p < .05$); however, this effect disappeared once response inconsistency was taken into account [12] ($\beta = 0.05$, $SE = 0.05$, $t = 1.00$, $p < .3$).

While there was a numerical interaction in the predicted direction, we cannot conclude from these data that the relevance of the stronger alternative to a contextual QUD indeed modulates online implicature processing. However, that the interaction was in the predicted direction is encouraging and may point to a power problem.[13] One way to add power is to include the data from the critical condition of Exp. 2. But what are the predictions for Exp. 2? Recall that in Exp. 2, participants received no cover story. That is, in this Experiment, participants had greater prior uncertainty about the QUD. However, the judgment data show that participants in Exp. 2 were much more likely than even in the *less*

[12] See Section 3.4.2 for a discussion of the effect of response inconsistency on response times. This effect is replicated here: increased response inconsistency leads to slower responses ($\beta = 0.09$, $SE = 0.03$, $t = 3.6$, $p < .0001$).

[13] Recall that 14 participants had to be excluded from the analysis due to experimenter error - it is possible that with more participants, this interaction might receive further support.

relevant condition of Exp. 2a to interpret *some* semantically. Thus, participants seem to have more consistently adopted the QUD that the stronger alternative is not relevant to. But this predicts that response times from Exp. 2 should further contribute to the interaction predicted for the two relevance conditions of Exp. 2a: NO responses should be even slower than in the *less relevant* condition, while YES responses should be even faster. Indeed, this is what Figure 3.12 suggests. The following analysis supports this interpretation in part.

In a second analysis, the data from the critical condition of Exp. 2 (also shown in Figure 3.12) was included, where the QUD was not manipulated explicitly. This added an additional level *unclear* to the relevance predictor. Based on the gradient increase in response time for YES responses and decrease for NO responses apparent in the left panel of Figure 3.12, relevance was Helmert-coded such that each more relevant level was compared to the mean of the less relevant levels. In order of decreasing relevance, the levels were *relevant*, *less relevant*, *unclear*.

The full model output is shown in Table 3.5. There are two results of interest. First, there is a significant main effect of response such that YES responses were slower than NO responses, even after controlling for response inconsistency. This is interesting because it suggests that, once proper controls for participants' uncertainty about the QUD are considered, pragmatic responses may be as fast and faster than semantic responses, which runs counter to the bulk of the findings on scalar implicature processing thus far. Indeed, it suggests that experimenters should be very careful about how they interpret aggregated response data: if it so happens that the majority of pragmatic responses are generated by participants who are uncertain as to the QUD or the task they are engaged in, this may artificially inflate the time it takes to process the pragmatic interpretation.

Second, the main results of interest are the interactions between response and each of the two relevance predictors: the interaction with the relevant.vs.rest contrast was not significant, while the one with the lessrelevant.vs.unclear contrast

Table 3.5: Full mixed effects linear regression model for combined critical condition data of Exps. 2 and 2a.

	Coef β	SE(β)	**t**	p
Intercept	7.31	0.03	278.3	**<.0001**
Relevant.vs.Rest	0.07	0.05	1.4	<.14
Lessrelevant.vs.Unclear	0.08	0.03	2.8	**<.01**
Response	0.08	0.03	2.4	**<.05**
Response inconsistency	0.04	0.01	3.3	**<.001**
Relevant.vs.Rest:Response	0.09	0.09	1.0	<.3
Lessrelevant.vs.Unclear:Response	0.12	0.06	2.0	**<.05**

was. The latter can be seen clearly in the left pair of bars in the left panel of Figure 3.12, where the response time ratios of YES to NO responses are flipped between Exp. 2 and the *less relevant* condition of Exp. 2a. However, inspecting the simple effects model reveals that the relevance of the stronger alternative mainly affects YES responses; the more relevant the stronger alternative, the slower participants were to respond semantically at the upper bound. Response times for pragmatic responses, while again numerically in the predicted direction, are not significantly affected by the relevance of the stronger alternative.

One final post hoc analysis was conducted to investigate whether there were differential effects of the relevance of the stronger alternative on pragmatic vs. semantic responders. Mean response times for semantic YES and pragmatic NO responses in the different relevance conditions are shown for pragmatic and semantic responders separatelty in the right panel of Figure 3.12 (inconsistent responders are excluded from the visualization). Interestingly, pragmatic responders' semantic responses become slower with increasing relevance of the stronger alternative, while for semantic responders it is the pragmatic responses that become slower. To test whether this interaction is significant, the same model just reported was run on the subset of participants who did not respond entirely inconsistently, with

the additional categorical fixed effect of responder type (pragmatic vs. semantic) interacting with context and response. Unfortunately, due to data sparseness (some cells contained only two observations), multicollinearity between predictors as measured by the variance inflation factor (VIF) was very large, rendering most of the model output uninterpretable.[14]

Summing up, top-down knowledge about the relevance of the stronger alternative affects implicature rates (in line with previous work, e.g. Zondervan, 2010). Response times are less clearly modulated by the QUD; but in the present case this may be due to lack of power. It is also possible that QUD effects are subtle enough that they are not expressed in response times, but may be reflected in a more sensitive measure of linguistic interpretation like eye movements. This is a very interesting avenue for future research.

Accounts that predict the interpretation of utterances of *some* to be a matter of initial default, context-independent processing cannot easily account for the (admittedly preliminary) response results of this study. However, these accounts *are* compatible with the implicature rate differences: both the Default and the Literal-First model allow for context - including, presumably, the relevance of the stronger alternative to a QUD - to enter the computation in a second step. Under the Default model, semantic responses given in the critical condition come about by effortful implicature cancelation. Under the Literal-First model, pragmatic responses come about by effortful implicature calculation. Thus these accounts are perfectly compatible with there being different implicature *rates* across experiments - one need only make the assumption that the experimental contexts are such that there is more implicature cancellation/computation in one over another.

However, neither of these accounts predicts the response time differences obtained. In particular, the Default model predicts that pragmatic NO responses

[14]The VIF was > 5 for all predictors except response inconsistency (1.3) and the interaction between response and responder type (2.3), which both reached significance. The full model is given in Appendix B.

should always be faster than semantic YES responses, while YES responses may themselves exhibit variation in response time depending on how much contextual information needs to be integrate to arrive at the implicature cancellation. While this seems to be the case in Exp. 2a, the results of Exp. 2 (along with the bulk of the response time literature on scalar implicature processing) suggest that pragmatic responses often incur a greater processing cost than semantic responses. In contrast, the Literal-First model predicts that semantic YES responses should always be faster than pragmatic NO responses; the non-default pragmatic responses may exhibit variation in response time depending on the amount of contextual information that needs to be integrated to derive it. Again, this is not what we observed - in Exp. 2a, where the stronger alternative was arguably more relevant than in Exp. 2, pragmatic responses were, if anything, faster than semantic responses.

The Constraint-Based account, on the other hand, explicitly allows for both implicature rates and response times to be modulated by contextual factors like the relevance of alternatives to a QUD. In particular, its underlying assumption that the interpretation process modulates listeners' belief distribution over alternative states of the world is consistent on the one hand with the implicature rate patterns observed: as the relevance of the stronger alternative increases, listeners' expectations that the speaker should use the stronger alternative where appropriate increase, and failure to do so constitutes a larger breach of the requirement to be as informativen as possible, leading to a stronger implicature.

3.7 Conclusion

The experiments presented in this Chapter provide support for the claim that scalar implicature, rather than being a relatively context-independent process, is in fact highly dependent on contextual information. I showed that implicature

rates are modulated by the naturalness of number alternatives to *some*, whether the speaker used the partitive, and the relevance of the stronger alternative to a contextual QUD. In addition, a) the partitive affected response times of responses reflecting the scalar implicature such that the partitive, which was more likely to give rise to an implicature, also led to faster implicatures; b) even when *some* did not require an implicature to be drawn, response times depended on the naturalness of *some*, such that more processing time was required to process *some* when there were rapidly available number terms the speaker could have used; and c) the relevance of the stronger alternative had a small effect on the time taken to process the implicature, such that when the stronger alternative was even somewhat relevant contextually, the upper-bound interpretation was computed more quickly than the lower-bound interpretation, while the reverse was true when the stronger alternative was not relevant.

Together, the evidence suggests that scalar implicature is a highly context-dependent phenomenon, consistent with the Constraint-Based account. However, proponents of two-stage models could argue that response time data is not fine-grained enough to tease apart whether there is not after all an early stage of context-independent interpretation - listeners may have immediately computed a default interpretation, but subsequent context integration processes may have prevented them from immediately responding, resulting in seeming delays. The next Chapter thus investigates the effect of naturalness and availability of number terms on scalar implicature processing using eye movements, a measure more closely time-locked to the incremental interpretation process. This will also allow for more direct comparison with the studies conducted by Huang and Snedeker (2009) and Grodner et al. (2010).

4 The effect of alternatives on eye movements

4.1 Introduction

In Exps. 3 and 4 I more directly test the hypothesis that instructions with exact number delay processing the upper-bound interpretation of *some*. The gumball paradigm was modified so that eye movements could be used to track the time course of interpretation. Figure 4.1 shows a schematic of the displays. An additional color of gumballs was introduced to allow for manipulating whether a partitioned or unpartitioned set of gumballs of either color moved to the lower chamber. Thus, visually distinct regions corresponded to different interpretations. On critical trials, the lower chamber contained a contrast between a partitioned set of, e.g., blue gumballs and an unpartitioned set of, e.g., orange gumballs. Participants then heard statements of the form *You got some of the blue gumballs*, and clicked on the gumballs in the lower chamber that were mentioned in the statement. Crucially, both the partitioned and the unpartitioned set are initially compatible with a semantic interpretation of *some*, while only the partitioned set is compatible with a pragmatic interpretation.

In Exps. 3a and 3b I collected naturalness ratings for the combinations of visual gumball machine displays and sentences used in Exps. 4a and 4b, where eye

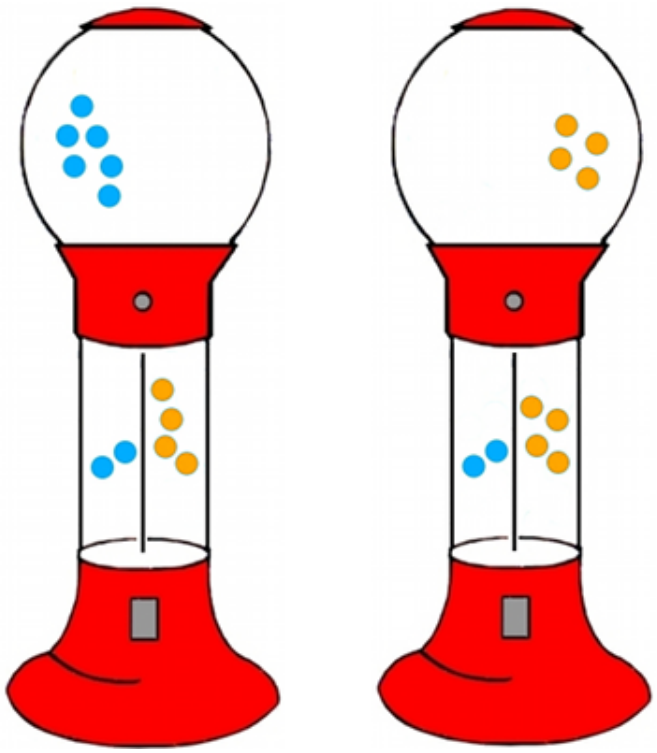

Figure 4.1: Sample displays that contain the same contrast between set sizes in the lower chamber but differ in whether the big or small set is partitioned. Display on the left could occur with sentences *You got some/two of the blue gumballs* or *You got all/four of the orange gumballs*, while the display on the right could occur with *You got all/two of the blue gumballs* or *You got some/four of the orange gumballs*.

movements were recorded. In Exps. 3a and 4a only utterances with the quantifiers *some* and *all* were used, whereas in Exps. 3b and 4b the number terms *two*, *three*, *four*, and *five* were included among the stimuli.

4.2 Exp. 3a: naturalness norms for Exp. 4a

4.2.1 Methods

Participants

Using Amazon's Mechanical Turk, 80 workers were paid $0.60 to participate. All were native speakers of English (as per requirement) who were naïve as to the purpose of the experiment.

Procedure and materials

On each trial, participants saw a display of a gumball machine with an upper chamber filled with 8 gumballs of one color of gumballs and either 2, 3, 4, or 5 gumballs of the other color. The lower chamber was empty. After 1.5 seconds a new display was presented in which a certain number of gumballs had dropped to the lower chamber. Participants heard a pre-recorded statement of the form *You got X of the C gumballs*, where X was a quantifier and C was a color adjective (either *blue* or *orange*). They were then asked to rate how naturally the scene was described by the statement on a seven point Likert scale, where seven was very natural and one was very unnatural. Participants were instructed to click a FALSE button located beneath the scale if they thought the statement was false.

Some trials contained literally false statements. For example, participants might get two of the eight blue gumballs (as in the left machine in Figure 4.1) and hear *You got all of the blue gumballs.* These trials were interspersed in order to have a baseline against which to compare naturalness judgments for *some* used with the unpartitioned set. If interpreted semantically (as *You got some and possibly all of the blue gumballs* in a scenario where the lower chamber contains 4 of 4 (all) blue gumballs and 2 of 8 orange gumballs), the *some* statement is true (however unnatural) for the unpartitioned set. However, if interpreted pragmatically

as meaning *You got some but not all of the blue gumballs*, it is false and should receive a FALSE rating. I refer to these trials as *garden-path* trials because the information from the quantifier is misleading as to the target and thus leads the listener down an interpretive garden path that is incompatible with the information from the adjective that follows. Two thirds of the stimuli were semantically true, pragmatically felicitous items (regular trials), while one third consisted of semantically false or pragmatically infelicitous *garden-path* trials.

Of the regular trials, eight trials did not contain a contrast in the lower chamber. That is, there were two orange gumballs and two blue gumballs out of two each in the *all* condition and out of eight each in the *some* condition. Similarly for three and three, four and four, and five and five gumballs. These no-contrast displays constituted the *late* baseline condition in Exps. 4, where the information from the quantifier did not provide disambiguating information regardless of whether it was interpreted semantically or pragmatically, and only the later-occurring color adjective (blue or orange) disambiguated the target set of gumballs.

The remaining eight regular trials all contained a contrast between a smaller set of gumballs of one color (2 or 3) and a larger set of the other (4 or 5) in the lower chamber. These were the *early* conditions in Exps. 4, where making use of the information from the quantifier is sufficient for picking out the target set (under the assumption that the implicature is drawn for *some* and regardless for *all*).

Thus, there were three (*early*, *late*, *garden-path*) x four (target set size: *2, 3, 4, 5*) x two (*all*, *some*) trials. Each participant rated the naturalness of statements with quantifiers as descriptions of gumball machine scenes on 24 trials. Trial order was randomized.

4.2.2 Results and discussion

Clicks of the FALSE button were coded as 0. Data from 24 participants were excluded because they did not use the FALSE button, indicating reduced attention to the task. I first present results from the *garden-path* condition and then focus on naturalness differences for *some* and *all* in the *early* condition.

On *garden-path* trials, the mean rating for *all* was close to 0 (0.4, SD=1.34), reflecting that most participants correctly judged semantically false sentences as false. The mean rating for *some* was higher at 2.55 (SD=2.55), indicating that participants were sensitive to the informativeness violation but did not judge underinformative uses of *some* as categorically false. On *garden-path* trials, 35% of responses to *some* were FALSE, compared to 85% FALSE responses to *all*.

Mean naturalness ratings in the *early* condition are shown in Figure 4.2. A mixed effects linear regression predicted naturalness ratings in the *early* condition from quantifier (*all* vs. *some*), target size (*big* vs. *small*) and their interactions. Predictors were centered before entering the analysis. The model contained the maximal by-participant random effects structure (random by-participant intercepts and random by-participant slopes for both main effects and the interaction term). Effect significance was assessed via model comparison of a model containing the predictor of interest with a minimally different model that did not contain that predictor, in keeping with Barr et al. (2013).

There was a main effect of target size such that descriptions of smaller sets were judged to beless natural than descriptions of larger sets (β = -0.27, SE = 0.12, t = -2.27, $p < .05$). This may be due to a prior preference for smaller sets to be referred to with number terms, which may in turn be due to rapid subitizing effects leading to automatic naming of set size for small, but not bigger sets. There was also an interaction of quantifier and set size ($\beta = 0.25$, $SE = 0.12$, $t = 2.04$, $p < .05$). Inspecting the simple effects (and as can be seen in Figure 4.2) revealed

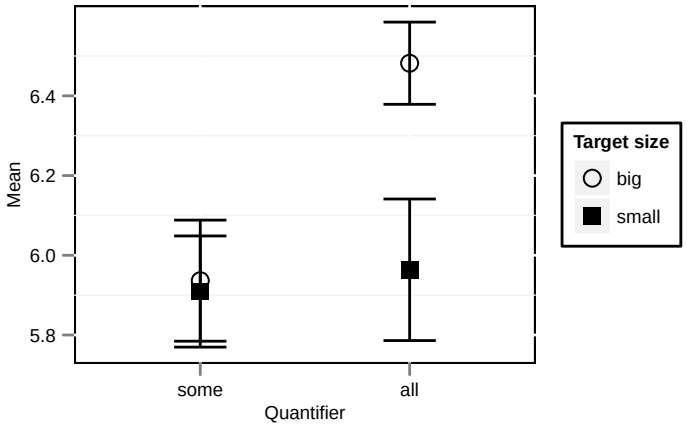

Figure 4.2: Mean naturalness ratings in the *early* condition for *some* and *all*, collapsing set sizes 2 and 3 into *small* and sizes 4 and 5 into *big*.

that naturalness was lower for *all* when the set was small than when it was big (β = -0.52, SE = 0.16, t = -3.19, $p < .01$), but there was no difference in naturalness for *some* with different set sizes (β = -0.02, SE = 0.18, t = -0.12, $p < .7$). The main effect of quantifier did not reach significance (β = -0.14, SE = 0.1, t = -1.46, $p < 0.15$).

In Exp. 3b, number terms were added to the stimuli to assess how availability of alternative descriptions with number would affect naturalness ratings for *some* and *all*.

4.3 Exp. 3b: naturalness norms for Exp. 4b

4.3.1 Methods

Participants

Using Amazon's Mechanical Turk, 80 workers were paid $0.80 to participate. All were native speakers of English (as per requirement) who did not participate in Exp. 3a and were naïve as to the purpose of the experiment.

4.3.2 Procedure and materials

The procedure was the same as that described for Exp. 3a with one difference: the number terms *two*, *three*, *four*, and *five* were included among the stimuli. Each participant rated naturalness of statements with quantifiers as descriptions of gumball machine scenes on 48 trials. Trial order was randomized. Half of the trials were identical to the 24 trials from Exp. 3a. The other half consisted in number trials. Each number term occurred six times: twice in the *early* condition, twice in the *late* condition, and twice in the *garden-path* condition. Of the two times it occurred, it occurred once with an *unpartitioned* and once with a *partitioned* set. For example, the statement *You got two of the C gumballs* (where *C* could be one of *blue* or *orange*) occurred in displays where there was either a 2-2 (*late* condition) or 2-4 (*early* and *garden-path* conditions) distribution of blue and orange gumballs in the lower chamber. If it occurred in the *late* condition, one set was partitioned (e.g., 2 out of 8 orange gumballs) while the other one was not (e.g., 2 out of 2 blue gumballs). In the *early* condition, the *two* statement was either about a partitioned set (in which case there were 2 out of 8 gumballs on one side and 4 out of 4 on the other) or an unpartitioned set (2 out of 2 vs. 4 out of 8). The *garden-path* trials were the same as the *early* trials but the color term in the statement was switched.

4.3.3 Results and discussion

As in Exp. 3a, clicks of the FALSE button were coded as 0. Data from 18 participants were excluded because they did not use the FALSE button. I first present results from the *garden-path* condition and then focus on naturalness differences for *some* and *all* in the *early* condition in Exp. 3b (numbers *present*) vs. Exp. 3a (numbers *absent*).

On *garden-path* trials, mean ratings for *some* and *all* were very similar to

ratings from Exp. 3a (2.57 and 0.59, respectively). Mean ratings for *two*, *three*, *four*, and *five* were similarly low (0.98, 0.88, 1.01, and 0.67, respectively). For *some*, 33% of responses were FALSE, compared to 75% FALSE responses for the other quantifiers. The slightly lower number of FALSE responses compared to Exp. 3a is due to somewhat lower FALSE rates for *two* and *three*, where an *at least* interpretation of the description is semantically true.

Mean ratings in the *early* condition are shown in Figure 4.3 alongside the ratings from Exp. 3a. The effect of number presence on the naturalness of *some* and *all* was assessed using a mixed effects linear regression predicting naturalness rating on the subset of the data that included only *early some* and *all* trials. Fixed effect predictors were quantifier (*all* vs. *some*), target size (*big* vs. *small*), number presence (*absent* vs. *present*) and their interactions. Predictors were centered before entering the analysis. The model contained the maximal by-participant random effects structure (random by-participant intercepts and random by-participant slopes for all within- participant predictors). As in Exp. 3a, the significance of an effect was assessed via model comparison of a model containing the predictor of interest with a minimally different model that did not contain that predictor.

There was a main effect of number term presence in the predicted direction: when number terms were included among the stimuli, the naturalness of both *some* and *all* was judged to be lower (β = -0.42, SE = 0.20, t = -2.15, $p < .05$). As in Exp. 3a, descriptions of smaller sets were judged to be less natural than descriptions of larger sets (β = -0.27, SE = 0.08, t = -3.58, $p < .001$). In contrast to Exp. 3a, the main effect of quantifier was significant such that *some* was judged as overall less natural than *all* (β = -0.43, SE = 0.14, t = -3.23, $p < .01$). Finally, there was a marginally significant three-way interaction between quantifier, target size, and number presence (β = -0.63, SE = 0.25, t = -1.81, $p < .08$). Inspecting the simple effects revealed that the two-way interaction between quantifier and

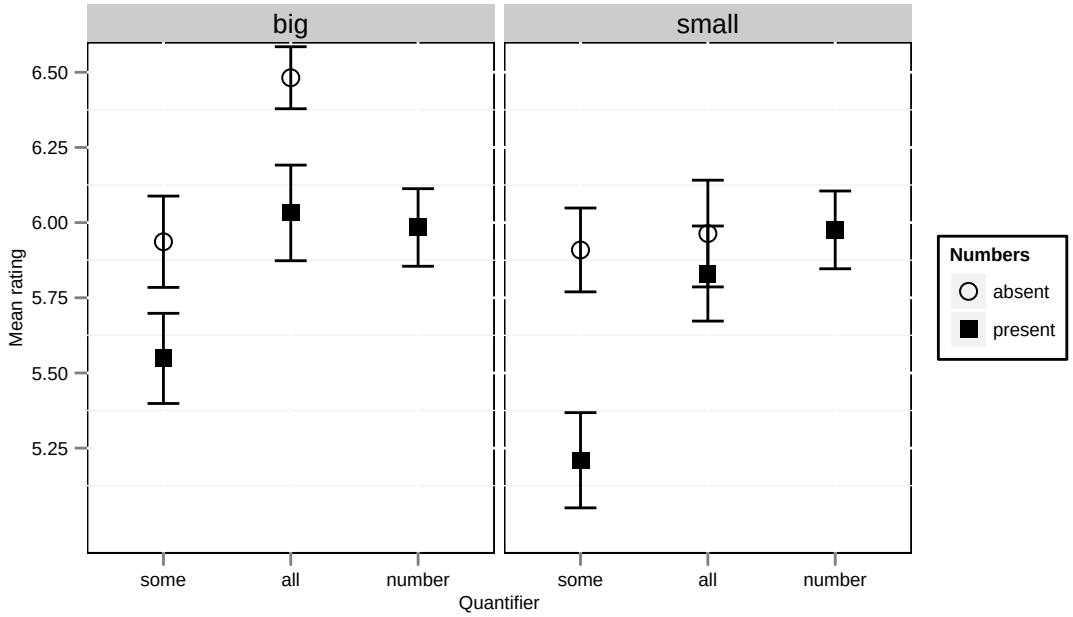

Figure 4.3: Mean naturalness ratings for *some*, *all*, and number terms in the *early* condition for different set sizes. Circles repeat values from Exp. 3a (numbers *absent*), while squares are values from Exp. 3b (numbers *present*).

target size that was present in Exp. 3a is not significant when numbers are present (β = -0.14, SE = 0.25, t = -0.55, $p < .58$); when numbers are present descriptions are rated as more natural for the big set than the small set regardless of the quantifier. Put differently, the number presence effect is equally large for *some* and *all* when the target set is big, but larger for *some* than for *all* when the target set is small. Allowing number as a third level of the quantifier variable in the analysis revealed that the naturalness of *all* and number terms was rated as equally high regardless of set size (β = 0.08, SE = 0.07, t = 1.08, $p < .25$).

To summarize the effect of number term presence on naturalness: when numbers were absent, *some* and *all* were equally natural for small sets. Big set *some* was as natural as small set *all* and small set *some*. However, big set *all* was the most natural. Adding numbers reduced the naturalness of *all* and *some*. Big set *all* became less natural with numbers. Nonetheless, its naturalness ratings were

roughly equivalent to the naturalness of numbers for both set sizes. The biggest effects of adding number occurred for small set *some*, which became much less natural than any of the other conditions.

4.4 Exp. 4a: eye movements in the absence of number terms

The goal of Exps. 4a and 4b was to investigate whether the naturalness of different quantifiers used with different set sizes affects the earliest stages of implicature processing, as predicted by the Constraint-Based account. Exps. 4a and 4b were conducted in the lab to allow eye movements to be recorded. This required using a repeated measures design. Rather than using each unique combination of target set size, quantifier, point of disambiguation/garden-path condition once, data from multiple trials in each condition was recorded. The choice of set sizes was motivated by the following considerations. First, I aimed to specifically investigate the effect of *some* used with very small (2, 3) sets as opposed to somewhat larger sets where the number term is not as rapidly available. In particular, 2 was the size of the *some*-sets used by Huang and Snedeker (2009); Grodner et al. (2010) used sets of size 2 and 3.

For the larger sets, all things being equal, it would have been preferable to choose sets where subitizing is clearly not possible. However, there were some additional constraints. First, the larger the contrast set, the larger the baseline fixation differences between set sizes. In order to avoid a baseline fixation probability for the big set that would be too high to reliably observe early effects of the quantifier, the size difference compared to the small sets had to be minimized. Second, in order to avoid processing delays associated with cohort effects, number terms that overlapped at onset with *some* had to be avoided. That is, in order to

avoid a temporary ambiguity between a number term and *some*, it was important not to use any numbers terms that began with the the phoneme /*s*/. This ruled out sets of size 6 and 7, for which the number terms *six* and *seven* are cohort competitors for *some*. This left the set sizes 4 and 5. While 4 is still inside the subitizing range, we saw in Section 3.4 that interference from the number term on the interpretation of *some* is minimal with set size 4 and absent with size 5.

I now consider the main predictions for the Constraint-based model, taking into account naturalness and availability of alternatives. I compare these predictions with the Literal-First and Default models, neither of which takes into account naturalness and availability. In the absence of number terms (Exp. 4a), the Constraint-Based model and the Default model both predict that listeners should begin to compute the implicature as soon as they hear *some*. In contrast, the Literal-First model predicts that pragmatic *some* will be delayed relative to *all*, which does not require an implicature. With number terms present (Exp. 4b), however, the Constraint-Based model predicts delayed effects for *some*, with the delays being most pronounced for small set *some*. The Literal-First model continues to predict delayed effects for pragmatic *some*, whereas the Default model predicts rapid computation of the implicature. In sum, based on the naturalness ratings, the Constraint-Based model makes the same predictions as the Default model when numbers are not included, whereas when numbers are included, it makes the same predictions as the Literal-First model.

4.4.1 Methods

Participants

Forty undergraduate students from the University of Rochester were paid $10 to participate. All were native speakers of English who had not participated in any of the previously reported experiments.

Procedure

The procedure and stimuli were similar to Exps. 3a and 3b. On each trial, participants saw the same types of initial displays of gumball machines as in Exp. 3a. After 2 seconds, the button in the center of the machine started flashing. Participants then clicked on the button. Upon clicking, a gray mask was displayed for 200ms, then the gumball machine was redisplayed with a certain number of gumballs of each color having dropped to the lower chamber. Clicking the central button ensured that all participants were looking at a central fixation point at the time of auditory stimulus onset. After 500ms, participants heard an auditory stimulus of the form *You got X of the C gumballs*, where X was one of *some* or *all* and C was one of *blue* or *orange*. Their task was to click on the side of the lower chamber that contained the gumballs mentioned in the statement if they thought the statement was true, and click on the central button otherwise. Once they clicked either on a side of the lower chamber or on the central button, a gray screen was displayed for one second and the experiment advanced to the next trial.

Participants first received a short tutorial explaining the task. They then completed four practice trials (that did not contain a *some* statement) before beginning the experimental trials. The total experiment consisted of 64 trials.

Participants' eye movements were recorded with an SR Eyelink II head mounted eye-tracker with a sampling rate of 250 Hz. Drift correction was performed every five trials. Auditory stimuli were presented over Sennheiser HD 570 headphones at a comfortable listening level.

Materials

The same visual stimuli as described in Exp. 3a were used. Visual stimuli were constructed such that on early trials there was always a contrast between a larger

and a smaller set in the lower chamber. If the smaller set was of size 2, the larger one was 4. If the smaller set was of size 3, the larger set was 5 gumballs. One of the two sets was always partitioned. The set that was partitioned originally started out as a set of 8 in the upper chamber. The left panel of Figure 4.1 shows a display that could have been used on an *early small some* trial or on an *early big all* trial, while the right panel shows a display that could have been used on *early big some* or *early small all* trials. The *early* trials were the trials of most interest because they were the ones on which the time course of the implicature from *some* to *not all* would emerge. The faster the implicature is computed, the earlier after the onset of *some* participants' looks should converge on the target set (the partitioned set). On early *all* trials, the quantifier is immediately disambiguating and participants should be able to quickly converge on the target set.

On *late* trials, there were either two partitioned (for *some*) or two unpartitioned (for *all*) sets of the same size in the lower chamber. Each of the two quantifiers occurred with each set size (2, 3, 4, 5) twice. On *late* trials, the quantifier thus did not disambiguate the target set of gumballs and participants had to wait until they heard the color adjective (*orange* or *blue*) further downstream to identify the target. This condition provided a baseline against which to determine whether the quantifier affected participants' interpretation on early trials.

Finally, the *garden-path* trials employed the same displays as *early* trials. However, on these trials the information provided by the quantifier was misleading. For *all*, the statement was semantically false. That is, in a situation like the left panel of Figure 4.1, participants might have heard *You got all of the blue gumballs.* In the instructions they were told to click the central button on the machine if the statement they heard didn't refer to a set of gumballs in the display. In the *garden-path some* condition, the statement was false if interpreted pragmatically as, e.g., *You got some, but not all of the orange gumballs* heard with the left display in Figure 4.1. However, it was true if interpreted semantically as *You got*

at least one of the orange gumballs. This allowed for participants to be classified as more pragmatic or more semantic responders depending on their responses on *garden-path some* trials. These trials also served as a further test for whether the inference from *some* to *not all* is delayed. An increase in looks to the partitioned set of gumballs before the switch to the gumballs compatible with the downstream adjective would constitute evidence for early implicature computation.

Target set color (orange, blue) and location (left, right) were each counterbalanced within quantifier condition to discourage participants from forming target associations with either color or location cues. In addition, for the quantifiers *some* and *all*, target set size was counterbalanced, i.e., *some* and *all* were used equally often to refer to a set of size 2, 3, 4, or 5. For each unique display type (i.e., for each unique combination of color and set size of left and right set of gumballs), two versions were created to further discourage participants from adopting display type based looking strategies. Importantly, each quantifier occurred with each unique display type. It was therefore not possible for participants to predict whether they would hear an *all* or a *some*-statement based solely on the display.

As already mentioned, each auditory stimulus was a sentence of the form *You got X of the C gumballs*. *X* was one of the quantifiers *some* and *all*. *C* was one of the color adjectives *blue* and *orange*. Stimuli were cross-spliced in the following way. Each sentence was recorded individually. The quantifier and color adjective were subsequently spliced out of the original recording and onto a recording of the sentence *You got most of the green gumballs*, where *most* and *green* were replaced by the experimental items. This ensured that quantifier onset was equal across all stimuli (at 462ms) and the onset of the color adjective was the same within each quantifier (*all*: 847ms, *some*: 856ms). Mean sentence length was 1844ms.

The cells in blue font in Table 4.1 show the number of trials each participant saw in each condition in Exp. 4a. Trial order was randomized.

Table 4.1: Distribution of trials over point of disambiguation (POD) / garden-path (GP), target set size, and quantifier conditions. Participants in the numbers *present* (Exp. 4b) condition saw all 96 trials, in the numbers *absent* (Exp. 2a, blue font) condition only the non-number trials.

		Quantifier						
POD/GP	Target	*some*	*all*	*two*	*three*	*four*	*five*	Total
Early	Small (2/3)	8	8	4	4			24
	Big (4/5)	8	8			4	4	24
Late	Small (2/3)	4	4	2	2			12
	Big (4/5)	4	4			2	2	12
Garden-path	Small (2/3)	4	4	2	2			12
	Big (4/5)	4	4			2	2	12
Total		32	32	8	8	8	8	96

4.4.2 Results and discussion

2% of the data were excluded because response time as measured from auditory stimulus onset was greater than three standard deviations from mean response time in each of the *early*, *late*, and *garden-path* conditions, respectively.[1] I first report the distribution of semantic vs. pragmatic responders in the garden-path *some* condition and then analyze the eye-movement data.

Semantic vs. pragmatic response distribution

In the underinformative *garden-path some* trials there were 78.2% pragmatic responses and 21.8% semantic responses. The distribution of participants over number of semantic responses given is shown in Table 4.2. I will refer to participants with fewer than four semantic responses as *pragmatic* responders, participants with more than four semantic responses as *semantic* responders, and participants

[1]Mean response time was calculated relative to these three conditions because especially the *garden-path* trials are expected to take longer, given that they are initially misleading. This was indeed the case. Mean response times in the *early*, *late*, and *garden-path* conditions were 1894ms, 1856ms, and 2489ms, respectively.

Table 4.2: Distribution of participants over number of semantic responses given (out of 8 possible) in *garden-path some* condition.

Number of semantic responses	0	1	2	3	4	5	6	7	8
Number of participants	20	6	1	3	6	0	1	1	2

with exactly four semantic responses as *inconsistent* responders.

More participants gave mostly pragmatic responses in this experiment than in Exp. 2, which also used the gumball task with numbers. One possible explanation is that the task in Exp. 4a is implicitly a referential task in which the implicature is more relevant than when there is only one color of gumballs and participants perform a truth-value judgment task. In order to determine the target as quickly as possible, participants should make optimal use of the information provided by the quantifier. By interpreting *some* as *some but not all*, an observation of *some* is more informative than if it is interpreted as *some and possibly all*, since the former is compatible with only one set of gumballs, while the latter is compatible with both.

Eye movements

Eye movement analyses were conducted on the time window beginning 200ms after quantifier onset and ending 200ms after adjective onset. This is the window in which looking patterns can plausibly be driven by the observed quantifier. I refer to this window as the quantifier window. Unless otherwise mentioned, all results were obtained from fitting mixed effects logistic regression models predicting looks to the target over looks to the target and competitor to different subsets of the data in the quantifier window. All models contained the maximal by-participant random effects structure.

I present analyses in the following order. First I establish that there was no difference between the *early* and the *garden-path* condition in the quantifier

window. I then compare the *early* condition to the *late* condition to establish that participants indeed could not make use of the quantifier in the *late* condition, but could do so in the *early* condition. Next I report the analysis of main theoretical interest, testing whether the naturalness results from Exp. 3a are reflected in eye movements. Finally, I present analyses showing different looking behavior in the quantifier window for semantic and pragmatic responders.

***Early* vs. *garden-path* condition** From the perspective of the participants, the *early* and *garden-path* conditions were identical until adjective onset: the same display types were used in the two conditions, with the only difference being that on *garden-path* trials, the information from the adjective did not match the information from the quantifier. Thus, if the information from the quantifier guides participants' early looks, there should be no difference in looking patterns up until the information from the adjective becomes available. To test this, a model was fit to the quantifier window data predicting target looks from condition (*early* vs. *garden-path*), time and their interaction. Only the main effect of time reached significance (β = 0.07, SE = 0.03, $p < .01$); neither the main effect of condition (β = -0.01, SE = 0.10, $p < .92$) nor the interaction (β = 0.03, SE = 0.05, $p < .54$) reached significance. This suggests that, while participants did use the information from the quantifier to identify the target (as evidenced in the main effect of time), there was indeed no difference in looking behavior between the *early* and *garden-path* condition. This remained true when controlling for target set size and quantifier. Given that the design of these two conditions was identical in the quantifier window and there were no measurable eye movement differences between these two conditions, I collapse over *early* and *garden-path* trials when performing analyses on the quantifier window.

***Early*/*garden-path* vs. *late* point of disambiguation** To determine whether participants were using information from the quantifier in the *early* and *garden-*

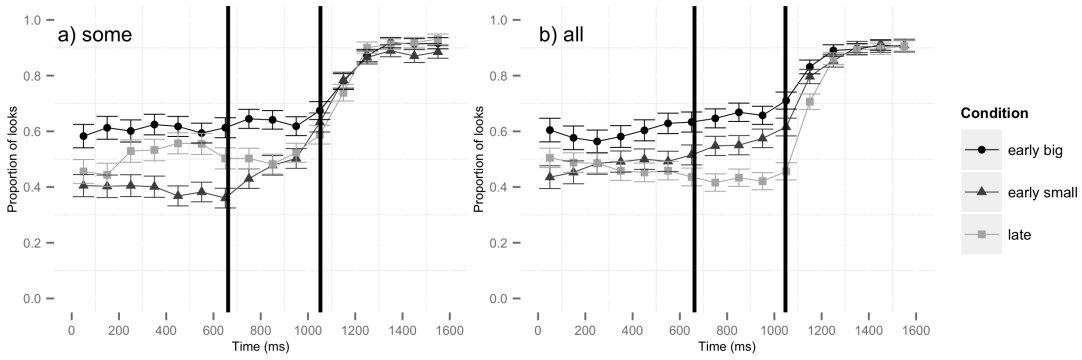

Figure 4.4: Proportion of looks to the target by condition for *some* (left) and *all* (right). Black vertical lines demarcate the quantifier window. The x-axis shows time in ms relative to auditory stimulus onset.

path but not in the *late* condition, one model each was fit to the *early*/*garden-path* and *late* subsets of the data. Target looks were predicted from a predictor coding time window, where time window was either the baseline window (the 400ms window ranging from quantifier onset minus 200ms to the start of the quantifier window) or the quantifier window. If participants were making use of the quantifier, there should be a significant effect of time window. The effect of time window was significant for the *early* condition ($\beta = 0.29$, $SE = 0.05$, $p < .0001$) but not for the *late* condition (β = -0.10, $SE = 0.08$, $p < .22$), suggesting that the quantifier information affected eye movements only when it could be used to disambiguate the target. Proportions of looks to the target for *early* and *late* conditions are shown in Figure 4.4.[2]

Naturalness effects The coarse-grained time window analysis on the *early* vs. *late* conditions shows that there are early effects of the quantifier. However,

[2]Note that in this graph as well as in all following graphs, data from the *garden-path* condition is not plotted because looking behavior after adjective onset diverges substantially from that on *early* trials (i.e., participants look to the set that is compatible with the adjective before typically looking to the central button on the machine that is clicked to indicate disagreement with the description). For visually interpreting the graphs, this means that what is visible to the eye in the quantifier window is only an approximation of the data that entered into the quantifier window analysis, where *early* and *garden-path* trials were collapsed.

these analyses do not reveal whether there are differences between quantifiers and set sizes in how quickly the target is identified. Therefore, a model was fit to the *early* and *garden-path* subset of the data (which are indistinguishable up until the adjective is observed) predicting target looks from fixed effects of quantifier type (*all* vs. *some*), target size (*big* vs. *small*), time (continuous), and their interactions. There were only significant main effects of target size and time, such that participants were more likely to look at big sets rather than small sets in this window (β = -0.59, SE = 0.11, $p < .0001$) and as time increased the log odds of looking to the target increased (β = 0.07, SE = 0.04, $p < .05$). The interaction between quantifier and set size observed for the naturalness ratings in Exp. 3a did not reach significance for eye movements (β = -0.06, SE = 0.09, $p < .51$). This is most likely due to participants' general preference to look at the larger of two sets, resulting in the main effect of target set size. Proportions of looks to the target in the early condition by target set size and quantifier are shown in Figure 4.5.

Importantly, the main effect of time combined with the lack of effect for quantifier suggests that participants look more towards the target in this window over time regardless of quantifier. That is, the information from the quantifier *some* is used immediately to disambiguate the target, suggesting in turn that computing the implicature from *some* to *not all* in this experiment does not come at a higher cost or greater delay than integrating the information from the quantifier *all*.

This result is inconsistent with the Literal-First hypothesis which predicts a delay in the processing of *some* relative to *all* but consistent with both the Default and the Constraint-Based accounts. Proponents of the Literal-First hypothesis might argue that rather than quickly computing the implicature, participants adopted a display-based strategy to disambiguate the target. However, this is unlikely because the same display types were used for both quantifiers. Participants could not use either the initial display in the upper chamber or the shifted

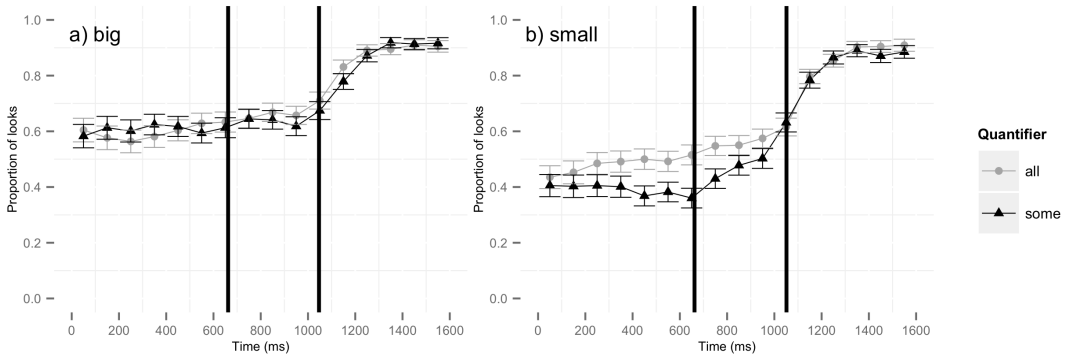

Figure 4.5: Proportion of looks to the target by quantifier for *big* set (left) and *small* set (right) targets in the *early* condition. Black vertical lines demarcate the quantifier window. The x-axis shows time in ms relative to auditory stimulus onset.

display in the lower chamber to predict whether they would hear an *all* or a *some* statement. However, it is possible that participants adopted a more sophisticated strategy. Perhaps after enough exposure, they learned to associate the partitioned set with *some* (in accordance with its pragmatics). This possibility was explored by comparing the results for the first and second half of the experiment to see whether looking patterns changed. A centered predictor coding *first* vs. *second* half of the experiment was allowed to interact with quantifier and target set size. The main effects of target set size and time remained, and there was additionally a marginal main effect of half, such that participants were more likely to look at the target in the second half than in the first half (β = -0.28, SE = 0.16, $p <$ .07). However, there was no interaction of half with quantifier, suggesting that participants did not adopt a *some*-specific strategy.

Looking behavior for semantic and pragmatic responders In Exp. 2 I found that response times of participants who most often responded pragmatically on *garden-path some* trials exhibited greater naturalness effects than participants who most often responded semantically. To test whether this was also the case for eye movements, a centered predictor coding responder type (*pragmatic* vs. *seman-*

tic) was allowed to interact with the target size, quantifier, and time predictors on the subset of the data that did not include entirely inconsistent responders. The main effects of target size and time remained significant. In addition, there was a main effect of responder type such that semantic responders were less likely to look to the target than pragmatic responders (β = -0.39, SE = 0.19, $p < .05$). In a model where responder type was additionally allowed to interact with experiment half, there was a marginally significant interaction between responder type and half, such that pragmatic responders looked more to the target in the second half than in the first, but there was no difference in looking behavior between experimental halves for semantic responders (β = -0.77, SE = 0.43, $p < .08$). Taken together, this suggests that semantic responders were less apt to use the information from the quantifier to identify the target set of gumballs than pragmatic responders. In addition, pragmatic responders became better at using the quantifier information as the experiment progressed, whereas semantic responders continued to rely on the adjective as the main cue to the target. Thus, rather than displaying greater naturalness effects (as observed in Exp. 2), participants' looking patterns revealed that pragmatic responders payed greater attention to the quantifier than semantic responders.

4.5 Exp. 4b: eye movements in the presence of number terms

Exp. 4b investigated the effect of available alternatives to *some* on the speed of implicature processing by adding number terms to the stimuli in addition to the quantifiers *some* and *all*.

4.5.1 Methods

Participants

Forty undergraduate students from the University of Rochester were paid $10 to participate. All were native speakers of English did not participate in any of the previously reported experiments.

Procedure and materials

The procedure was identical to Exp. 4a and the same materials were used, but Exp. 4b contained 32 additional number trials (shown in black font in Table 4.1). On these trials, the quantifier in the statement was either *two*, *three*, *four*, or *five*. Of these number trials, 16 were *early* trials, 8 were *late*, and 8 were semantically false *garden-path* trials. The same display types were used for each condition as in Exp. 4a. For example, either of the displays in Figure 4.1 could be used on *early two*/*four* trials. For number term trials, target set partitioning was counterbalanced (half of the number trials had the number term referring to a partitioned, half to an unpartitioned set). This was intended to discourage participants from forming associations between number terms and whether or not a set of gumballs was partitioned or unpartitioned.

Auditory stimuli were cross-spliced as in Exp. 4a. Quantifier onset was again 462ms after sentence onset. Adjective onsets were the same within quantifier but differed between quantifiers: *two*: 815ms, *three*: 894ms, *four*: 902ms, *five*: 909ms. Trial order was pseudo-random, such that the number of *early*, *late*, and *garden-path* trials was counter-balanced within each quarter of the experiment, as was the number of *some*, *all*, and number trials as well as target set size.

Table 4.3: Distribution of participants over number of semantic responses given (out of 8 possible) in the *garden-path some* condition in Exps. 4a and 4b.

Number of semantic responses	0	1	2	3	4	5	6	7	8
Number of participants (Exp. 4a)	20	6	1	3	6	0	1	1	2
Number of participants (Exp. 4b)	14	6	4	1	1	1	0	3	7

4.5.2 Results and discussion

Data for three participants were excluded because they did not respond correctly on semantic *garden-path* trials; instead of clicking the central button, they either clicked on the set of gumballs compatible with the quantifier or on the set of gumballs compatible with the adjective. This constituted 6.5% of the data. A further 1.5% of the data were excluded because response time as measured from auditory stimulus onset was greater than three standard deviations from the mean response time in each of the *early*, *late*, and *garden-path* conditions, respectively.

I structure the results in the following way, in order to investigate both the effect of adding number terms to the experimental items as well as between-quantifier and between-set size differences: first I report the distribution of semantic vs. pragmatic responders in the *garden-path some* condition and compare it to Exp. 4a. I then analyze differences in response times between Exps. 4a and 4b. Finally, I report the eye movement results.

Semantic vs. pragmatic response distribution

In the underinformative *garden-path some* trials there were 64.9% pragmatic responses and 35.1% semantic responses. Table 4.3 shows how many participants made each number of semantic responses.

I considered participants with fewer than four semantic responses to be pragmatic responders, and participants with more than four semantic responses to be semantic responders. A χ^2 test comparing Exps. 4a and 4b revealed that the dis-

tributions of responder types differed ($\chi^2(2) = 7.19$, $p < .05$). There are twice as many pragmatic responders than semantic responders in Exp. 4b compared with seven times as many pragmatic responders in Exp. 4a.

One possible explanation for this difference is that *some* becomes a less preferred alternative overall when number terms are available (as evidenced in the naturalness ratings obtained in Exp. 3b). However, the naturalness rating for *some* in the *garden-path* condition was the same regardless of whether or not numbers were present. That is, the naturalness difference between using *some* felicitously (with partitioned sets) vs. infelicitously (with unpartitioned sets) is not as large when numbers are available alternatives. This may have made the naturalness difference for *some* less salient overall and increased the likelihood that participants would accept the underinformative statement in Exp. 4b.

Response times

Mean response times in the early, late and garden-path conditions were 2092ms, 2051ms, and 2680ms, respectively. Response times were significantly slower than in Exp. 4a. This emerges in a mixed effects linear regression on the *some*/*all* subset of the data from both Experiments, predicting log-transformed response time from fixed effects of quantifier (*all* vs. *some*), number presence (*absent* vs. *present*), target set size (*big* vs. *small*), and their interactions. All predictors were centered before entering the analysis. The model included by-participant random slopes for all within-participant fixed effects.

There were significant main effects of quantifier and number presence, such that responses were slower to *some* than to *all* ($\beta = 0.04$, $SE = 0.01$, $t = 5.0$, $p < .0001$) and responses were slower when numbers were present than when they were absent ($\beta = 0.08$, $SE = 0.03$, $t = 2.8$, $p < .01$). No other effects reached significance. Both the quantifier and the number presence effect are predicted by

the naturalness data from Exps. 3a and 3b, where *some* was overall less natural than *all* and overall naturalness was lower when numbers were present.

There are at least two alternative explanations for why responses to *some* are slower than responses to *all*. One explanation is that the interpretation of *some* as *some but not all* is associated with a more complex verification strategy than the interpretation of *all* because both the size of the reference set (in the lower chamber) and the complement set (in the upper chamber) need to be verified.[3] Second, it is possible that the pragmatic interpretation is delayed due to a staged process of interpretation, consistent with the Literal-First hypothesis.

A further result to consider in this regard is the relationship between response times for semantic and pragmatic responses in the *garden-path some* condition. Note, however, that both the Literal-First hypothesis and the verification complexity explanation make the same prediction: pragmatic responses should be slower than semantic responses. This prediction was tested in two ways, corresponding to the analysis of semantic and pragmatic responses reported in Bott and Noveck (2004) (and as already performed for Exp. 2). First, I compared responses between participants that replied entirely consistently on *garden-path* trials (i.e., who gave either eight semantic or eight pragmatic responses). I then compared semantic and pragmatic responses for participants with inconsistent responses. In each mixed-effects linear regression, log-transformed response time was predicted from response type (*pragmatic* vs. *semantic*) and random by-participant intercepts. In both analyses, pragmatic responses were slower than semantic responses (between-participants: (β = 0.12, SE = 0.06, t = -2.2, $p < .05$), within-participants: (β = -0.14, SE = 0.03, t = -4.12, $p < .001$), replicating previous findings by Bott and Noveck (2004) (but recall that I only partly replicated this result in Exp. 2). If response type is allowed to interact with number presence

[3]See e.g., Bott et al. (2012), Grodner et al. (2010), and the Discussion of Chapter 3 for elaborations of this idea.

in the between-participants comparison, there is a significant interaction between response type and number presence (β = -0.21, SE = 0.12, t = -1.8, $p < .05$), such that pragmatic but not semantic participants were slower to respond when numbers were present. This suggests that pragmatic responders were more susceptible to naturalness effects (as evidenced in reduced naturalness of *some* when numbers were introduced into the items) than semantic responders.

The garden-path response time data cannot distinguish between the Literal-First and the verification-based explanation for the difference in response times between *some* and *all*. However, the increase in participants' looks to the target in the Exp. 4a quantifier window was similar after *some* and *all*. This suggests that participants began generating the implicature at the earliest moments of online processing. Therefore, the greater response time for *some* but not *all* statements is likely due to increased verification time and not increased time to compute the implicature.

Eye movements

As in Exp. 4a, eye movement analyses were conducted on the quantifier window starting 200ms after quantifier onset and lasting until 200ms after adjective onset. All reported results were obtained by fitting mixed effects logistic regression models predicting looks to the target over looks to the target and competitor to different subsets of the data in this time window. All models contained the maximal by-participant random effects structure.

I present analyses in the following order. I first compare the *early* condition to the *garden-path* condition to establish that there was no difference in looking behavior in the quantifier window. I then compare the *early* to the *late* condition to test whether participants were making use of the quantifier when they could. Next I conduct the analysis of main theoretical interest, testing whether the naturalness results from Exp. 3b are reflected in eye movements, with a focus

on the effect of adding number terms to the stimuli. I then analyze the effect of adding number terms separately for semantic vs. pragmatic responders. Finally, I present analyses of baseline looking preferences towards big/small and partitioned/unpartitioned sets even before quantifier onset to explore potential strategizing effects in Exp. 4b.

***Early* vs. *garden-path* condition** As in Exp. 4a, it was important to establish that there was no difference in looking patterns between the *early* and *garden-path* condition before the information from the adjective became available. I thus fit a model to the quantifier window predicting target looks from condition (*early* vs. *garden-path*), time, and their interaction. As in Exp. 4a, only the main effect of time reached significance (β = 0.18, SE = 0.02, $p < .0001$); neither the main effect of condition (β = -0.03, SE = 0.1, $p < .72$) nor the interaction (β = 0.03, SE = 0.05, $p < .54$) reached significance. This remained true when controlling for target set size and quantifier. Given the absence of any differences between these two conditions, I collapsed over *early* and *garden-path* trials when performing analyses on the quantifier window.

***Early* vs. *late* point of disambiguation** To determine whether participants were using information from the quantifier in the *early* and *garden-path* but not in the *late* condition, one model each was fit to the *early*/*garden-path* and *late* subsets of the data. Target looks were predicted from a predictor coding time window, where time window was either the baseline window (the 400ms window ranging from quantifier onset minus 200ms to the start of the quantifier window) or the quantifier window. If participants were making use of the quantifier, there should be a significant effect of time window. As in Exp. 4a, the effect of time window was significant for the *early* condition (β = 0.42, SE = 0.05, $p < .0001$) but not for the *late* condition (β = 0.05, SE = 0.08, $p < .53$), suggesting that participants were using the quantifier information when it could be used to disambiguate the

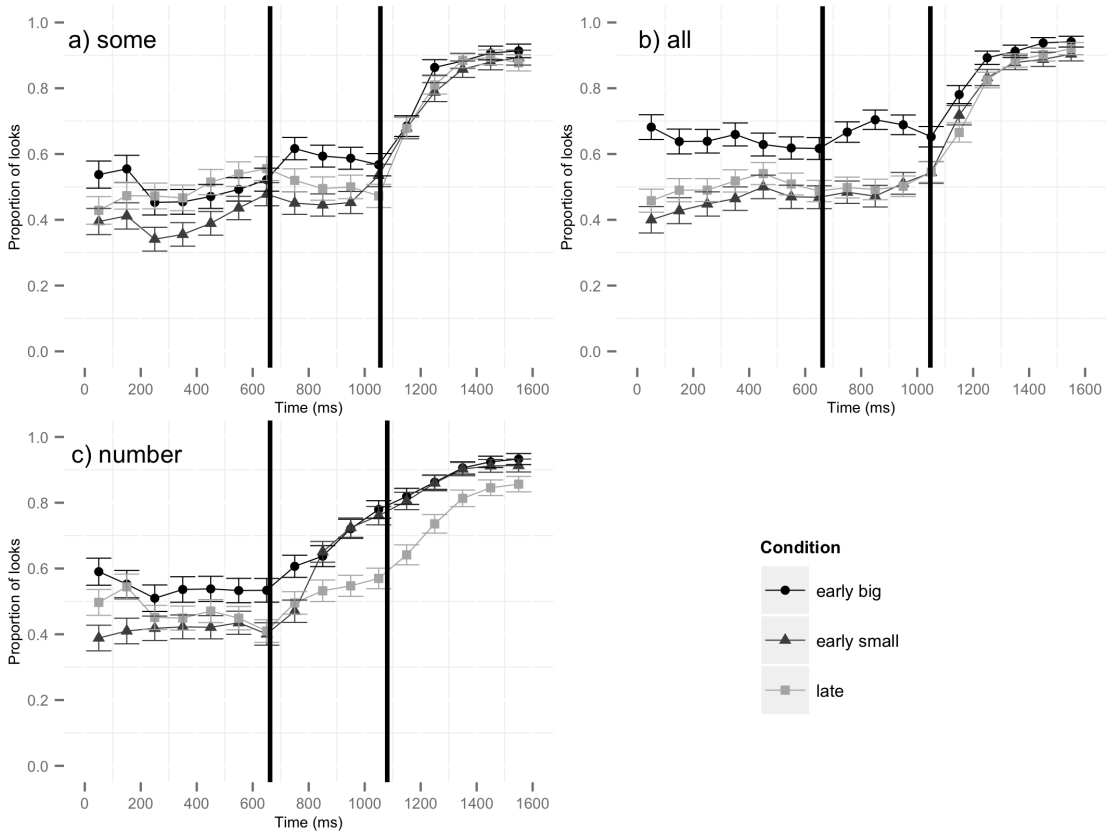

Figure 4.6: Proportion of looks to the target by quantifier and target set size for *early* and *late* conditions. Number terms are collapsed into *number* condition. Black vertical lines mark the quantifier window. The x-axis shows time in ms relative to auditory stimulus onset.

target, but not when it couldn't. Proportions of looks to the target for *early* and *late* conditions are shown in Figure 4.6.

Naturalness effects I next investigated whether there were differences between quantifiers and set sizes in how quickly the target was identified. In particular, visual inspection of Figure 4.6 suggests that there was a large increase in looks to the target in the quantifier window in both *big* set and *small* set number conditions, but the effects are much more attenuated for *all* and *some*. To investigate the difference in looking behavior between *some* and *all*, a model was fit to the *early* and *garden-path* subset of the data in the quantifier window, predicting tar-

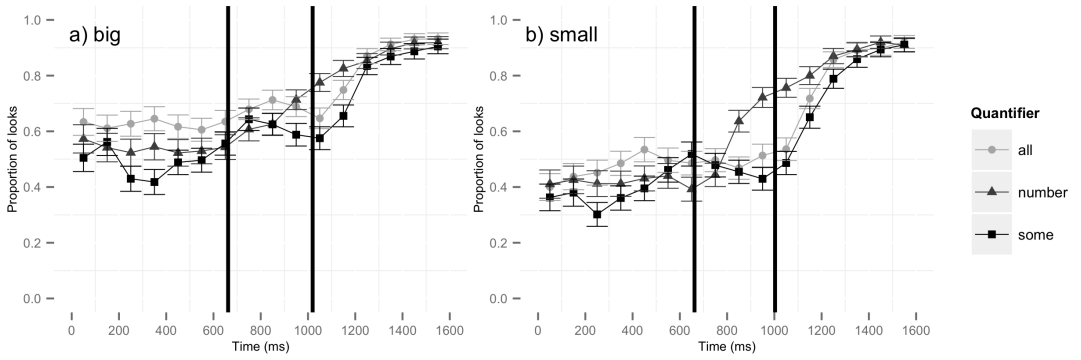

Figure 4.7: Proportion of looks to target on *early* trials for the *big* set (left) and *small* set (right) trials, by quantifier condition. First vertical black line indicates quantifier onset +200ms, second vertical black line indicates mean adjective onset +200ms.

get looks from fixed effects of quantifier (*all* vs. *some*), target size (*big* vs. *small*), time (continuous), and their interactions. Proportions of looks to the target in the *early* condition for *big* and *small* target sets by quantifier are shown in Figure 4.7.

The target size effect from Exp. 4a was replicated: participants looked more to the target when it was a big set than when it was a small set (β = -0.87, SE = 0.14, $p < .0001$). In addition, there was a main effect of quantifier such that participants were more likely to look to the target after *all* than after *some* (β = -0.09, SE = 0.04, $p < .05$). Thus, unlike in the numbers *absent* condition, *some* was delayed relative to *all*. Finally, there was a significant interaction between quantifier and target set size ($\beta = 0.19$, $SE = 0.07$, $p < .01$). Inspecting the simple effects model shows that the effect of quantifier is larger for the *big* set than for the *small* set: the advantage for *all* over *some* was greater when the target set was *big*. This is surprising given the naturalness effects from Exp. 3b, where there was an overall advantage for *all* over *some*, regardless of target set size. One explanation for this may be that participants' baseline looking preferences differed between quantifiers in this experiment. This can be seen visually in Figure 4.7. The difference in baseline fixation probability to the target within set size is

puzzling at first sight, given that counterbalancing was performed on how often each quantifier occurred with each set size, how often each set size was the target set size, and how often partitioned vs. unpartitioned sets were the target.

Participants' baseline preferences were investigated further by conducting mixed effects logistic regressions predicting looks to the target over looks to target and competitor in the window from display 2 onset to quantifier onset on trials where there was a contrast between a big and a small set in the lower chamber (*early* and *garden-path* trials, shown in Figure 4.8). The models contained fixed effects of target set size (*big* vs. *small*), target set partitioning (*partitioned* vs. *unpartitioned*), and their interaction. One model each to the Exp. 4a (numbers *absent*) and Exp. 4b (numbers *present*) data.

In the numbers *absent* model, there was only a main effect of target size such that participants preferred to look to the big set rather than the small set before quantifier onset (β = -0.70, SE = 0.16, $p < .0001$). This effect replicated in the numbers *present* model (β = -0.81, SE = 0.11, $p < .0001$) but there was also a main effect of target set partitioning such that participants looked more towards unpartitioned rather than unpartitioned sets (β = -0.42, SE = 0.12, $p < .001$).

The simple big set bias in Exp. 4a does not lead to differences in baseline preferences between quantifiers because the number of times each quantifier occurred with a big and a small set was counter-balanced. However, the added bias to prefer to look at unpartitioned sets in Exp. 4b did affect baseline preferences. In particular, *all* was always used with unpartitioned sets while *some* was always used with partitioned sets. Thus, these baseline preferences will lead to more looks to the target in the *all* than in the *some* conditions (because of the bias to look at unpartitioned sets). This bias might also account for the greater advantage for *all* with big sets (where the *all*-target was both a big and an unpartitioned set) compared to small sets (where the *all*-target is only an unpartitioned set).

In sum, comparing these results to those from Exp. 4a, we see that while there

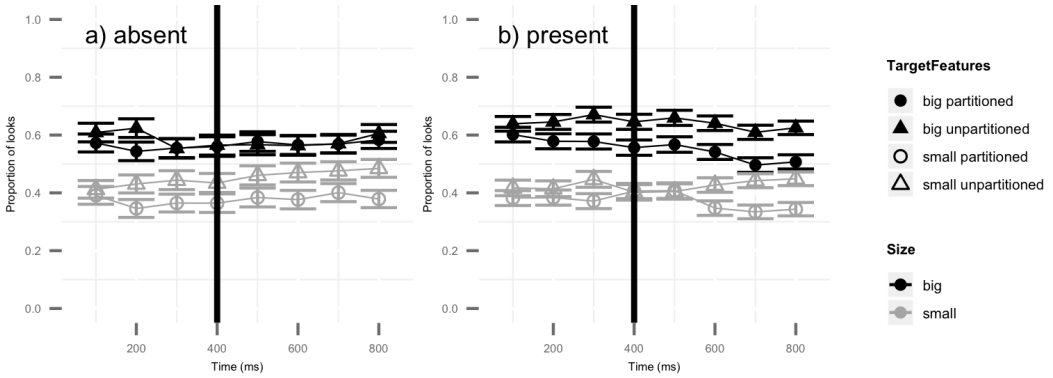

Figure 4.8: Proportion of looks to the target before quantifier onset when numbers were *absent* (left, Exp. 4a) and *present* (right, Exp. 4b). Vertical black line indicates onset of auditory stimulus. The x-axis shows time in ms relative to visual display change.

was no delay in looks to the target after *some* compared to *all* when numbers were absent, including number terms introduced a delay for *some*. Thus, Exp. 4a replicated the results from Grodner et al. (2010) (numbers *absent*), while Exp. 4b replicated the results from Huang and Snedeker (2009) (numbers *present*). However, introducing number terms also led participants to develop strategies for visually inspecting the display even before quantifier onset.

Introducing number In the comparison of Exps. 3a and 3b, introducing number led to an overall reduction in naturalness for both *some* and *all* in addition to rendering *some* less natural than *all*. To investigate the effect of introducing number on looking behavior in the quantifier window, a model was fit to *early* and *garden-path* subset of the data in the quantifier window predicting target looks from quantifier (*all* vs. *some*), target set size (*big* vs. *small*), and number presence (*absent* vs. *present*). The significant main effects of quantifier and target size found in the naturalness ratings replicated, such that there were more looks to the target for the big set (β = -0.70, SE = 0.08, $p < .0001$) and for *all* (β = -0.06, SE = 0.03, $p < .05$). However, while the main effect of number presence was numerically in the predicted direction, it failed to reach significance (β =

-0.11, SE = 0.08, $p < .20$). Instead, there was a three-way interaction between quantifier, set size, and number presence. Conducting separate analyses for big and small sets revealed that when the target set was big, there were more looks for *all* as well as an interaction between quantifier and number presence such that there was a trend for participants to look less to the target when numbers were present after *some* (β = -0.24, SE = 0.14, $p < .11$), but not after *all* (β = 0.24, SE = 0.17, $p < .15$). In contrast, when the target set was small, there was only a marginal main effect of number presence such that there were fewer looks to the target when numbers were present regardless of the quantifier (β = -0.23, SE = 0.08, $p < .08$). That is, introducing number led to delays for both *big* and *small* set *some* and for *small* set *all*, but not for *big* set *all*. One post hoc explanation is that *all* has a general preference to be used with big sets (as can also be seen in the naturalness data from Exp. 3a). If the bias for *all* to be used with a big set is strong enough, it might have trumped any possible interference effect from numbers.

Visual inspection of the number term effect as visualized in Figure 4.9 suggests that the strongest effects of number begin at the end of the quantifier window and last until around 200ms after the end of the quantifier window. Thus, the same model just reported was fit to the 200ms window after the quantifier window. The results were striking. While the target size effect remained significant (β = -0.33, SE = 0.07, $p < .0001$), the difference between *some* and *all* disappeared (β = -0.04, SE = 0.02, $p < .12$). Instead, there was a highly significant main effect of number presence such that the presence of number terms decreased the log odds of looking to the target (β = -0.29, SE = 0.07, $p < .0001$). No other effects reached significance.

Semantic vs. pragmatic responders Differences in looking behavior between pragmatic and semantic responders were analyzed both in the quantifier window

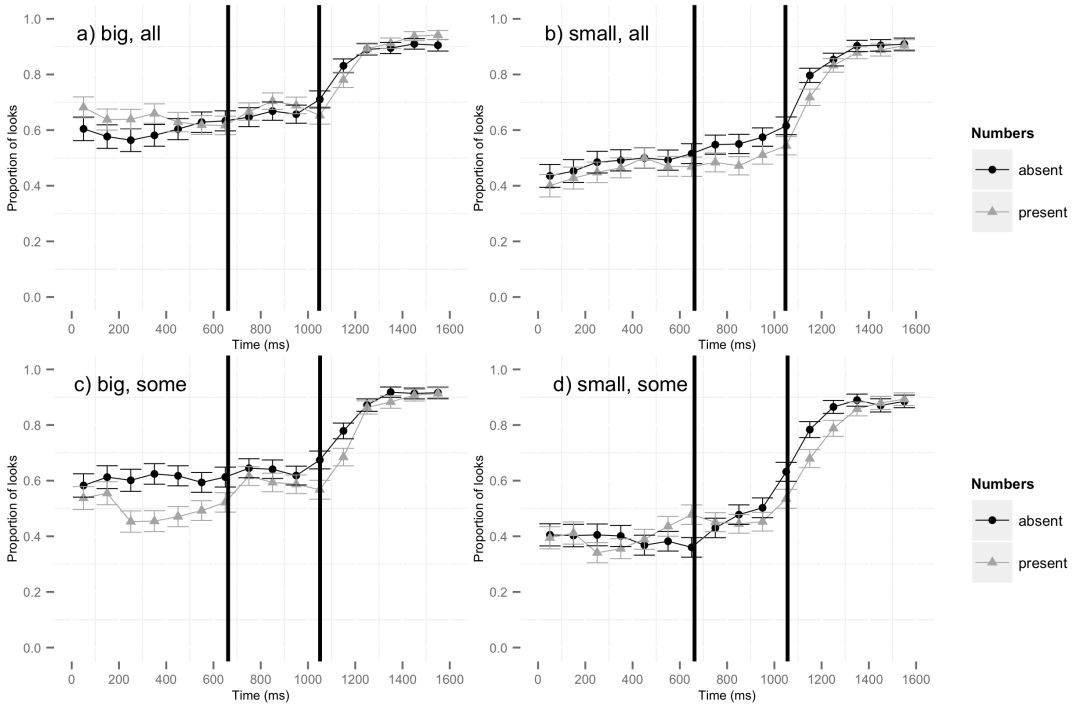

Figure 4.9: Number presence effect by early target size and quantifier conditions. Black vertical lines mark the quantifier window. The x-axis shows time in ms from auditory stimulus onset.

and in the 200ms window immediately following the quantifier window. The same model was fit to the data in both windows, predicting target looks from quantifier (*all* vs. *some*), target set size (*big* vs. *small*), and number presence (*absent* vs. *present*). In addition, all of these terms were allowed to interact with responder type (pragmatic vs. semantic). Given the results from Exp. 2, pragmatic responders are predicted to exhibit the naturalness effects found in Exps. 3a and 3b more clearly than semantic responders.

Unfortunately, there were very few semantic responders (14, of which only 4 were in the numbers *absent* condition), compared to 55 pragmatic responders. Seven inconsistent responders were excluded from the analysis. There was a robust main effect of responder type in the quantifier window such that semantic responders looked less to the target in the quantifier window than pragmatic responders (β = -0.37, $SE = 0.16$, $p < .05$). However, with this small data set, the model was unable to detect any finer potential differences between semantic and pragmatic responders' behavior to different quantifiers or set sizes. The target set size and quantifier effect remained robust, such that there were more looks to the target for the big set (β = -0.70, $SE = 0.08$, $p < .0001$) and for *all* (β = -0.06, $SE = 0.03$, $p < .05$).

In the 200ms window directly following the quantifier window, there was both a main effect of number presence (β = -0.39, $SE = 0.13$, $p < .01$) and an interaction between number presence and responder type ($\beta = 0.70$, $SE = 0.33$, $p < .05$). Inspecting the simple effects model yields an effect of number presence for pragmatic responders in the predicted direction (β = -0.54, $SE = 0.14$, $p < .001$) but not for semantic responders ($\beta = 0.16$, $SE = 0.30$, $p < .60$). The difference in number presence effect for semantic vs. pragmatic responders is visualized in Figure 4.10. Pragmatic responders initially make faster use of the quantifier as a cue to the target than semantic responders. However, in the window where the number presence effect is strongest, the effect is driven exclusively by pragmatic

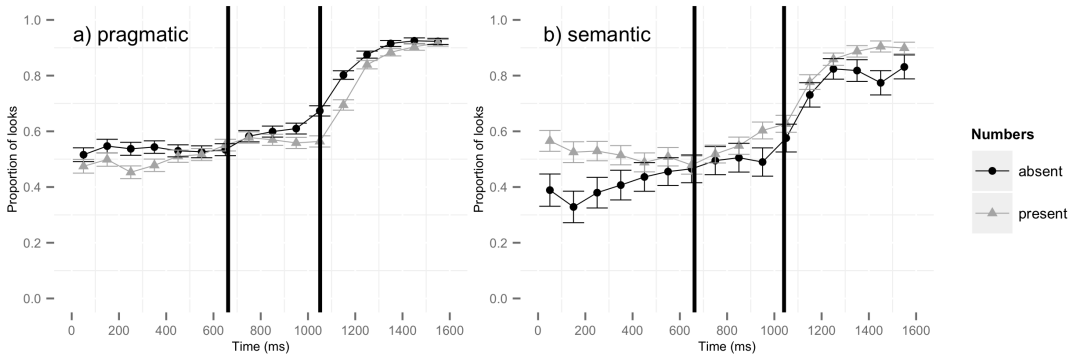

Figure 4.10: Proportion of looks to the target for pragmatic (left) vs. semantic (right) responders, collapsing over quantifier and target set size. Black lines mark the quantifier window. The x-axis shows time in ms relative to auditory stimulus onset.

responders. When numbers are present, pragmatic responders' increase in looks to the target is delayed relative to semantic responders', despite the latter being initially less likely to look to the target. This suggests that semantic responders are employing a strategy of waiting for the disambiguating information from the adjective, and then use that information immediately without any residual uncertainty as to the target. Pragmatic responders on the other hand initially use the information from the quantifier but are also more delayed by the availability of alternatives.

4.6 Discussion of Exps. 3 and 4

Exps. 3 and 4 demonstrate that the naturalness of using the quantifier *some* in different contexts affects the speed with which listeners generate a scalar implicature. I summarize the results from the two naturalness (3a and 3b) and the two eye-tracking (4a and 4b) experiments. I group the results into three categories: (1) between-quantifier naturalness effects; (2) within-quantifier naturalness effects; and (3) differences between semantic and pragmatic responders. I discuss the implications of these results for Literal-First, Default, and Constraint-Based

accounts of scalar implicature processing.

4.6.1 Between-quantifier naturalness effects

When numbers were absent, participants generated a scalar implicature from *some* to *not all* immediately after observing the quantifier. They were just as fast to start converging on the target set of gumballs after hearing *some* as they were after hearing *all*. In contrast, when number terms were present there was a general advantage for *all* over *some* immediately after the quantifier, suggesting that generating the implicature was delayed. The number *absent* results are compatible with the Default model, which predicts rapid, effortless implicatures, but not with the Literal-First hypothesis, which predicts slow, costly implicature. In contrast, the number *present* results are compatible with the Literal-First but not with the Default model. Thus neither of the two accounts that assume that initial processing is context-independent can account for the full set of data.

In contrast, the Constraint-Based account allows for the early stages of pragmatic processing to be affected by contextual constraints. The constraint I have focussed on in this chapter is the naturalness of *some* and its plausible alternatives. When numbers were *absent*, the naturalness of *some* and *all* did not differ (except for the big set, where *all* was more natural) and indeed this is where participants rapidly started converging on the target for both *some* and *all*. When numbers were *present*, *all* was more natural than *some*. This naturalness difference was also reflected in the eye movements.

4.6.2 Within-quantifier naturalness effects

Immediately after the quantifier, there were fewer looks to the target when numbers were present than when they were absent. This was true for both *big* and *small* set *some* but only for *small* set *all*. After the quantifier window (i.e., once

information from the adjective was available), there was a large delay in looking to the target for all quantifiers and set sizes when numbers were *present* compared to when they were *absent*.

Neither the Default nor the Literal-First accounts predict that differences in eye movements after observing a quantifier should depend on the presence or absence of number terms. In contrast, the Constraint-Based account is consistent with this result. In particular, introducing number might increase the granularity of the scale that listeners expect for descriptions of scenes. Having number terms available means that speakers can use an exact number term for set size. The number presence effect suggests that listeners are aware that there are better alternatives that the speaker could have used but chose not to. Note, however, that this effect was not restricted to *some* - reasoning about alternatives also affected the interpretation of *all*.

As can be seen clearly in Figure 4.7, while there are small early effects for *all*, both *some* and *all* are delayed relative to the speed with which participants converge on the target after hearing the better, number alternatives.

4.6.3 Semantic vs. pragmatic responders

Finally, both the response time and eye movement patterns differed between semantic and pragmatic responders. Semantic responders were slower to converge on the target overall (i.e., regardless of quantifier and target set size). This suggests that they relied more heavily on the adjective for interpreting the description. In contrast, pragmatic responders made early use of the quantifier. Including number terms affected pragmatic responders but not semantic responders.

This effect was also present in the response time data. Thus, pragmatic responders are overall more susceptible to naturalness differences than semantic responders, mirroring the finding in Exp. 2.

4.6.4 Response time results

An interesting methodological point arises from comparing the response time and eye movement results. In response times, despite not urging participants to respond as quickly as possible, previous findings that semantic responses to *some* are processed more quickly than pragmatic responses (e.g. Bott & Noveck, 2004) were replicated here. I argued that this could be explained either by positing an initial stage of semantic processing before pragmatic enrichment, but it could also reflect a more complex verification process associated with the *some but not all* interpretation. The data presented here support the latter interpretation: while semantic responses were faster than pragmatic responses even when numbers were absent, participants' eye movements indicated that the implicature was generated without a delay. This suggests that the response time difference indeed stems from verification difficulty rather than from costly implicature computation. This is the first scalar implicature study that has studied semantic and pragmatic responders' response times and eye movements in tandem.

4.6.5 Conclusion

These studies, in combination with Exps. 1 and 2, paint a detailed picture of how scalar implicatures are processed: both the robustness and the speed with which listeners settle on either an upper-bound or lower-bound interpretation of utterances containing *some* depend on multiple contextual cues to the speaker's intention. The focus in the past two chapters has been on investigating the effect of the naturalness of *some* and its scalar and non-scalar alternatives on the time course of scalar implicature processing: the speed with which listeners arrive at the upper-bound interpretation is highly dependent on whether there is a "better" - more natural, more expected - way of describing the world than by using a weak scalar term. The scalar implicature is computed more slowly as the relative

naturalness of alternative utterances increases.

Accounts of scalar implicature that do not allow for context to affect the early stages of processing, such as the Default and the Literal-First model, cannot explain these results. Instead, the results are consistent with contextualist accounts of scalar implicature like the Constraint-Based account presented in Section 2.2.2 that do not posit an initial stage of context-independent processing.

As we better understand the relevant contextual constraints and the distribution of usage of pragmatic *some*, it is likely that pragmatic inference, like ambiguity resolution, will turn out to be consistent with approaches to language processing that are grounded in constraint-based and information theoretic principles. In the next chapter, I make an attempt at quantifying some of these distributional properties of scalar implicatures arising from the use of *some* in spontaneous speech.

5 Distributional properties of *some* scalar implicatures in naturally occurring speech

5.1 Introduction

This chapter consists of an investigation of some of the distributional properties of scalar implicatures in naturally occurring speech. The goals are two-fold: first, I provide an estimate of the frequency of scalar implicatures from *some* to *not all* by collecting implicature ratings for naturally occurring instances of *some*. This constitutes a test of the Frequency Assumption introduced in Section 1.2. Second, I explore the effect of three different cues on scalar implicature strength, as measured by the aforementioned implicature ratings. The cues under investigation are a) syntactic partitivity of the *some*-NP, b) quantifier strength, and c) discourse accessibility of the *some*-NP, which includes c.i) linguistic mention of the embedded NP referent, c.ii) topicality of the *some*-NP, and c.iii) modification of the head of the *some*-NP.

In the remainder of this Chapter, I first reiterate the Frequency Assumption, provide likely reasons for why it has thus far not been tested, and discuss briefly its importance as a central assumption of the Default model and in the interpretation

of "costly implicature" results as support for the Literal-First hypothesis in the remainder of Section 5.1. Next I present the experiments in Section 5.2 and Section 5.3. Finally, I discuss the implications of the results in detail in Section 5.4.

5.1.1 The Frequency Assumption

Recall the Frequency Assumption, formulated in (18) and repeated here as (45).

(45) The Frequency Assumption (lexical version)
Scalar implicatures arise more often than not when a scalar item is used.

The Frequency Assumption features prominently in two areas of the scalar implicature literature: i) it is one of the basic assumptions of the Default model of scalar implicatures (Levinson, 2000) and ii) it is an (often implicit) assumption made by many researchers interpreting experimental results from experiments that test the time course of scalar implicature processing. In particular, it is an important assumption for interpreting "costly implicature" effects as support for the Literal-First hypothesis. In Section 1.2 I discussed how the Frequency Assumption most likely came about; I pointed out that there is much intuitive support for the Frequency Asssumption, and yet there is a conspicuous lack of attempts to find empirical support for it. In the following, I discuss some of the likely reasons for the lack of tests of the Frequency Assumption. I defer discussion of the theoretical importance of the Frequency Assumption until Section 5.4.

There are at least two possible reasons for why the Frequency Assumption has not previously been tested. The first is that it has received so much intuitive support that there is no need to test it. The second is that it is *hard* to test. I discuss each of these in turn and briefly delve into the theoretical importance of testing the Frequency Assumption.

The Frequency Assumption has received much intuitive support based on considerations of regularity discussed in Section 1.2. However, intuition may be misleading. That in and of itself is motivation for testing the Frequency Assumption. In addition, there is a conceptual problem with the Frequency Assumption: it is not entirely clear at what level the Frequency Assumption is supposed to hold. Does it hold only for lexicalized scales or are ad hoc scales included in the assumption, and is there a clear boundary between the two?

I will not attempt to address these questions. Instead, I will take the case for which the Frequency Assumption is most likely to hold: the case of scalar implicatures from *some* to *not all*, which constitutes the prime example of scalar implicatures arising from a lexicalized scale. That is, I will attempt to address the following version of the Frequency Assumption:

(46) The Frequency Assumption (*some* version)
Scalar implicatures arise more often than not when the lexical item *some* is used.

Note that I am tying the Frequency Assumption to the surface form *some* rather than a particular use of *some*. This may strike some as odd and as an unfair characterization of the Frequency Assumption because it is well-known that there are different uses of *some*, some of which have been argued to not give rise to implicatures. For example, the strong, quantificational *some* which allows stress as in (47a) does allow for implicatures, while the weak, indefinite, often unstressed *sm* as in (48a), it has been argued, does not. Both of these occur with mass or plural count nouns, but *some* can also occur with singular count nouns as in (49a), where the implicature is ruled out syntactically.

(47) a. Some parents are great.
b. Some, but not all, parents are great.

(48) a. My son needed sm money.

b. ? My son needed sm but not all money.

(49) a. Some guy predicted the end of the world today.

b. * Some, but not all, guy predicted the end of the world today.

It seems clear that the *some*s of (48a) and (49a) should not count towards the Frequency Assumption. Indeed, the strong use of *some* may be one of the factors Horn had in mind when he restricted his version of the Frequency Assumption to "unmarked contexts". However, I will continue to speak of the Frequency Assumption about *some* rather than about *some* conditioned on a particular use of *some*, for the following reasons. First, in practice it is hard to draw a clear boundary between different uses of *some*. This is discussed in more detail in Section 5.3.2. Second, from the listener's (or reader's) perspective, instances of *some* do not come tagged as $some_{\text{strong}}$, $some_{\text{weak}}$, etc. Rather, the use of *some* (and its associated potential for giving rise to a scalar implicature) is something that listeners need to infer from context. I take the listener's perspective here and ask how, given only the linguistic signal and information about context (but not tagged uses of *some*), listeners arrive at an upper- or lower-bound interpretation.

This brings us to the second reason for why the Assumption has not been tested previously: testing the Frequency Assumption is a difficult enterprise even when reduced to only *some-not all* implicatures. The formulation of the Frequency Assumption so far suggests that there is *something* we can count; there is a frequency of some categorical entity that we intend to determine. While it may seem straightforward that what we want to count is the number of times a speaker's use of *some* gives rise to a scalar implicature, two immediate methodological difficulties arise.

The first problem involves the definition of an implicature: implicating is something that speakers do. While successfully implicating requires the listener to make

the correct inferences about the speaker's intended meaning, the crucial component of an implicature is that speakers intend to convey the implicated content in part by the listener recognizing that intention. A listener may draw a scalar inference in a case where a speaker did not intend for him to do so.[1] Thus, simply studying a listener's interpretation of utterances with *some* is not enough. To study the distribution of scalar *implicatures* rather than scalar *inferences*, it is necessary to have access to the speaker's intentions. Empirically, this is an easy problem for armchair linguistics, where we can imagine ourselves and our own intentions in producing certain utterances. However, the situation is more difficult if we want to collect implicit distributional information about implicatures experimentally from multiple participants. Intuitively, one would want to provide participants with a) the speaker's communicative intention, b) the speaker's utterance, and c) the listener's interpretation, in order to arrive at assessments of whether the speaker succeeded in communicating his intention (including having the right (scalar) implicatures go through).

Here, I make the simplifying assumption that inferences about speaker meanings are usually correct. That is, I will assume that when a listener makes a scalar inference, it was so intended by the speaker. The extent to which this is a simplifying assumption is unclear, but one argument in favor of there being only few mismatches between speaker meanings and listener interpretations is that commu-

[1] An example made famous by Larry Horn is one from the movie *When Harry Met Sally*, where the following dialog takes place between Harry and his friend Jess, whom Harry is trying to set up with Sally. Jess has just pointed out that by saying that Sally has a good personality, Harry must mean that she is not attractive. Harry disagrees and goes on to say:

(50) Harry: [For all I have said,] she could be attractive with a good personality, or not attractive with a good personality.
Jess: So which one is she?
Harry: Attractive.
Jess: But not beautiful, right?

Here, Jess makes the scalar inference that Sally is attractive, but not beautiful, which Harry did not intend for him to make. Therefore, Harry cannot be said to have implicated that Sally is not beautiful. Thus, in this example there is a scalar inference (on Jess's part) but no scalar implicature (on Harry's part).

nication would not be robust if listeners routinely made wrong inferences. Since communication usually proceeds without problems (i.e., misunderstandings about scalar inferences like the one in Footnote 1 are rare), it is likely that listeners' inferences usually reflect what speakers intended them to infer.

In the experiment presented in Section 5.2, implicature ratings were collected from participants who observed only a speaker's utterance. That is, we will only have access to b) and c) from above (the speaker's utterance and the listener's interpretation), but not a) (the speaker's intention). Thus, generalizations that are drawn will be generalizations about scalar *inferences* rather than *implicatures*; and they will be particularly about scalar inferences from *some* to *not all*.

The second challenge for testing the Frequency Assumption stems from the categorical treatment of scalar implicatures in most of the literature: typically, scalar implicatures are treated as an all-or-none phenomenon, whereby a speaker either does or does not intend to convey the upper-bound interpretation or a listener does or does not take the speaker to implicate it (Gazdar, 1979; Horn, 1984). In practice, however, implicatures seem to be rather a matter of degree; in fact, one of the basic properties of scalar implicatures is precisely that they are *implicatures* - they are not part of what is said, and they can be canceled. Their content is not explicitly encoded. This suggests that there is more uncertainty about implicated content than there is about coded content[2]; another way to think about this is that listeners take a speaker to implicate X *to a certain degree* or with *a certain probability*, rather than taking them to either implicate X or not.

The Constraint-Based and probabilistic accounts of scalar implicature introduced in Chapter 2 treat scalar implicature as a gradient, rather than as an

[2]Grice himself seems to have had something not entirely dissimilar from this in mind when he wrote, at the end of "Logic and Conversation", that there may often be multiple explanations for a speaker's utterance (where an "explanation" here is the result of calculating "what has to be supposed in order to preserve the supposition that the Cooperative Principle is being observed", Grice, 1975, p. 58). When there are multiple explanations, Grice observes, "the implicatum will have just the kind of indeterminacy that many actual implicata do in fact seem to possess" (Grice, 1975, p. 58).

all-or-none phenomenon. Listeners are taken to have some degree of belief in the implicated content after inferring a belief distribution over states of the world by integrating information from the linguistic signal, knowledge of alternative utterances, and knowledge about how likely the speaker is to produce a particular alternative in a particular context (among other things).

In collecting implicature ratings, I side with these probabilistic approaches. Rather than asking whether or not a scalar implicature arises, I ask to what *degree* it arises. This constitutes a departure not only from most of the theoretical literature on scalar implicatures, but also from most previous experimental approaches (but c.f. Degen et al., 2013; M. C. Frank & Goodman, 2012; N. D. Goodman & Stuhlmüller, 2013) and from the treatment of scalar implicature in Chapters 3 and 4, where I made the simplifying assumption that listeners ultimately settle on either the upper-bound or the lower-bound interpretation.

One of the main assumptions of the Constraint-Based account is that the strength of an implicature depends on the amount of probabilistic contextual support for the implicature. That is, the more converging evidence for the implicature is provided by contextual cues, the stronger the implicature should be. In Chapters 3 and 4, I tested the prediction of the Constraint-Based account that the *speed* of scalar implicatures depends on multiple contextual constraints. Here I test the prediction that scalar implicature *strength* should be probabilistically modulated by multiple contextual cues.

I will begin by presenting the results of a corpus study in which implicature ratings for utterances containing *some* were collected in order to test the Frequency Assumption (Exp. 5). In a second step, implicature ratings were correlated with a number of syntactic, semantic, and pragmatic contextually available cues (Exp. 6). The theoretical importance of the Frequency Assumption and the implications of the experimental results are discussed at length in Section 5.4.

5.2 Exp. 5: frequency of scalar implicatures

The studies were conducted in three steps. First, a database of *some*-NPs was generated by extracting all instances of utterances containing the word *some* from the Switchboard corpus. Second, for each instance of *some* an implicature rating was obtained in a web-based study. Finally, the effect of the different cues on the probability of generating a scalar inference was investigated in regression analyses predicting implicature ratings from the cues.

I first present the database of *some*-NPs that was created. Next, I report the methodology for obtaining implicature ratings for each case in the database. Finally, in Exp. 6 I present three substudies corresponding to the three cues mentioned above, including the theoretical motivation for studying each cue.

5.2.1 The database

I used TGrep2 (D. Rohde, 2005) to extract all 1748 occurrences of *some*-NPs that were not part of a disfluency from the Penn Treebank (release 3, M. P. Marcus, Santorini, & Marcinkiewicz, 1993; M. Marcus, Santorini, Marcinkiewicz, & Taylor, 1999) subset of the Switchboard corpus of telephone dialogues (Godfrey, Holliman, & McDaniel, 1992). The corpus contains approximately 800 thousand spoken words in over 100 thousand utterances from about 650 telephone dialogues on various topics between two participants who did not know each other. The TGrep2 Database Tools (Jaeger, 2006; Degen & Jaeger, 2011) were used to organize the *some*-utterances into a database.

Because only those cases that do not syntactically prohibit a scalar implicature were interesting for the purpose of the study, 359 cases (20.5%) of *some*-NPs headed by singular count nouns were excluded.[3] In a *some*-NP, singular count

[3]However, in the grand scheme of things one would not want to exclude these cases of *some*, but rather include head noun number as a cue that listeners can use to restrict their

nouns are compatible with two different meanings. The more common meaning is the specific indefinite reading, which cannot give rise to a scalar inference (see example (51)). Singular count nouns in *some*-NPs can, however, also receive a coerced interpretation as shown in (52a). Under this reading, the implicature, made explicit in (52b) is possible, but these cases seem to be very infrequent (e.g., in a random sample of 50 singular count noun *some*-NPs, only three were cases of coercion, and they all occurred in the partitive, as in (53)).

(51) a. She stuck my name on some list.

b. *She stuck my name on some, but not all list.

(52) a. John kicked some cat off the street.

b. John kicked some, but not all cat off the street.

(53) Well, I had some of that problem.

A further 26 cases where the *some*-NP consisted only of *some* were also excluded:

(54) Some say that coffee is healthy.

This was done because for these cases it is not possible to investigate the effects of the discourse accessibility cues tested in Section 5.3.3, which assumes that *some* occurs with an overt NP. However, it is worth noting that in these cases the implicature seems to generally go through.

After the exclusion, 1363 cases of utterances containing *some*-NPs remained. For these cases, implicature ratings were collected in a web-based study, which I report next.

---

interpretation of *some* - that is, a singular count noun can be seen as a strong, but nevertheless probabilistic, cue *against* the upper-bound interpretation.

5.2.2 Collecting implicature ratings

Implicature ratings were collected using Amazon's Mechanical Turk service. In accordance with probabilistic views of scalar implicature, rather than asking categorically whether or not a speaker intended to convey the negation of the stronger alternative, participants were asked to provide gradient judgments. These gradient judgments can then be interpreted as a measure of scalar inference strength, given the amount of probabilistic support for it in context.

Methods

Participants 243 participants were recruited on Amazon's Mechanical Turk and paid $0.80 for each block of 10 items. Participants who completed at least three blocks received a one-time bonus of $0.20.

Procedure and materials On each trial, participants saw an utterance containing a *some*-NP (the *target utterance*) together with ten utterances from the immediately preceding discourse context (or until the beginning of the dialogue if there were fewer than ten utterances in the previous context). The target utterance was presented in red. Below the target utterance, an almost identical utterance (the *comparison utterance*) was presented which differed only in that the implicature was made explicit by inserting *but not all* after *some*. The comparison utterance was presented in green font. Two example pairs of (a) target and (b) comparison utterances are shown in (55) and (56).

(55) a. I like, I like to read some of the philosophy stuff.

b. I like, I like to read some, but not all of the philosophy stuff.

(56) a. And I'll take some time and do that with her.

b. And I'll take some, but not all time and do that with her.

Participants were then asked: "How similar is the statement with 'some, but not all' (green) to the statement with 'some' (red)?" They provided similarity judgments on a seven point Likert scale where 1 was "very different meaning" and 7 was "same meaning". The more the implicature is part of the speaker's originally intended meaning, the less of a difference explicitly encoding the content of the implicature should make, and the higher the similarity judgments are expected to be. Conversely, if the content of the implicature was not part of the speaker's originally intended meaning, making it explicit should lead to a larger perceived shift in meaning and the two utterances should be rated as very dissimilar.

Participants first completed two practice trials to become acquainted with the task and understand the range of the scale before completing the experimental trials. One of the practice trials was a clear case of an implicature, shown in (57), while the other one, shown in (58), clearly could not give rise to an implicature.

(57) a. I had some of the banana yogurt.

b. I had some, but not all of the banana yogurt.

(58) a. There are probably some peanuts in the pantry.

b. There are probably some, but not all peanuts in the pantry.

Participants were told that cases like (57) should receive a high rating and cases like (58) should receive a low rating, but were not instructed on which particular value to assign.

Items were divided into blocks of 20 items each. Each block was rated by ten participants. 11 items appeared in two different blocks in order to ensure that each block consisted of 20 items. Because of this, most items received 10 ratings each and 11 items received 20 ratings each.

Table 5.1: Distribution of participants (bottom row) over completed number of blocks (top row).

1	2	3	4	5	6	7	8	10	12	14	16	18	18	20	22	24	26
26	120	3	12	2	35	1	13	6	3	3	5	2	2	1	1	1	1

34	44	46	48	54	100
2	1	1	1	1	1

Results and discussion

The distribution of participants over number of rated blocks of items is shown in Table 5.1. Mean number of completed blocks per participant was 5.72 (median: 2).

Mean overall similarity rating was 3.9 (median: 4). The distribution of raw ratings and (aggregated) mean by-item ratings is shown in Figure 5.1. Under the Frequency Assumption, there should be more high than low ratings reflecting overall strong support for the pragmatic interpretation. However, only 44.7% of ratings were higher than the midpoint of the scale, while 46.6% of ratings were lower than 4. Looking only at the endpoints of the scale, only 14.7% of the data were highest ratings while 19% were lowest ratings. Thus, contrary to the Frequency Assumption, most utterances with *some* do not strongly support a scalar inference.

Examples from the lower, medium, and upper end of the scale are shown in (59) - (61) (numbers are mean similarity ratings):

(59) *Low similarity rating (little support for implicature)*

a. That would take **some planning**. 1.4

b. And this would give them a chance to have **some positive self-esteem**. 1.4

c. You sound like you've got **some small ones** in the background. 1.5

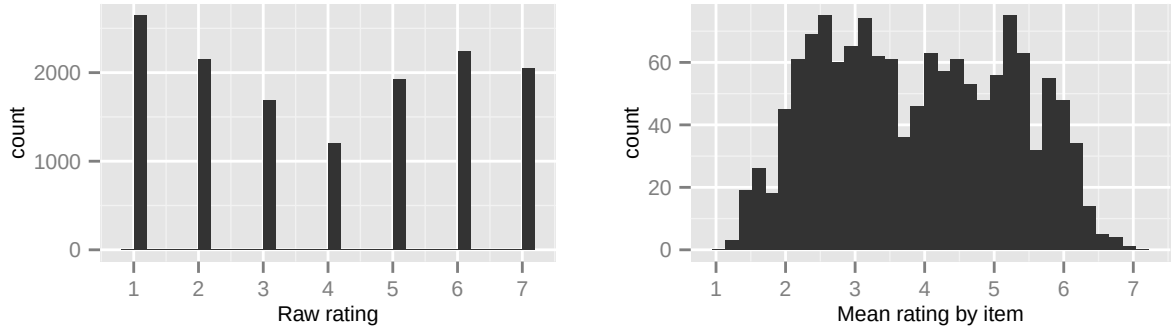

Figure 5.1: Distribution of raw ratings (left panel) and mean per-item ratings (right panel)

(60) *Medium similarity rating (medium support for scalar implicature; or ignorance implicature)*

a. And **some ways**, it might be kind of scary. 4

b. I'd love to have, have **some animals**. 4

c. It would certainly help them to appreciate **some of the things that we have here** 4

(61) *High similarity rating (much support for implicature)*

a. But I think that at **some times** it can be the right thing to do. 6.7

b. I sold **some of them**. 6.8

c. I like **some country music**. 6.9

A number of questions immediately arise. One question concerns the status of discrete interpretations. One way to interpret the distribution of implicature ratings is that there is an unstructured continuum from upper-bound to lower-bound interpretations of utterances containing *some*. An alternative possibility is that there are discrete, but noisy underlying interpretations giving rise to the distribution. In this case, good candidates for underlying interpretations are the standard lower-bound and upper-bound interpretations shown in (63) and (62). However, a third interpretation, the ignorance implicature as in (64), arises when

all that listeners take the speaker to know is that the speaker does not know *whether* the stronger alternative is true, as discussed in the Introduction (see Gazdar, 1979).

(62) Upper-bound interpretation: $K_s W \wedge K_s \neg S$

(63) Lower-bound interpretation: $K_s W$

(64) Ignorance implicature: $K_s W \wedge \neg K_s S$

Here, K_s stands for *the speaker knows*, which takes a propositional argument. W denotes the weak alternative, S denotes the strong alternative.

Note that participants were not asked to estimate whether the speaker was an authority on whether the stronger alternative was true. Intuitively, cases of ignorance implicatures should receive intermediate ratings, reflecting the listeners' uncertainty about whether the speaker definitely did or did not intend an upper- or lower-bound interpretation. However, there is no independent way to tease apart cases of ignorance implicatures from cases of standard lower- or upper-bound interpretations in this dataset without collecting in addition participants' estimates of speaker authority.

Barring that, we can test whether there are multiple underlying interpretations giving rise to the data by fitting finite Gaussian mixture models (Jacobs, Jordan, Nowlan, & Hinton, 1991; Richardson & Green, 1997) with different numbers of components to the data, and then determine the optimal model by using log-likelihood ratio tests. The finite Gaussian mixture model with k components may be written as:

$$p(y|\mu_1, \ldots, \mu_k, \sigma_1, \ldots, \sigma_k, \lambda_1, \ldots, \lambda_k) = \sum_{j=1}^{k} \lambda_j \mathcal{N}(\mu_j, \sigma_j), \tag{5.1}$$

where μ_j are the means, σ_j the standard deviations, λ_j the mixing proportions (which must be positive and sum to one), and $\mathcal{N}$ a Gaussian with specified mean

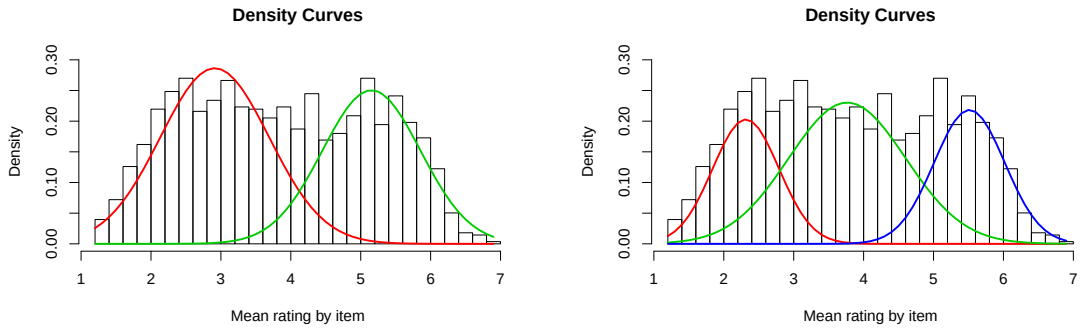

Figure 5.2: Density curves for models assuming two (left) or three (right) Gaussian components. The model on the right was identified as optimal in a log-likelihood ratio test.

and standard deviation. Observations $y = \{y_1, \ldots, y_n\}$ are scalar observations: the mean per-item similarity ratings in our dataset.

Separate models were fit to the mean by-item ratings for $k = 1, 2, 3, 4$ components using the `mixtools` package (Benaglia, Chaveau, Hunter, & Young, 2009) in R, which uses expectation maximization (EM) to estimate the optimal parameter values. Means (μ), standard deviations (σ), mixing proportions (λ), and log likelihood for each of the models with one, two, three, and four components are shown in Table 5.2. Density curves for the optimal two- and three-component models are shown in Figure 5.2.

The optimal two-component model fits the data significantly better than the optimal one-component model ($\chi^2(3) = 197.9, p < .0001$). This suggests that there are at least two underlying interpretations giving rise to the distribution over mean ratings. Let us assume that these two interpretations correspond to the lower- and upper-bound interpretation. We can now turn to the mixing proportions given by λ, which specifies the extent to which each of the components contribute to the data. We see that the lower-bound component contributes 56% of the data, while the upper-bound component contributes only 44%. That is, under the interpretation of the components as lower- and upper-bound, this con-

Table 5.2: Means (μ), standard deviations (σ), mixing proportions (λ), and log likelihood for each model. First column indicates number of components. P-values are derived from applying log likelihood ratio test comparing model in row to model in previous row.

			Component 1			Component 2			Component 3			Component 4		
Model	logLik	p	μ	σ	λ	μ	σ	λ	μ	σ	λ	μ	σ	λ
1	-2377		3.9	1.34	1									
2	-2278	$< .0001$	2.9	0.78	0.56	5.2	0.70	0.44						
3	-2254	$< .0001$	2.3	0.47	0.24	3.8	0.84	0.48	5.5	0.51	0.28			
4	-2251	$< .14$	2.1	0.39	0.16	3.0	0.54	0.29	4.3	0.62	0.29	5.6	0.48	0.26

stitutes further evidence against the Frequency Assumption: more of the data is captured by the lower-bound component.

However, the model fit is again significantly improved by adding a third component ($\chi^2(3) = 48.5, p < .0001$, see right panel of Figure 5.2), but not by adding a fourth one ($\chi^2(3) = 5.5, p < .14$). That is, the optimal model consists of three components, one at either end of the scale contributing 24% and 28% of the data each, and one covering mostly the center and contributing 48% of the data. How is this to be interpreted? One natural interpretation is that this large middle component corresponds to ignorance implicatures. That it has a relatively large variance compared to the other two components suggests that there is a lot of variation in how much uncertainty there is about whether the speaker is an authority on the truth of the stronger alternative. Let us entertain for the time being that these three components indeed correspond to the three interpretations in (62) - (64). Where does this leave us with regards to the Frequency Assumption?

The answer is that the Frequency Assumption becomes untenable. Under this interpretation of the components, only 28% of the data is generated by an upper-bound interpretation. This is an interesting and surprising result, given the prevalence of the Frequency Assumption about *some*. It suggests that the amount of probabilistic support for scalar implicatures is generally quite low. If we further take seriously the interpretation of the intermediate component as one reflecting ignorance implicatures, we can further conclude that most of the time, listeners are actually uncertain about whether speakers know whether the stronger alternative is true, and thus derive at most ignorance implicatures.

A second interpretation of the three components is as upper-bound (rightmost), lower-bound (middle), and weak *sm* (leftmost). Under this interpretation of the components, the upper-bound interpretation still contributes only 28% of the data; that is, the conclusion that support for scalar implicatures from *some* to *not all* is generally low remains unchanged. However, under this interpretation,

weak *sm* and lower-bound strong *some* would contribute different portions of the data. Ignorance implicatures would then presumably contribute to the noise in the lower-bound interpretations.

A compelling test of whether the former rather than the latter interpretation of the components is warranted would be to collect ratings about speaker knowledge. If the interpretation of the results as reflecting lower-bound, ignorance, and upper-bound interpretations is correct, participants' uncertainty about the speaker's knowledge should be largest at the mean of the component (around 3.8) and decrease towards each end of the scale.

Some readers may worry about the fact that under this interpretation, weak *sm* and lower-bound strong *some* are collapsed into one category. This worry may be justified; but it may also be that for listeners this distinction has no cognitive counterpart. In this case, collapsing the two uses of *some* into one category is exactly the right thing to do.

Finally, a caveat of these results is that the assumption of normality does not necessarily hold. However, it is nevertheless striking that the best model does return three components and that these three components have a natural interpretation as the three interpretations that have been discussed in the literature on scalar implicatures (Gazdar, 1979; Levinson, 2000).

The interpretations of the components are speculative and more work certainly needs to be done, but this is an interesting demonstration that we can begin to investigate the prevalence of scalar implicatures with the use of corpora.

5.3 Exp. 6: corpus studies

I next turn to investigations of the individual and joint effects of different contextual cues on implicature strength. The investigated cues are a) the partitive form,

b) quantifier strength, and c) discourse accessibility as quantified by linguistic mention, topicality, and modification.[4]

5.3.1 Cue 1: the partitive form

Consider the difference between (65) and (66).

(65) Alex ate some of the cashews.

(66) Alex ate some cashews.

There is a clear intuitive difference in how strongly each of these gives rise to the implicature that Alex did not eat all the cashews. In the example with the overt partitive form *of*, intuition strongly suggests that Alex did not eat all the cashews, while in the example without the partitive this intuition is much weaker. There are several reasons why this is the case.

First, it is well-known that there are additional constraints on using the partitive structure that are not at play for non-partitive quantifiers (Abbott, 1996; Barker, 1998; Barwise & Cooper, 1981; De Hoop, 1997; Ionin, Matushansky, & Ruys, 2006; Jackendoff, 1977; Ladusaw, 1982; Reed, 1991). Jackendoff (1977) originally formulated the constraint as one of definiteness of the NP embedded under *some (of)*:

(67) Partitive Constraint I

The complement NP in a partitive must be definite.

Subsequently, this formulation of the constraint was shown to be too strong: there are well-documented cases of indefinite, but specific partitives, as in *one of many people who saw the accident* or *half of a cookie* (Ladusaw, 1982).

[4]The corpus annotations used in Exp. 6 were first compiled by Degen and Jaeger (in prep.).

Reed (1991) re-formulated the constraint as one of discourse accessibility. She proposed that the embedded NP must refer to a discourse accessible group; rather than *evoking* a discourse group, the embedded NP must *refer back to* an already mentioned (or inferable) discourse group. The function of the partitive structure is to evoke a subgroup of that discourse group. Under a discourse accessibility account like Reed's, the strong preference for the embedded NP to be syntactically definite is explained by the embedded NP's discourse function: "the need to access a discourse group creates a preference for, but not a restriction to, definite NPs in the embedded position" (Reed, 1991, p. 216).

Whence, then, the intuition that partitive *some* more strongly gives rise to the implicature that Alex did not eat all of the cashews than non-partitive *some*? Consider what the implicature presupposes: in order to infer that the speaker intended to convey that X is the case of *some, but not all* Y, there must be some group Y, mutually known by both interlocutors, that can be partitioned. Such groups are precisely Reed's discourse accessible groups. That is, the partitive's intuitively high propensity to give rise to scalar implicatures is a consequence of the discourse accessibility constraint on NPs embedded under partitives. It is only with discourse accessible NP referents that scalar implicatures should be able to arise.

Note that this does not prevent utterances with non-partitive *some* from giving rise to scalar implicatures, i.e. using the partitive is not *necessary* to get scalar implicatures from utterances with *some*. As long as the embedded NP is discourse accessible, the scalar inference is possible, whether or not the *some*-NP is overtly partitive. For example, it seems that if (66) were uttered in a context with a contextually given set of cashews, the speaker should more strongly be taken to mean that Alex did not eat all the cashews than if such a set was not given.

The a priori difference between partitive and non-partitive *some* in their probability of resulting in a scalar inference can thus be summarized as follows: scalar

implicatures can only arise with discourse accessible embedded NP referents. The partitive structure can only be used with discourse accessible embedded NP referents, while non-partitive *some* can be used with both accessible and inaccessible referents. Thus, the a priori probability of a scalar implicature is higher for partitive *some* (which *always* occurs with accessible embedded NP referents) than for non-partitive *some* (which only *sometimes* occurs with accessible embedded NP referents).

However, the occurrence of the partitive itself is not *sufficient* for a scalar implicature to arise, either. For example, in Exp. 2 implicature rates were higher for statements with partitive than non-partitive *some*. However, implicature rates were not at 100% for either construction, suggesting that the partitive does not categorically *force* the proper part reading, as has often been noted in the literature (e.g., Horn, 1997).

Thus, the presence of the partitive should be a strong, but nonetheless probabilistic, cue that increases the probability of a scalar inference from *some* to *not all*.

Data analysis

Here and in the following, I report the results of linear mixed-effects regression models (Baayen, 2008) to test the effect of different cues on implicature ratings while simultaneously accounting for conditional dependencies between data points from the same rater. These dependencies are captured in so-called random effects, which offer a convenient way to account for violations of the assumption of independence of each data point (for an introduction directed at language researchers, see Jaeger, 2008). This kind of independence is not granted in datasets in which different participants contribute multiple data points; in our case, different participants may have a systematically different perception of how large the shift in meaning is when the implicature is made explicit. Thus one (forgiving) partici-

pant may have given systematically higher similarity ratings than another (less forgiving) participant. Random effects allow us to account for this individual participant variability and thus more clearly interpret the effects of the cues under investigation.

All statistical analyses were mixed-effects linear regressions predicting implicature rating from fixed effects of interest (the cues under investigation) and random by-participant intercepts. Reported p-values were obtained by MCMC sampling using the `pvals.fnc()` function in R (Baayen, 2008). The partitive and the quantifier predictor were allowed to interact, as were the three discourse accessibility predictors. I report the main effect of each cue individually. The interaction between partitive and quantifier strength is discussed in Section 5.3.2. The interaction between the different discourse accessibility predictors is discussed in Section 5.3.3. The full model is summarized in Appendix C.

Results

Of the entire dataset, 26.5% of cases were partitives, of which in turn 26.8% were headed by pronouns or demonstratives as in (68) and (69).

(68) Uh, **some of that** unfortunately is legal.

(69) And for **some of them** it was just kind of, I don't know, not so much a holiday.

As can be seen in Figure 5.3, the overtly partitive cases received higher implicature ratings than the non-partitive cases ($\beta = 1.01, SE = 0.05, t = 22.05, p < .0001$). Compared to the global mean rating of 3.9, the partitive mean was higher at 5, while the non-partitive mean was lower at 3.5. Similarly, the median rating for partitive cases was 5, while the non-partitive median was 3.

Compared to the 44.7% of cases that globally received ratings above the midpoint of the scale, conditioning on overt partitivity increases that number to 67.8%.

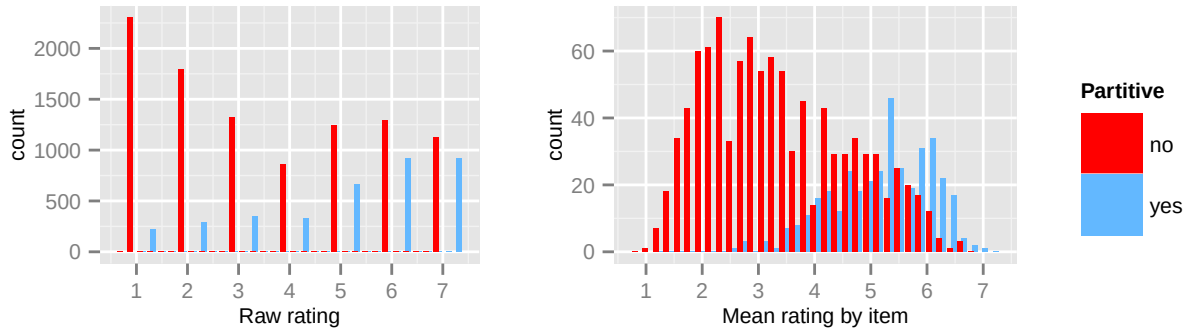

Figure 5.3: Distribution of raw ratings (left panel) and mean per-item ratings (right panel) for overtly partitive (green) and non-partitive (red) cases.

This suggests two things: a) support for the implicature is much higher when the *some*-NP is overtly partitive and b) the partitive is nevertheless not sufficient for getting a scalar inference: only 25% of ratings were 7s, and 23% of ratings were still below the midpoint of the scale. Examples of partitive cases that received low similarity ratings are shown in (70) - (72) alongside their mean similarity rating.

(70) I wish my mother had had **some of those opportunities**, because, I think she would have really, she rea-, would have succeeded in a lot of ways, that men, that women were not able to succeed in her generation. 2.4

(71) But when you get into **some of these health clubs** where you just stand around and wait. . . 2.9

(72) I just go to be entertained and am not really interested in some of the, like, the Terminator or **some of the Schwarzenegger stuff**. 2.9

In all three cases, the interpretation of *some* seems to be lower-bound despite the presence of the partitive.

5.3.2 Cue 2: quantifier strength

The word *some* is ambiguous between a weak, indefinite, or non-presuppositional reading, often written as *sm* because it tends to be unstressed, and a strong, quantificational, or presuppositional, reading (Milsark, 1974, 1977; Barwise & Cooper, 1981; Horn, 1997; Ladusaw, 1994; Israel, 1999). Consider the example in (73).

(73) Some prospectors got the plague.

The sentence in (73) can mean either that there is an indefinite number of prospectors who got the plague (weak, sometimes also called cardinal interpretation) or that some prospectors got the plague but others presumably did not (strong, sometimes also called partitive or proportional interpretation). In general, determiners can either be unambiguously weak (e.g., *a/an* and *no*) or strong (e.g., *all* and *most*), or ambiguous between the two readings (e.g., *some*).

The distinction between weak and strong determiners is central to the distribution of scalar implicatures from *some* to *not all* because it has been noted that the use of strong, but not weak determiners, gives rise to scalar implicatures (Ladusaw, 1994). Indeed, the partitive form (which, as noted in the previous section, is associated with higher implicature rates than non-partitives) tends to only occur with strong determiners (e.g., Ladusaw, 1994; Horn, 1997).

However, the weak/strong distinction has been notoriously difficult to pin down (e.g., Horn, 1997). The goal here is not to give an exhaustive review of the rich literature on weak and strong determiners, but rather to identify an operationalization of the weak/strong distinction that will facilitate a quantitative test of whether strong *some* is more likely to give rise to scalar inferences than weak *some*. To foreshadow, the presuppositionality difference between weak and strong *some*-NPs (e.g., Lumsden, 1988) will be employed to arrive at empirical ratings

Table 5.3: Diagnostics for identifying strong vs. weak uses of *some* (based on Horn, 1997)

	Strong *some*	Weak *some*
a)	presuppositional	non-presuppositional
b)	partitive or proportional	cardinal
c)	likely to give rise to scalar implicature	unlikely to give rise to scalar implicature

of the strength of each use of *some* in the database. I begin by elaborating on some of the properties that have been observed to correlate with the distinction.

Table 5.3 summarizes the diagnostic tests relevant to our purposes, provided in a review by Horn (1997). The property that we crucially depend on in collecting strength ratings from participants is one made by e.g. de Jong and Verkuyl (1985) and Lumsden (1988). They propose that there is a presupposition on strong determiners that their restriction not be empty and their domain of quantification be part of the domain of discourse. That is, under the strong interpretation of (73), there needs to be some set of prospectors in the domain of discourse of whom it is being predicated that they got the plague. Under the weak reading, the domain of discourse need not contain a set of prospectors - the set is introduced (the discourse group evoked, in Reed, 1991's terms) by the *some*-NP.

The weak/strong distinction correlates with other properties which are not directly relevant to our purposes, e.g., the propensity to occur in existential *there* constructions (Milsark, 1974; McNally & Geenhoven, 1998) and the ability to occur with individual-level predicates (Carlson, 1977; Milsark, 1977). Importantly, the literature provides counterexamples to each of these diagnostics (see e.g., Horn, 1997; McNally & Geenhoven, 1998). Rather than being strict constraints or part of the definition of strong determiners, it seems that these properties are approximate diagnostics and I will treat them as such.

To arrive at an estimate of the strength of *some* for each of the cases in the database, the presuppositionality difference was exploited in a web-based study

collecting participants' judgments about the use of *some*.

Collecting quantifier strength ratings

To quantify determiner strength, participants rated the similarity of each original utterance from the dataset to the same utterance without *some (of)* on a seven-point Likert scale. The reasoning behind this choice was built on the presuppositional nature of strong NPs: the weak use of *some* does not have a non-empty restriction presupposition associated with it, while the strong one does. Thus, in removing *some*, the change in meaning should be greater for strong than for weak *some*-NPs. Consider examples (74) and (75).

(74) *Weak use*

a. But my son needed sm money.

b. But my son needed money.

(75) *Strong use, partitive*

a. And some of the people in our church use birth control.

b. And the people in our church use birth control.

(76) *Strong use, non-partitive*

a. Some history books are pretty scary.

b. History books are pretty scary.

Mutual entailment holds between the a and b sentences in (74) but not in (75) and (76), i.e., there is intuitively a greater difference in meaning between the a and b forms in (75) and (76) than in (74). Thus, the higher the similarity rating given for a particular case, the weaker the use of *some* in this case. Conversely, the lower the rating the stronger the use.

Evaluation of quantifier strength ratings

An independent test of the goodness of the collected quantifier strength ratings was performed by investigating the correlations between quantifier strength ratings and three properties that have been proposed as diagnostic tests for the weak/strong distinction: overt partitivity (discussed in detail above), the propensity to occur in existential *there* constructions (Milsark, 1974; McNally & Geenhoven, 1998) and the ability to occur with individual-level predicates (Carlson, 1977; Milsark, 1977).

I briefly touch on the latter two properties. The inability of strong NPs to occur in existential *there* constructions while weak ones can, is an observation originally made by Milsark (1974). Consider the following examples:

(77) There are sm prospectors in camp.

(78) * There are some of the prospectors in camp.

Milsark proposed the presuppositionality property of strong NPs as an explanation for this asymmetry: in existential constructions, the existence of a non-empty discourse group is asserted. Thus, to be pragmatically acceptable the group cannot have been previously evoked. If strong NPs contain discourse-accessible embedded NPs, this would make the NP pragmatically infelicitous in the existential *there* construction. Thus, for example, no partitives should occur in existential constructions since they are only licensed with discourse-accessible embedded NPs. This is supported by the contrast between (77) and (78).

In practice, however, language users do not seem to categorically adhere to this constraint. Our database of sentences with *some*-NPs from the Switchboard corpus contains two cases of partitive *some* occurring in existential constructions:

(79) . . . because there is *some of those English words* that you just don't exactly know what they mean. . .

(80) ... but there is *some of the day-to-day needs* that they just are not able to deal with physically anymore...

And indeed, many further counterexamples have been noted in the literature (see, e.g., McNally & Geenhoven, 1998, and references therein). McNally and Geenhoven (1998) note that the fact that both weak and strong determiners can occur in existential sentences if certain conditions on the descriptive content of the NP are met suggests that an alternative demarcation of the weak/strong distinction is required (though the general correlation between existential constructions and weak determiners is taken to be real).

The second difference in distribution of weak and strong NPs (in the sense of Milsark, 1974) is exemplified in (81) and (82): the type of predicates that weak and strong NPs can occur with differ (Milsark, 1977). Strong NPs can occur with both *individual level* and *stage level* predicates (Carlson, 1977). Weak NPs, on the other hand, cannot occur with individual level predicates. Individual level predicates are those that "name some trait possessed by the entity and which is assumed to be more or less permanent, or at least to be such that some significant change in the character of the entity will result if the description is altered" (Milsark, 1977, p. 13, who refers to these as *property denoting*). Examples are *tall, intelligent, a lawyer*. Stage level predicates are those "which are subject to change without there being any essential alteration of the entity" (Milsark, 1977, p. 12, who refers to these as *state denoting*). Examples are *sick, drunk, open*.

(81) a. Sm men are sick. (weak & stage level)

b. * Sm men are tall. (weak & individual level)

(82) a. Some men are sick. (strong & stage level)

b. Some men are tall. (strong & individual level)

Predictions Mean similarity ratings for sentence pairs containing weak *some*-NPs should be higher than for strong *some*-NPs. Given the link between determiner strength and overt partitivity, propensity to occur in existential *there* constructions, and ability to occur with individual and stage level predicates, this makes the following three specific predictions:

1. Mean similarity ratings for utterances with partitive *some* should be on average lower than for simple *some* (because partitive NPs should be strong NPs, which should have low similarity ratings).

2. Mean similarity ratings for utterances where the *some*-NP occurs in an existential construction should be on average higher than for other utterances (because NPs occurring in existential constructions should be weak NPs, which should have high similarity ratings).

3. Mean similarity ratings for utterances where the *some*-NP occurs with individual level predicates should be on average lower than for utterances with stage level predicates (because individual level predicates should only occur with strong NPs, which should have low similarity ratings).

Methods

Participants. Participants were recruited over Amazon's Mechanical Turk. Each participant was paid $0.05 for a block of 10 items. Anyone who completed more at least four blocks received a bonus of $0.05, anyone who completed at least than 8 blocks received an additional bonus of $0.10. In total, 137 Mechanical Turk workers participated in this experiment.

Materials and procedure. Sentences with *some* were rated for similarity to the same sentence not containing *some (of)* on a 7-point scale, where 1 meant "very different meanings" and 7 was "exactly the same meaning". Because quantifier ratings were originally collected as part of a separate project investigating the

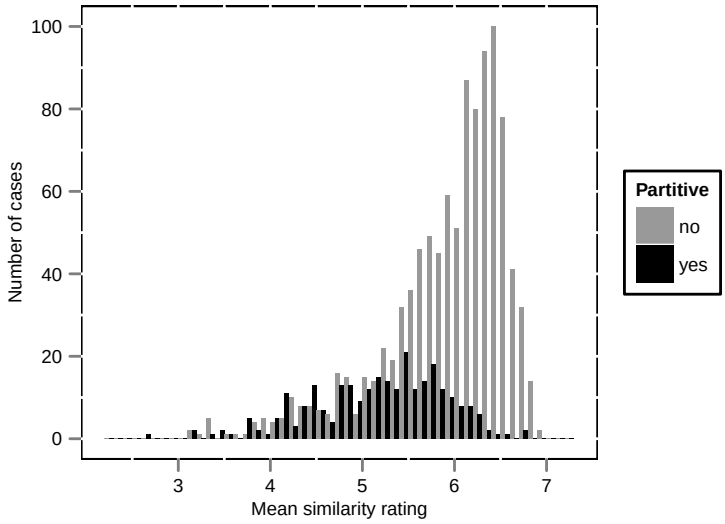

Figure 5.4: Distribution of simple and partitive *some* cases over mean quantifier strength ratings.

choice of partitive *some of* over simple *some* and pronoun heads cannot occur in the simple form, pronoun cases were not included in this study. Thus, only the 1290 cases that were not pronoun cases received quantifier strength ratings. Sentences were grouped into blocks of 10 items that reflected the distribution of partitives (22%) and non-partitives (78%) in the total dataset. Order of partitive cases was randomized in each block. Each block occurred in two orders (one was the reverse of the other). That is, there were a total of 129 blocks of 10 items that were rated 10 times each.

Results

Partitives. As predicted, the global mean of mean similarity ratings for partitive *some*-NPs was lower than for simple *some*-NPs (5.11 vs. 5.88, see Figure 5.4). Median values were 5.2 and 6.1, respectively.

Existentials. Of the 1290 cases in the database, 97 were cases of *some*-NPs in existential constructions. One case, shown in (83), had a particularly low

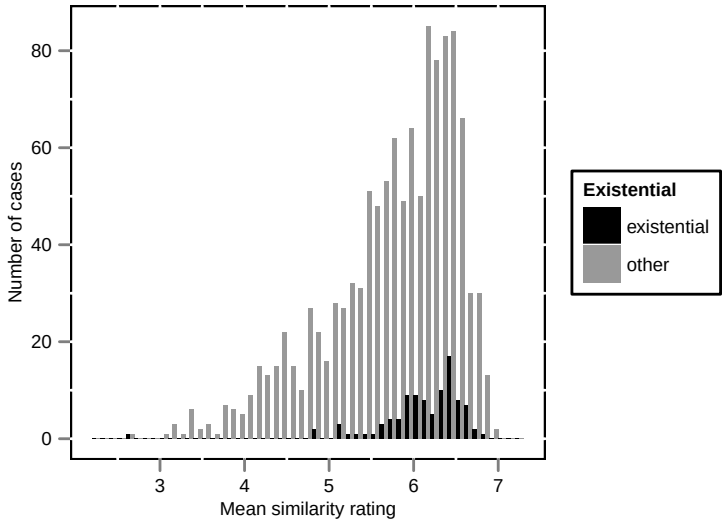

Figure 5.5: Distribution of existential and other *some*-NPs over mean similarity (quantifier strength) ratings.

similarity rating of 2.6. However, this rating seems to be driven not by the strength of *some*, but rather by the a priori ungrammaticality of the utterance (which is most likely a transcription error). This case was thus excluded from the analysis.[5]

(83) I mean, there are sort some inherent limits there.

As predicted, mean similarity ratings for utterance pairs were higher for *some*-NPs in existential constructions compared to the rest of the utterances (mean of 6.1 vs. 5.7, median of 6.2 vs. 5.9, see Figure 5.5).

Both the partitive and the existential results were confirmed by a linear mixed effects regression predicting mean similarity rating from centered partitive and existentiality predictors. Partitive *some* led to lower mean similarity ratings than simple *some* ($\beta = -0.65$, $SE = 0.05$, $t = -14.23$, $p < .001$) and utterance pairs with existential constructions were rated as more similar than other utterances ($\beta = 0.20$, $SE = 0.07$, $t = 2.95$, $p < .01$). This remained true in a model controlling for log-transformed sentence length. Longer sentences tended to lead

[5]Not excluding this outlier did not change the results qualitatively.

to higher similarity ratings, presumably because more of the actual linguistic form is identical than in shorter sentences and thus perceived meaning differences are smaller ($\beta = 0.3$, $SE = 0.03$, $t = 9.49$, $p < .001$).

Predicate types. To test whether there was a difference in mean similarity ratings for individual vs. stage level predicates, we extracted a random sample of 100 cases of plural simple *some*-NPs from the database and manually annotated them for whether they occurred with individual level predicates as in (84) or stage level predicates as in (85).

(84) a. I think some parents go a little bit overboard.

b. And some kids do need to be spanked at home.

(85) a. I have some spiro gyro tapes.

b. And he said, um, let's give her some race horse shots.

There were 14 cases where there was not enough context to clearly determine the predicate type. Of the remaining 86, 14 occurred with individual level predicates and 72 with stage level predicates. On the annotated subset of the data, we performed a linear mixed effects regression predicting mean similarity from a centered predictor coding whether a *some*-NP occurred with a stage level predicate. As predicted, similarity ratings were higher for stage level predicates ($\beta = 1.25$, $SE = 0.17$, $t = 7.39$, see Figure 5.6).

Discussion All three of the tested diagnostics provide support for the decision to quantify the weak/strong distinction in mean similarity ratings via the *some*-omission test: lower mean similarity ratings for partitive than simple *some*, higher similarity for existential constructions, and lower similarity for individual level predicates are all predicted under the assumption that low similarity ratings pick out strong *some*-NPs.

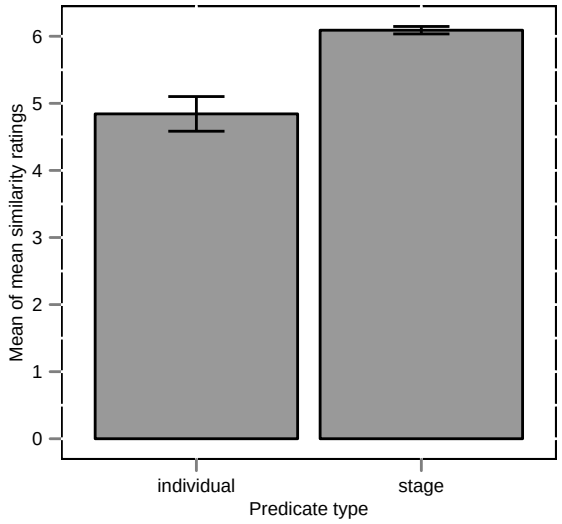

Figure 5.6: Mean of mean similarity (quantifier strength) ratings for *some*-NPs occurring with individual vs. stage level predicates.

At the same time, however, if the distinction between weak and strong *some* is categorical, we should see two categorically distinct distributions in Figure 5.4: one for partitives and other strong NPs and one for weak NPs. This is not the case. While the partitive distribution is shifted to the left of the rest of the cases, there is no clear boundary. There are at least three possible explanations for this: the first is that while the distinction is categorical and the *some*-omission test appropriately captures the distinction, our Mechanical Turk workers' intuitions were either unclear or some workers simply did not pay enough attention. While it is possible that collecting data over the web might lead to noisier data than in the lab, we collected ten ratings per item to minimize the effects of noise.

The second possibility is that, while the weak/strong distinction is categorical, the *some*-omission test is not optimal for capturing it. While this is certainly a possibility, given the theoretical difficulty in pinning down exactly the difference between weak and strong uses I believe the final possibility is more likely: that the weak/strong distinction itself is not categorical. That is, rather than thinking

of an NP as either strong or weak, what a speaker chooses to linguistically encode may lead to more or less perceived NP strength on the comprehender's side. For example, an NP that is partitive and occurs with an individual level predicate may be perceived as stronger than one that is not partitive but occurs with an individual level predicate, which may be perceived as stronger than a partitive occurring with a stage level predicate, which in turn may be stronger than a partitive in an existential construction (strength conflict), which will probably be perceived as stronger than a non-partitive NP in an existential construction or a non-partitive occurring with a stage-level predicate. Admittedly cherry-picked examples that seem to confirm this pattern are provided below, in increasing order of mean similarity rating (i.e., in decreasing order of strength).

(86) But, um, some of these people are just out and out brutal. (3.9)

(87) Well, um, some history books are pretty scary. (4.3)

(88) I think it would be nice if some of the parents would be around for a few more hours. (5.2)

(89) There's some of those English words that you just don't exactly know what they mean (5.9)

(90) Oh, there's, there's some nice things in Baltimore, you know. (6.5)

(91) But my son needed some money. (6.9)

Another reason for taking a gradient view of the weak/strong distinction is the absence of a proposal in the literature that fully captures the distributional facts. As pointed out by some (e.g., Horn, 1997; McNally & Geenhoven, 1998), these properties should be interpreted as correlating with the distinction rather than determining it.

The quantifier strength ratings collected here thus correlate strongly with the diagnostics proposed in the literature, but there does not seem to be a clear-cut

weak/strong distinction. The quantifier strength ratings are used in the following to assess the effect of quantifier strength on perceived implicature strength.

Results of quantifier strength effect on implicature ratings

Because we did not collect strength ratings for the cases where the head of the embedded NP was a deictic expression like a pronoun or a demonstrative, strength ratings for these cases were not available. In order to include these 97 cases, strength ratings were generated in a principled way. To understand how, note first that the partitive is mandatory for pronouns and demonstrative heads - see examples (92) - (94) together with their mean implicature rating.

(92) And some *(of) them fizzled out. 6.6

(93) Some *(of) it sounds more like pop music. 5.9

(94) But some *(of) those are pretty big. 5.6

It is thus not implausible that theses cases would receive strength ratings similar to those of the other partitive cases. Based on this assumption, ratings were generated by sampling from the strength rating distribution of the 269 partitives. That is, the resulting strength distribution for pronoun/demonstrative cases was approximately the same as that of the other partitive cases. These strength ratings were used for the rest of the analysis.[6]

Implicature ratings were higher with increasing quantifier strength (β = -0.54, SE = 0.03, t = -21.10, $p < .0001$). This is shown in Figure 5.7. The stronger the use of *some*, the stronger the support for a scalar inference. Conversely, the weaker the use, the less likely the implicature. This is compatible with the general observation in the literature that strong uses of *some* can give rise to the implicature, but it is important to note that this is not a perfect correlation

[6]Excluding the deictic head cases does not change any of the qualitative results or significance of effects.

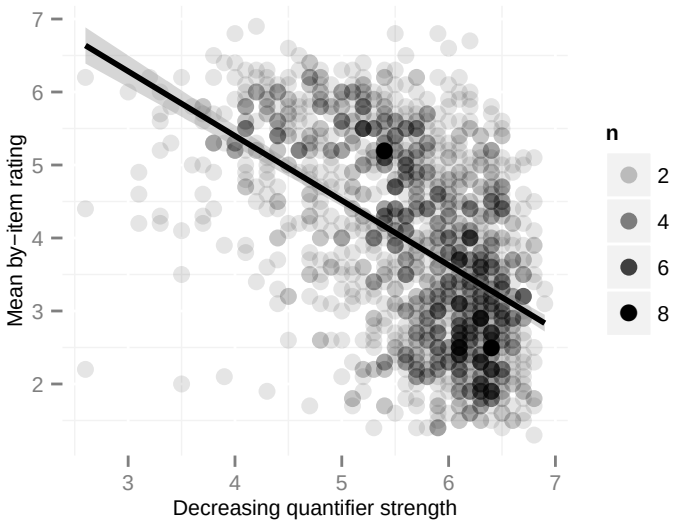

Figure 5.7: Mean by-item implicature rating as a function of decreasing quantifier strength. Opacity of each point indicates the contributing number of data points (i.e. darker dots = more data).

(Pearson's $r = -0.51$). That is, some uses of the quantifier were judged as strong but did not strongly support the scalar inference, whereas others were judged to be weak but nevertheless the implicature ratings suggest that support for a scalar inference was strong. We look at some examples of each of these cases.

(95) Strong quantifier, low implicature rating (first number: mean quantifier strength rating; second number: mean implicature rating)

a. I'd like to go to Sundance and Park City and **some of those**. 2.6; 3.6

b. What are **some of the things they don't recycle** . 4.1; 3.8

c. Maybe this would be a way to get that feeling back, if we've lost **some of that**. 4.1; 3.9

Cases of strong quantifiers that nevertheless give rise to scalar implicatures only weakly, if at all, are not in principle surprising: standard lower-bound interpretations, where the implicature does not arise because the stronger alternative is not contextually relevant, should give rise to just this pattern. The example in

(95a) seems to be of this type. In contrast, the weak implicature support in (95b) and (95c) seems to have a different source: in (95b), the *some*-NP is embedded in a wh-question, while in (95c) it is in the antecedent of a conditional. Both of these are instances of downward-entailing (DE) environments, which have been known to cancel and even flip implicatures (Levinson, 2000; Chierchia, 2004; Chierchia et al., 2008).

It will be important to annotate the entire dataset for whether the *some*-NP occurs in a DE environment or not. So far, annotation of a random sample of 50 cases only yielded two cases where the *some*-NP occurred in a polar interrogative. While polar interrogatives are not typically treated as DE environments, they share with DE environments that they license Negative Polarity Items (NPIs). Thus, if this is a good estimate of DE or NPI-licensing contexts, roughly 4% of *some*-NPs occur in DE environments, for which implicature ratings should be low. The following two are the polar interrogative cases with their mean implicature rating.

(96) Or do **some of them** play the same song? 4.7

(97) But is it a legal, uh, solution for **some companies**? 5.4

Both of these mean ratings are higher than the global mean in the dataset, suggesting that at least in polar interrogatives, the implicature is not categorically ruled out. However, a complete test of the effect of DE context on ratings in our dataset remains to be conducted.

I turn next to examples of cases where quantifier use was judged as weak but implicature ratings were nevertheless high.

(98) Weak quantifier, high implicature rating (first number: mean quantifier strength rating; second number: mean implicature rating)

a. It's hurting, you know, it's hurting Germany , for example, too, and **some other parts of Europe where they, where they have high industry.** 6.4 5.7

b. And, after I, I graduated, I read **some of the old classics that I just bluffed my way through** and have found that I enjoy them quite a bit, too. 6.2 6

c. But I think that at **some times** it can be the right thing to do. 6.2 6.7

d. And then on the other hand, I've seen **some people** go into the nursing home and just so happy you know. 5.8 5.7

There seem to be two different things going on here. In the a and b cases, use of the quantifier is weak in that it is introducing two new discourse groups - *other parts of Europe* and *old classics*. However, the modifying post-nominal material introduces a contrast with a (presumably non-empty) complement set - *parts of Europe where they don't have high industry* and *the old classics that I did not bluff my way through*. In these cases, then, the upper-bound interpretation may not arise as a standard implicature, but as a consequence of the non-empty complement set presupposition introduced by the post-nominal modification.

Similarly, in the c) and d) cases the upper-bound interpretation does not seem to be due to the standard Quantity reasoning, but instead to the prior probability of the state of the world signaled by the upper-bound interpretation being high: world knowledge tells us that it is more likely that it is not at all times the right thing to do rather than that it is (whatever *it* may refer to in this case). And it is more likely that not all people go into the nursing home and are happy rather than that they all are.

Thus, while implicature support is strongly correlated with quantifier strength, factors like monotonicity properties of the context that the *some*-NP is embedded

in, discourse expectations, and world knowledge affect scalar implicatures.

5.3.3 Cue 3: discourse accessibility

As discussed above, Reed (1991) proposed a discourse accessibility constraint on the partitive - the partitive can only be used with embedded NPs referring to discourse accessible referents. Relatedly, strong uses of *some* have been argued to be both covertly partitive and have a discourse accessibility presupposition on the embedded NP. In this section I investigate the effect of discourse accessibility on scalar implicatures above and beyond overt partitivity and quantifier strength.

Several factors contribute to discourse accessibility: here I investigate a) linguistic mention of the embedded NP referent, b) topicality of the *some*-NP, and c) modification of the embedded NP.

Several researchers have noted that scalar implicatures seem to be affected by information structure. For example, Breheny et al. (2006) found that more scalar implicatures are generated in Greek for sentences in which the *some*-NP is in subject position compared to when it is in object position. Their explanation is that scalar implicatures should only arise when the scalar trigger is in focus, i.e. addressing a contextually relevant issue, which in turn makes the stronger alternative with *all* salient in context. Because of the strong tendency of Greek (and weaker tendency of English) for subjects to be more likely to contain old rather than new information and thus more likely to be addressing a contextually relevant issue, scalar implicatures are more likely to arise for *some*-NPs in subject position than in positions that are lower on the obliqueness hierarchy (e.g. objects, adjuncts, etc., Breheny et al., 2006).

Similarly, Zondervan (2008) showed that the implicature triggers *or* and *most* are more likely to give rise to scalar implicatures if they are in the focus of an implicit or explicit Question Under Discussion (QUD, Roberts, 1996).

Taken together, this predicts effects of both linguistic mention and topicality on scalar implicatures: implicature ratings should be higher with both previously mentioned and topic *some*-NPs. Additionally, adding pre- or post nominal modification to an NP that refers to a new (previously unmentioned) entity or group makes this group accessible (Prince, 1981; Reed, 1991). Consider the following example:

(99) When we arrived at the hotel we didn't know where to go so we asked *the guy at the front desk*.

The restrictive modifier *at the front desk* makes the novel mention of *the guy* discourse-accesssible by providing *uniquely identifying* information (Reed, 1991; Webber, 1983).

It is thus to be expected that these different markers of discourse accessibility interact: for example, modification should increase implicature ratings more for relatively discourse-inaccessible (new or non-subject) cases than for relatively discourse accessible (old or subject) cases. To reflect this, predictors for linguistic mention, subjecthood, and modification were allowed to interact in the regression model. I first discuss the main effects of each of the three factors before turning to the interactions.

Linguistic mention

Nouns in the Switchboard are annotated for whether they are *old* (previous mention), *new* (novel mention), or *mediated* (not previously mentioned but contextually inferable). In the *some*-database, there were 142 old, 767 mediated, and 454 new cases. Figure 5.8 shows the distribution of ratings over different mention categories. In the mean ratings histogram (right panel), the distribution of new NPs is skewed towards the lower end of the rating scale while the distribution

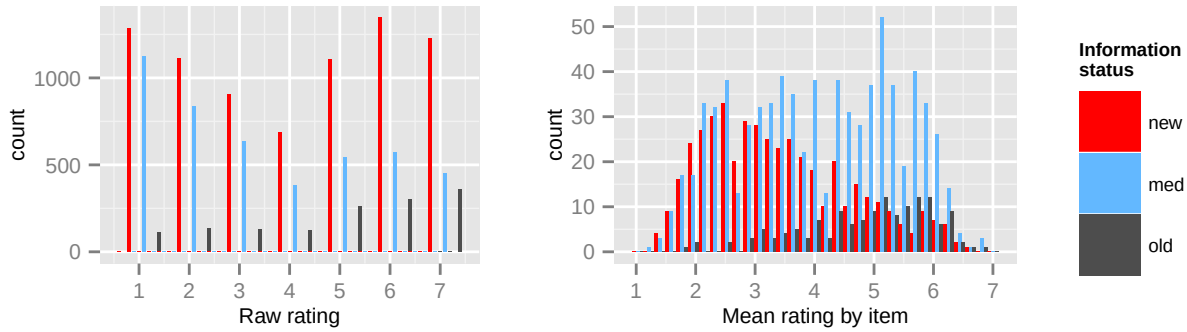

Figure 5.8: Distribution of raw ratings (left panel) and mean per-item ratings (right panel) for items with old, mediated, and new embedded NP referents.

of old NPs is skewed towards the upper end. The situation for mediated NPs is visually less clear.

For the purpose of ease of analysis, embedded NPs with old and mediated head nouns were collapsed into one category.[7] As predicted, implicature ratings were higher for old than new NPs ($\beta = 0.32, SE = 0.04, t = 8.62, p < .0001$).

One surprising finding is that there were many new NPs that nevertheless received high implicature ratings. I discuss this further below.

Topicality

In the Switchboard corpus, NPs are annotated for whether they are sentential subjects as in (100) or in topicalized constructions like left-dislocations as in (101).

(100) *Some people* are motorboaters, you know, which I think is fine. 5.5

(101) *Some of those people*, they don't deserve to be let loose. 4.8

(102) I've heard *some horrible, horrible stories* about high school teachers. 3.1

(103) We actually do some work with *some people down at Georgia Tech.* 4.5

[7] Old and inferable information tends to pattern together in discourse (Birner, 1997).

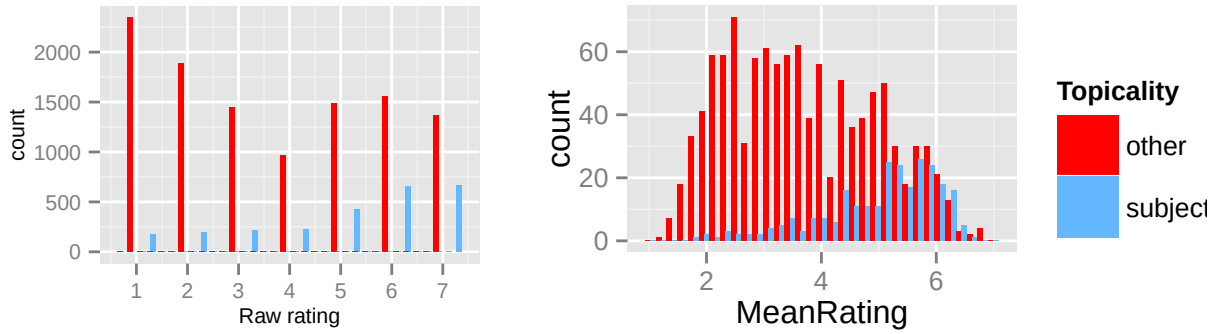

Figure 5.9: Distribution of raw ratings (left panel) and mean per-item ratings (right panel) for subject and other *some*-NPs.

Because there were only 19 cases of topicalized NPs, these were collapsed into the subject NP category. There were thus 257 subject and 1106 other NPs in the *some*-database. Other NPs were, for example, cases of direct objects as in (102) or prepositional adjuncts as in (103). Figure 5.9 shows the distribution of ratings over these two categories. The distribution of subject NPs is clearly skewed towards the upper end of the scale while the distribution of other NPs is skewed towards the lower end. This difference is significant: subject NPs are associated with higher implicature ratings ($\beta = 0.40, SE = 0.06, t = 7.25, p < .0001$).

Modification

Finally, each case in the database was coded as either *modified* or *unmodified*, depending on whether the embedded NP had pre- or post-nominal modification or not. For example, the examples in (104) and (105) both fell into the modified category, while the case in (106) was classified as unmodified, resulting in 667 modified and 696 unmodified cases. In addition, partitive cases with possessive embedded determiners were categorized as modified because in those cases, the determiner provided additional information on the relation between the head noun of the embedded NP and already discourse accessible entities, as in (107) where the possessive provides a link between relatives (new) and the speaker's family

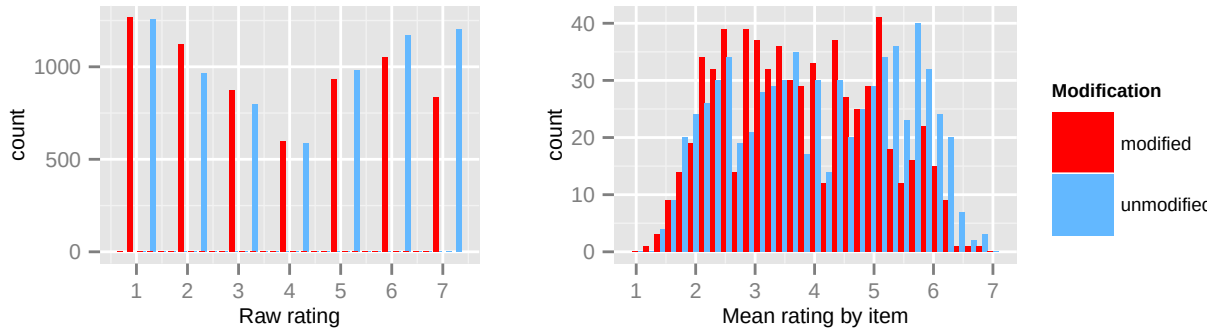

Figure 5.10: Distribution of raw ratings (left panel) and mean per-item ratings (right panel) for modified and unmodified *some*-NPs.

(old). There were 12 of these cases in the database overall.

(104) And then I've seen *some of the Star Trek movies*. 6.5

(105) So, we're a little farther removed from like Dallas and *some of the areas where they probably have more of the homeless and that type of thing*. 5.2

(106) We had *some friends* over as recently as Saturday night. 3.4

(107) Christmas time, uh, *some of our relatives* would come up from Alabama. 6.3

Figure 5.10 shows the distribution of ratings for the two modification categories. Unmodified NPs received higher ratings than modified NPs ($\beta = 0.11, SE = 0.04, t = 3.14, p < .001$). This is due to the interaction with other variables, which I discuss in the next section.

Interactions between discourse accessibility factors

The model coefficients for the two-way and three-way interactions between discourse accessibility predictors are shown in Table 5.4. All interactions were significant, though the one between modification and mention only marginally so. The three-way interaction is visualized in Figure 5.11. Simple slopes analysis revealed

Table 5.4: Model coefficients for interactions of discourse accessibility predictors.

Predictor	β	SE	t	p
Topicality:Mention	0.18	0.12	1.43	< 0.08
Modification:Mention	0.33	0.073	4.54	$< .0001$
Modification:Topicality	0.27	0.11	2.53	$< .01$
Modification:Topicality:Mention	0.53	0.25	2.14	$< .05$

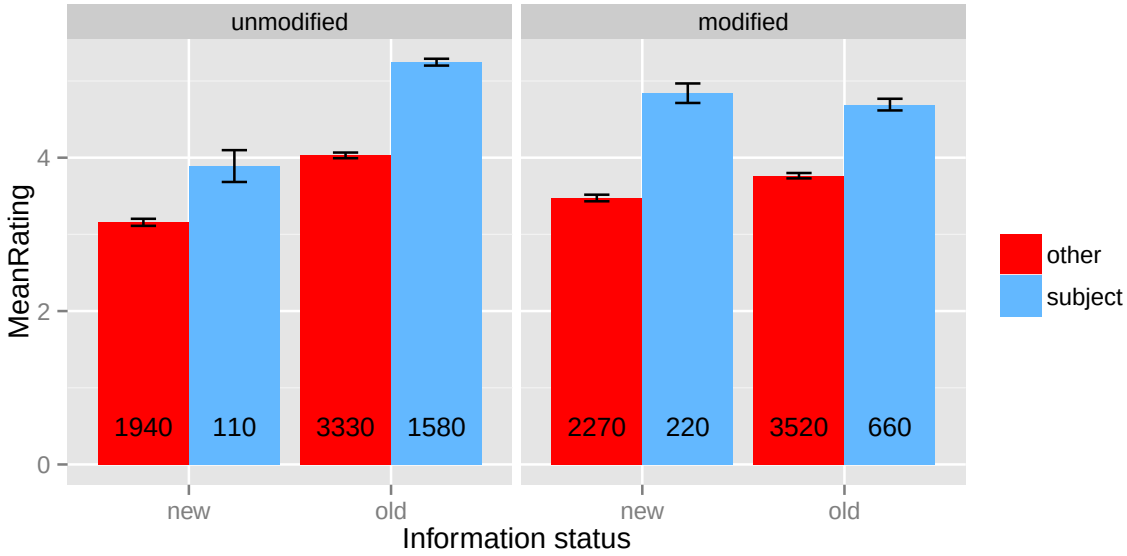

Figure 5.11: Mean similarity ratings by linguistic mention (old/new embedded NP referent), topicality (subject/other *some*-NP), and modification (modified/unmodified embedded NP). Numbers in bars indicate the number of contributing observations.

that the two-way interaction between linguistic mention and topicality was significant for unmodified ($\beta = 0.11$, $SE = 0.05$, $t = 2.34$, $p < .05$), but not for modified NPs ($\beta =$ -0.02, $SE = 0.04$, $t =$ -0.49, $p < .7$); for modified NPs, there was only a main effect of topicality, such that modified NPs in subject position received higher implicature ratings than modified NPs in other positions. For unmodified NPs, there was an interaction such that both old and subject NPs received higher ratings, but the difference between subject and other NPs was greater for old than for new NPs.

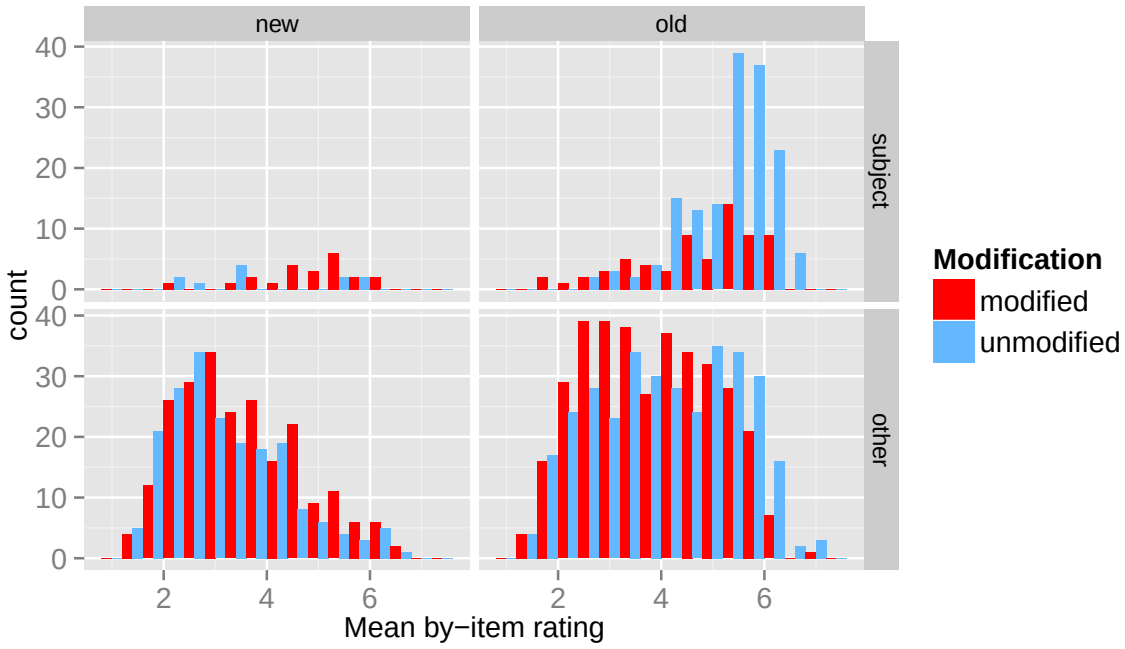

Figure 5.12: Distribution of modified and unmodified cases over mean by-item ratings by linguistic mention (left vs. right column) and topicality (top vs. bottom row).

These interactions provide an explanation for why modification, as we saw in Section 5.3.3, *lowers* rather than increases implicature ratings on average: when an entity is already discourse accessible, either because it was previously mentioned or because its referring NP is in subject position (or both), modification is not necessary to establish discourse accessibility. In contrast, when discourse accessibility is otherwise low, modification can be used to make an entity discourse accessible, which in turn should increase implicature ratings. That this pattern holds can be observed most clearly in Figure 5.12. Where discourse accessibility is high (old and subject NP, upper right corner), there are many more *unmodified* than modified cases on the higher end of the scale; where discourse accessibility is low (new and other NP, lower left corner), there are more *modified* than unmodified cases on the higher end of the scale. The other two cells are intermediate between the two, though there are very few cases of new subjects. A χ^2 test over the linguistic mention x topicality contingency table replicates this well-documented tendency for grammatical subjects to favor old over new information ($\chi^2(1) = 58.73, p < .0001$).

5.3.4 Discussion of results

Let us take stock. The corpus analyses have revealed that the Frequency Assumption does not seem to hold for scalar inferences from *some* to *not all*. However, the strength of an implicature, or the degree to which a speaker is taken to implicate the negation of the stronger alternative, increases on average when *some* is used in the partitive, when its use is relatively strong (as established by an independent test), and when the embedded NP referent is relatively discourse accessible (i.e. when it has been previously mentioned or is contextually inferable, the *some*-NP is in subject position, or the embedded head noun is modified).

The overall lack of support for the upper-bound interpretation of *some* is a surprising result given the previous literature. Follow-up work testing both

different scales and different corpora is needed to establish the robustness and generalizability of the result.

Further work should also estimate the degree of uncertainty that listeners believe speakers to have about the stronger alternative. As discussed in Section 5.2.2, this knowledge will be useful in predicting ignorance implicatures.

Finally, in Section 5.3.2 I briefly touched on additional factors that are likely to influence implicature ratings, e.g. monotonicity properties of the context the *some*-NP occurs in, prior world knowledge about how likely different states of the world are, and expectations of relevance of the stronger alternative to a contextual Question Under Discussion. Some of the variance in the *some*-dataset is no doubt due to these kinds of factors interacting with the cues that I have quantified explicitly. Trying to understand the exact nature of these interactions in this dataset lies outside the scope of this Chapter but is an important goal of future work.

In the final section of this Chapter, I discuss the theoretical implications of these results with a view towards the accounts of scalar implicature presented in Chapter 2.

5.4 Discussion of Exps. 5 and 6

In this section I discuss how the results of Exps. 5 and 6 impact Levinson's Default theory and address the role that the Frequency Assumption plays in the interpretation of experimental results showing that scalar implicatures incur a cognitive cost, which has been viewed as support for the Literal-First hypothesis. Finally, I argue that the obtained results are most consistent with probabilistic, contextualist accounts of scalar implicature.

5.4.1 Levinson's Default theory

Levinson (2000) Default theory of scalar implicatures is built around the distinction between Generalized and Particularized Conversational Implicatures (GCIs vs. PCIs).[8] For Levinson, the distinction provides a solution to the articulatory bottleneck problem: by making the highly frequent, regular, and almost context-independent GCIs (including scalar implicatures) cognitively cost-free, the speed of communication is increased.

Thus, the first piece of establishing the role of the Frequency Assumption for Default theory is this: GCIs (including scalar implicatures) are assumed to be more frequent than PCIs. However, note that this does not yet entail the Frequency Assumption; even if GCIs arise more frequently than PCIs, this does not yet entail that within GCI triggers (i.e., lexical items that trigger a GCI, such as an item from a scale) the GCI is more frequent than not. To arrive there, it is useful to look at the processing component of Levinson's theory. The crucial step involves reinterpreting the GCI-PCI distinction in processing terms.

If the goal is to balance out the processing cost associated with PCIs by making GCIs cost-free and their cancellation costly,[9] the number of cancelled GCIs should not be larger than the number of non-canceled GCIs, since this would result in a net processing *cost* instead of a *gain*. Applied to scalar implicatures, which are treated as the prime example of GCI: the number of cases in which a scalar implicature does not arise should not be larger than the number of cases where it does. But this is exactly the Frequency Assumption!

The Frequency Assumption thus constitutes one of the central assumptions of Levinson's theory. Overturning it would remove the central rationale of Levinsonian Default theory: if the Frequency Assumption does not hold, Levinson's

[8] See Section 1.2 for a discussion of the GCI-PCI distinction and Section 2.1.1 for an overview of Levinson's account.

[9] See discussion on p. 17 for how this would constitute a solution to the articulatory bottleneck problem.

solution to the articulatory bottleneck problem is no longer one. If the frequency of scalar implicatures is low enough, having them be costless while their cancelation is effortful would lead to an overall *increase* in interpretive processing effort - precisely the opposite of Levinson's intention.

The results of Exp. 5 suggest that the Frequency Assumption indeed does not hold, at least for *some*, and at least in the Switchboard corpus, constituting a challenge to Levinson's theory. This work is only a first step towards a broader test of the Frequency Assumption. A more complete test would include investigating multiple scalar items over multiple genres of spoken and written language. However, it is striking that for *some*, the prime example of a lexical item that triggers scalar implicatures, average support for the implicature is not high. As far as I can tell, this is incompatible with the Default model.

In addition, the Default model does not predict that there should be gradient differences in the *strength* of a scalar inference based on contextual cues like partitivity, quantifier strength, or discourse accessibility. Under the Default model (in accordance with the standard assumption in the mainstream linguistic literature), a scalar implicature is either canceled or not. There is no room to talk about the degree to which a speaker implicates the upper-bound meaning based on context. However, the results of Exp. 6 clearly demonstrate that scalar implicature strength is subtly modulated by a great number of contextual cues.

5.4.2 Huang & Snedeker's Literal-First hypothesis

The second area that would be negatively affected by overturning the Frequency Assumption is the interpretation of experimental results showing that scalar implicatures (contrary to Levinson's prediction) incur a processing cost. These results have been taken to provide support for the Literal-First hypothesis:[10] the idea

[10]See Section 2.

that scalar implicature processing involves a two-step staged process whereby the semantic lower-bound interpretation is computed before the upper-bound pragmatic one (Huang & Snedeker, 2009). However, the cognitive cost associated with processing scalar implicatures can only be attributed to a staged literal-first process if the cost cannot be explained away by reference to another mechanism. Here is where the Frequency Assumption becomes important.

From a vast body of literature on frequency and predictability effects in other domains of language processing, it is well-known that more frequent or predictable words or structures are processed more quickly than less frequent or predictable words or structures. For example, more frequent words are recognized more quickly and more accurately than less frequent words (Dahan, Magnuson, & Tanenhaus, 2001; Marslen-Wilson, 1987; Seidenberg & McClelland, 1990). Similarly, more frequent, contextually predictable, and less surprising words and structures are read more rapidly than less frequent and predictable ones (Ehrlich & Rayner, 1981; Hale, 2001; Levy, 2008; Mcdonald & Shillcock, 2003).

Thus, under the Frequency Assumption and assuming that frequency or predictability effects apply at the pragmatic level just as at other levels of linguistic processing, arriving at the upper-bound pragmatic interpretation of *some* should be less cognitively costly than arriving at the lower-bound semantic interpretation. However, if the Frequency Assumption does not hold (and if in fact the semantic interpretation is more frequent than the pragmatic one), the inverse pattern is predicted. Thus, the actual time course pattern found in the literature is compatible both with the Literal-First hypothesis but also with an inverse frequency effect: scalar implicatures may be processed slowly not because there is an obligatory stage of semantic processing before pragmatic processing, but instead because of the infrequent occurrence and associated difficulty of computing the implicature.

The results of Exp. 5 call the validity of the Frequency Assumption into question. This poses a problem for the testability of the Literal-First hypothesis, since

any observed processing cost associated with an implicature may be explained either by frequency or by stages of processing. It is in fact unclear to me how one would tease these two explanations apart. Huang and Snedeker (2009) were aware of this issue when they reported, as an aside, a small corpus study they conducted on the British National Corpus (BNC). They extracted a random sample of 50 occurrences of *some*, "looked for cases that unambiguously referred to a subset" (Huang & Snedeker, 2009, p. 410), and found that this subset interpretation accounted for 42% of the sentences which they took as evidence that "the upper-bounded inference is often associated with the interpretation" (Huang & Snedeker, 2009, p. 410) of *some*. Three problems are of note. First, it is not clear how a subset interpretation was defined, i.e. what interpretive test was applied to each of the sentences. Second, 50 cases is a very small sample - there are a total of 147051 occurrences of *some* in the BNC (so 50 cases constitute 0.03% of the total number of cases of *some*). That is, the validity of this corpus study is very low. However, even when taking the results of this tiny dataset at face value, the Frequency Assumption does not hold, just as it did not hold in the study presented here. Given the scalar item *some*, more often than not the scalar implicature is not present.

5.4.3 Probabilistic, constraint-based scalar implicature

In the previous sections I argued that the results of Exps. 5 and 6 pose a challenge to both the Default and the Literal-First model. In this section I will make a positive argument for the type of account of scalar implicature that the results *do* support. In particular, the results are consistent with accounts that view the outcome of scalar implicature processing as probabilistic, and the process itself as highly context-dependent.

In Chapter 1, I argued that a theory of scalar implicature should be able to account for both judgment data (i.e. listeners' ultimate interpretations of an utter-

Table 5.5: Proportion of responses reflecting an upper-bound interpretation across experiments, in chronological order.

Experiment	Proportion	Experiment	Proportion
Noveck and Posada (2003)	63%	Bott and Noveck (2004), Exp. 3	61%
Geurts and Pouscoulous (2009)	3-94%	Zondervan (2010)	41-85%
Degen and Tanenhaus (2011)	9-54%	Chapter 3 (this thesis)	18-93%
Chapter 4 (this thesis)	65-78%		

ance) as well as time course data (i.e. the online computation of inferences).[11] The rating results presented in this Chapter provide support for the claim that scalar implicatures are *probabilistic* (M. C. Frank & Goodman, 2012; Russell, 2012). This is further backed up by the widely varying implicature rates observed in forced-choice truth value judgment tasks across many experiments, both between scalar items as well as within scalar items used in different contexts (see Table 5.5 for an incomplete list). Thus, rather than being an all-or-none phenomenon, speakers are taken to implicate the negation of the stronger proposition to varying degrees, a point made eloquently by Russell (2012):

> "[T]he degree to which an implicature is "felt" is simply dependent on the probability, calculated by the hearer, of the speaker's belief in a

[11] An additional type of data that a theory of scalar implicature should capture is acquisition data, but this lies outside the scope of this thesis. Note, however, that the results of Exp. 5 and 6 provide the beginning of an explanation for one of the main findings in the acquisition literature on scalar implicatures: that children learn the literal meaning of scalar terms before they learn the pragmatically enriched meaning (Barner et al., 2011; Katsos & Bishop, 2011; Noveck, 2001). Children typically acquire the use of more frequent words and structures before less frequent ones (J. C. Goodman, Dale, & Li, 2008; Huttenlocher, Haight, Bryk, Seltzer, & Lyons, 1991; Schwartz & Terrell, 1983). If the lower-bound interpretation is more frequent than the upper-bound interpretation in spontaneous speech, as I have shown here for *some*, acquiring the literal interpretation before the pragmatically enriched one falls out naturally as yet another frequency effect in acquisition. This also makes the interesting prediction that the age of acquisition of the upper-bound interpretation should vary between scales as a function of the average support for the upper-bound vs. lower-bound interpretation, given a particular scale.

> stronger proposition. The calculation of this probability depends, in turn, on the relevance, defined probabilistically, of the speaker's utterance compared to alternative utterances, and on comparative simplicity." – Russell (2012, p. 150)

This brings us to the second feature of scalar implicature that I hope to have provided further evidence for in this Chapter: scalar implicatures are highly sensitive to context, e.g. to salient alternative utterances, the relative relevance of those utterances to a contextually given QUD, the production costs of those alternatives, beliefs about the speaker's knowledge state, etc. Note that these are all features of the context that are independent of the actual lexical item that is involved in triggering the scalar reasoning process. The results of Exp. 6 suggest that the strength of scalar inferences associated with *some* additionally depend on a variety of *some*- (or quantifier-)specific cues like partitivity, quantifier strength, and discourse accessibility, which do not play a role in cases such as the biking past Mount Hope/Averill example (17).

Thus, a unified account of scalar implicature should take into account both abstract features common to all types of scalar implicatures while also allowing the idiosyncrasies of particular scales to enter into the computation. This is important because the additional information contained in the scale-specific cues may provide a "short-cut" to the speaker's intended meaning that obviates the need for complicated reasoning about speaker intentions. For example, only a quarter of the cases in the *some*-database were partitives, but most of those in turn received high implicature ratings. Thus, when observing a partitive after *some*, it's a good bet that the speaker intended the upper-bound interpretation. Compare that to a cue like linguistic mention of the embedded NP - here we observed that implicature ratings are higher for cases of linguistic mention, but the variance in ratings is much larger. That is, linguistic mention is a less informative cue than the partitive in that the variance in speaker intentions is greater for previously mentioned NPs

than it is for partitives. Put differently, relying only on the partitive to estimate whether a speaker intended the upper-bound interpretation will lead to fewer errors than relying on linguistic mention. If listeners tracked the joint statistics of this plethora of cues along with speaker intentions, they would be equipped with a powerful basis for using context to inform pragmatic inference. The Constraint-Based account introduced in Section 2.2.2 makes exactly this assumption.

The probabilistic accounts mentioned above have begun to formally implement these desiderata. For example, the output of models like those of M. C. Frank and Goodman (2012); N. D. Goodman and Stuhlmüller (2013) and Russell (2012) can be interpreted as the listener's subjective probability that the speaker intended the upper-bound interpretation, given contextual constraints. Contextual constraints that have been implemented are e.g. the degree to which the speaker is assumed to know whether the stronger alternative holds (Franke, 2009; N. D. Goodman & Stuhlmüller, 2013), the cost of the assumed to be mutually known alternatives (Bergen, Goodman, & Levy, 2012; Franke, 2009; Russell, 2012), or the degree to which alternative utterances are assumed to be relevant to a contextually given QUD (Russell, 2012).

The great advantage of these models is that they make quantitative, empirically testable predictions about the degree to which an implicature should arise, given very explicit assumptions about contextual factors that are assumed to play a role. A current restriction of these models is that they make predictions about interpretations given *an entire utterance*. This is perfectly sufficient, of course, if all that one is interested in is understanding which factors affect listeners' final judgments. However, in practice utterances don't present themselves as whole units, but rather as incrementally unfolding speech/written text. Consequently, cues that may be used to identify the speaker's intention become available at different points in time. For example, the QUD may be given and generate some expectations about upcoming linguistic input even before the start of the utter-

ance. Information about the partitive becomes available before information about the head noun. Different discourse accessibility cues become available at different points in time in the utterance also, as do other cues. For example, if the *some*-NP is the grammatical subject of the utterance, that cue will become available before knowledge about linguistic mention (which is associated with the head noun), which in turn will become available before information about post-nominal modification. Thus, cues to the upper-bound interpretation that differ in reliability become available at different points in the unfolding speech stream, which may affect both listeners' final interpretation of the utterance as well as their certainty about the speaker's intention during the unfolding speech stream.

Current probabilistic models of scalar implicature make fine-grained predictions about listeners' final interpretations. The Constraint-Based account, making many of the same assumptions as the probabilistic accounts, adds the notion of incrementality. Combining these approaches promises to yield a formally explicit, psychologically plausible theory of scalar implicature based on (implicit or explicit) reasoning about alternative utterances the speaker could have made, whereby language-specific features of the context (e.g., the conventional meaning of the utterance) interact with domain-general inference and cue integration mechanisms, thus implementing Grice's vision of treating talking "as a special case or variety of purposive, indeed rational, behavior" (Grice, 1975).

5.5 Conclusion

In this Chapter I made a first step towards testing the assumption that scalar implicatures are very frequent, a basic tenet of the Default model of scalar implicature and an important assumption in the interpretation of "costly implicature" results as supporting the Literal-First hypothesis. At least for the case of implicatures from *some* to *not all*, which are the focus of this dissertation, it seems that

average support for scalar implicatures is low. However, this support, as measured in participants' implicature ratings, selectively increased when the *some*-NP contained a partitive, when the *some*-NP was in subject position, when the embedded NP was modified, when the embedded NP referent had been previously mentioned or was inferable, and when the use of *some* was relatively strong.

This work suggests that scalar implicatures behave much more like Particularized rather than like Generalized Conversational Implicatures - the strength with which they are felt is very much dependent on contextual cues. Thus one of the main predictions of the Constraint-Based account is borne out. These results are also broadly compatible with the probabilistic accounts of scalar implicature discussed in Chapter 2.

This work could be expanded in many different ways: empirically, the results should be replicated for different scales and corpora. Theoretically, current probabilistic accounts of scalar implicature should be extended to incorporate the use of multiple cues in pragmatic inference; while the Constraint-Based account should receive an explicit formal implementation. An important question that requires both empirical and theoretical investigation is how the timing of cues affects listeners' interpretations of utterances with *some* online.

6 Concluding remarks

The goal of this dissertation was to examine the role of context in the processing of scalar implicatures. In Chapter 1 I discussed reasons for why scalar implicature have traditionally been treated as seemingly context-independent inferences. In particular, the high degree of regularity with which scalar implicatures arise relative to other types of inferences has led researchers to posit a categorical distinction between scalar (and other highly regular) implicatures and more irregular implicatures: this is the GCI-PCI distinction. This dissertation has attempted to show that, contrary to the most widely held assumption, even the most regularized of pragmatic inferences - scalar implicatures from *some* to *not all* - are in fact highly context-dependent.

In the following I briefly summarize the experimental evidence presented in Chapters 3, 4, and 5, that supports the claim of context-dependence (Section 6.1). I then present the conclusions that this work allows (Section 6.2). Finally, I highlight some interesting open questions for future research (Section 6.3).

Table 6.1: Summary of main results. The following abbreviations are used: "SI" = "Scalar implicature from *some* to *not all*"; "are faster" = "are computed more quickly"; "FA" = "Frequency Assumption". Cues appear in small caps.

Result	Ch. (Exp.)
I. Listeners are aware of the relative naturalness of alternatives.	
1. Naturalness of *some* is lower for small set sizes when number alternatives are available than when they are not.	3, 4 (1a, 1b, 3a, 3b)
2. Naturalness of non-partitive *some* is higher than that of partitive *some* at the upper bound.	3 (1a, 1b)
3. Naturalness of *some* is lower than that of *all* at the upper bound.	3, 4 (1a, 1b, 3a, 3b)
4. Naturalness of *all* is lower for small set sizes when number alternatives are available than when they are not.	4 (3a, 3b)
II. SI *strength* is modulated by multiple contextual cues.	
1. Implicature strength is higher when the PARTITIVE is used.	3, 5 (2, 6)
2. Implicature rates increase with increasing RELEVANCE OF THE STRONGER ALTERNATIVE to a contextual QUD.	3 (2, 2a)
3. Implicature strength increases with increasing QUANTIFIER STRENGTH.	5 (6)
4. Implicature strength increases with increasing DISCOURSE ACCESSIBILITY of the embedded NP.	5 (6)
III. SI *time course* is modulated by multiple contextual cues.	
1. Implicatures are faster when PARTITIVE is used.	3 (2)
2. Implicatures are faster with increasing RELEVANCE OF THE STRONGER ALTERNATIVE to a contextual QUD.	3 (2, 2a)
3. Implicature speed decreases with decreasing UNCERTAINTY ABOUT THE QUD.	3 (2, 2a)
4. Implicatures are faster when NUMBER ALTERNATIVES are not available than when they are, especially for small SET SIZE.	4 (4a, 4b)
IV. Average support for SIs is low, pace FA.	5 (5)
V. Processing speed of literal content is affected by alternatives.	
1. The lower-bound interpretation of *some* is processed more slowly with decreasing naturalness of *some* (compared to available alternatives).	3 (2)
2. *All* is processed more slowly when number terms are available alternatives.	4 (4a, 4b)

6.1 Summary of results

Table 6.1 provides a summary of the main results reported in this dissertation. Both truth-value judgment tasks and gradient similarity rating tasks in conjunction with corpus analyses revealed that implicature strength is modulated by multiple contextual cues that provide more or less probabilistic support for the upper-bound interpretation of utterances with *some*. Examples of such cues are use of the partitive, the naturalness or expectedness of *some* compared to number terms, the relevance of the stronger alternative *all* to a contextual Question Under Discussion, quantifier strength, and the discourse accessibility of the embedded NP.

In addition, response time and eye movement analyses revealed that the time course of the inference process is similarly modulated by multiple cues. In particular, cue values that increase implicature strength were found to be associated with more rapid implicatures.

An additional important result is that not only the processing of pragmatic *some* is affected by the naturalness and availability of alternatives; literal content that has typically been treated in the scalar implicature literature as providing a constant baseline to compare non-literal content to was also found to be sensitive to alternatives. In particular, the naturalness of *all* with small sets decreased when number alternatives were available; conversely, eye movements reflected that participants were slower to converge on the *all*-target when number terms were available alternatives. Similarly, the lower-bound interpretation of *some* was shown to be affected by the availability of more natural alternatives. Naturalness of *some* was low for very small sets and for the unpartitioned set, where number terms and *all* are more natural alternatives; conversely, the lower-bound interpretation was processed more slowly for these set sizes than where *some* was most natural.

Finally, a test of the Frequency Assumption revealed that on average, support

for scalar implicatures from *some* to *not all* is low, counter to claims in the literature.

6.2 Implications

The main conclusion to be drawn from this work is that scalar implicature - both the outcome of the inference process and the inference process itself - is much more context-dependent than previously assumed. In computing scalar implicatures, listeners are highly sensitive to both scalar and non-scalar alternatives that the speaker could have produced, but didn't. In addition, average support for scalar implicatures "in the wild" is low. This stands in stark contrast to the received view of scalar implicature (especially the seemingly lexicalized *some-not-all* implicature) as a prime example of GCI. Where GCIs are taken to be highly regularized, frequent, and context-independent, this work has shown that scalar implicatures from *some* to *not all* are anything but. If anything, they seem to function more like PCIs. This by itself is not surprising - ad hoc scalar implicatures have been recognized as being similar to PCIs for a long time. However, this work shows that even the most regularized, lexicalized scalar implicatures have properties that make them more similar to PCIs than to GCIs.

Of course, identifying one type of scalar implicature that seems to not be a case of GCI does not mean that the distinction between GCI and PCI itself should be abandoned.[1] For example (barring the cases of unambiguously ad hoc scales), it may turn out that all other scales typically considered in the literature do give rise to scalar implicatures independently of context, and do so in a categorical way. However, I believe this is unlikely: the ⟨all, some⟩ scale has unambiguously been treated as the most clearly lexicalized scale. If any scale should be involved

[1] Arguments against the distinction itself were presented in Chapter 1 and throughout the dissertation.

in GCIs, it is this one. This makes it unlikely that the less lexicalized scales will be *less* context-dependent and more likely that they will be even *more* context-dependent. This is a matter for future work.

The context-dependence exhibited by scalar implicatures has implications for theoretical accounts of scalar implicature processing. On the one hand, it does not in and of itself pose a challenge for two-stage models. Both the Default and the Literal-First model allow for information from the context to be integrated in a second step and either cancel or generate the implicature. However, the obtained time course data suggest that context enters even the earliest moments of implicature processing. This is at odds with both of the two-stage accounts, which predict that either the upper- or the lower-bound interpretation should always be arrived at more slowly than the other one, respectively. The work presented here suggests that processing speed for both interpretations varies. Proponents of two-stage accounts may argue that the time course measures used are not sufficiently sensitive to the actual earliest moments of processing. While this cannot be ruled out, it is an argument that both a) renders all of these accounts untestable in the limit and b) distracts from the more interesting enterprise of identifying the contextual cues that listeners use in performing pragmatic inference and building explicit models of how those cues interact to give rise to perceived speaker meaning. In contrast, under the Constraint-Based and other contextualist accounts that view scalar implicature processing as a problem of probabilistic constraint-based updating of belief distributions over states of the world, this is precisely the resulting research program.

This work also provides further evidence for the recently emerging claim that scalar implicatures themselves are probabilistic: when not experimentally forced to commit to the upper- or lower-bound interpretation, listeners' gradient judgments suggest they perceive the implicated content as being conveyed more or less strongly. This is compatible with accounts that assume that listeners maintain

uncertainty over possible states of the world even after observing an entire utterance, e.g. the Constraint-Based account and other probabilistic accounts discussed in Chapter 2, in contrast to accounts that treat scalar implicature as a categorical phenomenon, e.g. the Default model, Literal-First hypothesis, and Relevance Theory.

Finally, an important issue concerns the role of alternatives in scalar implicature processing. What this work has shown is that not only scalar, but non-scalar alternatives to *some* affect the strength and speed of scalar implicature processing. This adds further complexity to the already difficult problem of defining constraints on scalar alternatives. While it is clear that the space of alternatives cannot be entirely unconstrained, it will be a challenge to understand exactly what those constraints are. Presumably, the alternatives to a given scalar item will be selected based on both top-down, goal-driven expectations about relevance and informativeness (yielding, for the *some* case, both *all* and number terms as alternatives in the right contexts) as well as on low-level features of words or phrases like syntactic or string predictability, i.e., the probability of a particular alternative expression occurring, given the listener's language model and a linguistic context. For example, a probabilistic context-free grammar might be used to incrementally generate the most likely alternatives at different points in an unfolding utterance; the relative probability of those alternatives, in conjunction with considerations of relevance, informativeness, and complexity and the assumption that speakers generally try to say true things, might yield a pruned set of most salient contextual alternatives. Priming-based methods probing the activation of particular alternatives at the point in the utterance where the scalar item would have otherwise been used could be used to evaluate such models of alternatives. This is purely speculative, of course.

6.3 Future directions

This dissertation provides the basis for a research program dedicated to studying pragmatic inference as the result of optimal integration of multiple probabilistic cues to speaker meaning. Future work should consist in three related lines of work.

First, the Constraint-Based account should receive a formally explicit computational implementation that makes clearly testable quantitative predictions. As discussed in Chapter 2, already existing Bayesian models of pragmatic inference are a natural candidate. This will require generating good cue estimates and integrating considerations of timing differences in cue availability.

Second, this dissertation has been concerned only with scalar implicatures from *some* to *not all*. Future work should investigate the context-dependence of other scales to address whether the results obtained here are just idiosyncrasies of the ⟨all, some⟩ scale or whether (as predicted by the Constraint-Based account), all scalar implicatures are a probabilistic result of multiple cue integration.

Third, as touched on above, this work has made clear that a broader theory of alternatives in scalar implicature is required. Developing such a theory, perhaps along the lines sketched above, should be a main priority.

In sum, as we better understand the relevant contextual constraints and the distribution of usage of scalar terms and their associated inferences, I expect that pragmatic inference, like ambiguity resolution, will turn out to be consistent with approaches to language processing that are grounded in constraint-based and information theoretic principles. Unconstrained inference might be slow and costly. But inference that is constrained by rich conversational context and natural use of linguistic forms might be remarkably easy. If so, then a unified account of the speed and efficiency of language processing might indeed be possible.

References

Abbott, B. (1996). Doing Without a Partitive Constraint. In J. Hoeksema (Ed.), *Partitives: Studies on the Syntax and Semantics of the Partitive and Related Constructions* (pp. 25 – 56). Berlin and New York: Mouton de Gruyter.

Altmann, G., & Kamide, Y. (1999). Incremental interpretation at verbs: restricting the domain of subsequent reference. *Cognition*, *73*(3), 247–264.

Armstrong, S. L., Gleitman, L. R., & Gleitman, H. (1983). What some concepts might not be. *Cognition*, *13*(3), 263–308.

Atkinson, J., Campbell, F. W., & Francis, M. R. (1976). The magic number 4 +/- 0: a new look at visual numerosity judgements. *Perception*, *5*(3), 327–34.

Atlas, D. J., & Levinson, S. C. (1981). It-Clefts, Informativeness, and Logical Form: Radical Pragmatics (Revised Standard Version). In P. Cole (Ed.), *Radical Pragmatics* (pp. 1–61). New York: Academic Press.

Baayen, R. (2008). *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R* (Vol. 2; R. H. Baayen, Ed.) (No. 3). Cambridge: Cambridge University Press.

Baayen, R., Davidson, D., & Bates, D. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412.

Bach, K. (1999). The Myth of Conventional Implicature. *Linguistics and Philosophy*, *22*, 327–366.

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition*, *113*(3), 329–49.

Barker, C. (1998). Partitives , Double Genitives and Anti-Uniqueness. *Natural Language & Linguistic Theory*, *16*(4), 679–717.

Barner, D., Brooks, N., & Bale, A. (2011). Accessing the unsaid: the role of scalar alternatives in children's pragmatic inference. *Cognition*, *118*(1), 84–93.

Barr, D. J. (2008). Pragmatic expectations and linguistic evidence: Listeners anticipate but do not integrate common ground. *Cognition*, *109*(1), 18–40.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure in mixed-effects models: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255 – 278.

Barwise, J., & Cooper, R. (1981). Generalized Quantifiers and Natural Language. *Linguistics and Philosophy*, *4*(2), 159–219.

Benaglia, T., Chaveau, D., Hunter, D. R., & Young, D. S. (2009). mixtools: An R Package for Analyzing Finite Mixture Models. *Journal of Statistical Software*, *32*(1).

Benz, A., & Rooij, R. (2007). Optimal assertions, and what they implicate. A uniform game theoretic approach. *Topoi*, *26*(1), 63–78.

Bergen, L., Goodman, N. D., & Levy, R. (2012). That's what she (could have) said: How alternative utterances affect language use. In *Proceedings of the Thirty-Fourth Annual Conference of the Cognitive Science Society.*

Bergen, L., & Grodner, D. J. (2012). Speaker knowledge influences the comprehension of pragmatic inferences. *Journal of experimental psychology. Learning, memory, and cognition*, *38*(5), 1450–60.

Birner, B. J. (1997). The linguistic realization of inferrable information. *Language and Communication*, *17*(2), 133 – 147.

Bott, L., Bailey, T. M., & Grodner, D. (2012). Distinguishing speed from accuracy in scalar implicatures. *Journal of Memory and Language*, *66*(1), 123–142.

Bott, L., & Noveck, I. (2004). Some utterances are underinformative: The onset and time course of scalar inferences. *Journal of Memory and Language*, *51*(3), 437–457.

Breheny, R. (2008). A New Look at the Semantics and Pragmatics of Numerically Quantified Noun Phrases. *Journal of Semantics*, *25*(2), 93–139.

Breheny, R., Ferguson, H. J., & Katsos, N. (2013). Taking the epistemic step: Toward a model of on-line access to conversational implicatures. *Cognition*, *126*(3), 423–40.

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*(3), 434–63.

Brown, C., Hagoort, P., & Kutas, M. (2000). Postlexical integration processes in language comprehension: evidence from brain-imaging research. In M. S. Gazzaniga (Ed.), *The Cognitive Neurosciences* (2nd ed., pp. 881–895). Cambridge: MIT Press.

Brown-Schmidt, S., Gunlogson, C., & Tanenhaus, M. K. (2008). Addressees distinguish shared from private information when interpreting questions during interactive conversation. *Cognition*, *107*(3), 1122–34.

Carlson, G. N. (1977). A Unified Analysis of the English Bare Plural. *Linguistics and Philosophy*, *1*(3), 413–456.

Carnap, R. (1950). *Logical Foundations of Probability*. Chicago: University of Chicago Press.

Carston, R. (1998). Informativeness, Relevance and Scalar Implicature. In R. Carston & S. Uchida (Eds.), *Relevance Theory: Applications and Implications* (pp. 179–236). Amsterdam: John Benjamins.

Chambers, C. G. (2002). Circumscribing Referential Domains during Real-Time Language Comprehension. *Journal of Memory and Language*, *47*(1), 30–49.

Chambers, C. G., Tanenhaus, M. K., & Magnuson, J. S. (2004). Actions and

Affordances in Syntactic Ambiguity Resolution. *Journal of experimental psychology: Learning, memory, and cognition*, *30*(3), 687 – 696.

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* (Vol. 3, pp. 39–103). Oxford University Press.

Chierchia, G., Crain, S., Teresa, M., Guasti, M. T., Gualmini, A., & Meroni, L. (2001). The Acquisition of Disjunction: Evidence for a Grammatical View of Scalar Implicatures. In A. H.-J. D. Et al. (Ed.), *BUCLD 25 Proceedings* (pp. 157–168). Somerville, MA: Cascadilla Press.

Chierchia, G., Fox, D., & Spector, B. (2008). The Grammatical View of Scalar Implicatures and the Relationship between Semantics and Pragmatics. In K. von Heusinger, C. Maienborn, & P. Portner (Eds.), *Handbook of Semantics* (pp. 1–43). Mouton de Gruyter.

Cooper, R. M. (1974). The control of eye fixation by the meaning of spoken language: A new methodology for the real-time investigation of speech perception, memory, and language processing. *Cognitive Psychology*, *6*(1), 84–107.

Dahan, D., Magnuson, J. S., & Tanenhaus, M. K. (2001). Time course of frequency effects in spoken-word recognition: evidence from eye movements. *Cognitive psychology*, *42*(4), 317–67.

De Hoop, H. (1997). A semantic reanalysis of the partitive constraint. *Lingua*, *103*, 151 – 174.

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.

Degen, J., & Franke, M. (2012). Optimal Reasoning About Referential Expres-

sions. In S. Brown-Schmidt, J. Ginzburg, & S. Larsson (Eds.), *Proceedings of the 16th Workshop on the Semantics and Pragmatics of Dialogue* (pp. 2 – 11).

Degen, J., Franke, M., & Jäger, G. (2013). Cost-Based Pragmatic Inference about Referential Expressions. In *Proceedings of the 35th Annual Conference of the Cognitive Science Society.*

Degen, J., & Jaeger, T. F. (2011). *The TGrep2 Database Tools.* Unpublished manuscript.

Degen, J., & Jaeger, T. F. (in prep.). *Meaning and processing pressures are some (of the) reasons for choosing the partitive.*

Degen, J., & Tanenhaus, M. K. (2011). Making inferences: the case of scalar implicature processing. In L. Carlson, C. Hölscher, & T. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 3299–3304).

de Jong, F., & Verkuyl, H. (1985). Generalized Quantifiers: the Properness of their Strength. In J. van Benthem & A. ter Meulen (Eds.), *Generalized Quantifiers: Theory and Applications* (pp. 21–43). Dordrecht: Foris Publications.

Dell, G. S., Schwartz, M. F., Martin, N., Saffran, E. M., & Gagnon, D. A. (2000). The Role of Computational Models in Neuropsychological Investigations of Language: Reply to Ruml and Caramazza (2000). *Psychological Review*, *107*(3), 635–645.

Ehrlich, S. F., & Rayner, K. (1981). Contextual Effects on Word Perception and Eye Movements during Reading. *Journal of Verbal Learning and Verbal Behavior*, *20*, 641 –655.

Elman, J. L., Hare, M., & McRae, K. (2004). Cues, constraints, and competition in sentence processing. In M. Tomasello & D. Slobin (Eds.), *Beyond Nature-Nurture: Essays in Honor of Elizabeth Bates* (pp. 111–138). Mahwah, NJ:

Erlbaum.

Frank, A., & Jaeger, T. F. (2008). Speaking rationally: Uniform information density as an optimal strategy for language production. In *The 30th Annual Meeting of the Cognitive Science Society* (pp. 939 – 944).

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

Frank, M. C., Goodman, N. D., & Tenenbaum, J. B. (2009). Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science*, *20*(5), 578–85.

Franke, M. (2009). *Signal to Act: Game Theory in Pragmatics*. Unpublished doctoral dissertation, Universiteit van Amsterdam.

Franke, M. (2011). Quantity implicatures, exhaustive interpretation, and rational conversation. *Semantics and Pragmatics*, *4*(1), 1–82.

Gazdar, G. (1979). *Pragmatics: Implicature, Presupposition, and Logical Form*. Academic Press.

Geurts, B., & Pouscoulous, N. (2009). Embedded implicatures?!? *Semantics and Pragmatics*, *2*, 1–34.

Godfrey, J., Holliman, E., & McDaniel, J. (1992). Switchboard: A Telephone Speech Corpus for Research and Development. In *Proceedings of ICASSP-92* (pp. 517 – 520).

Goodman, J. C., Dale, P. S., & Li, P. (2008). Does frequency count? Parental input and the acquisition of vocabulary. *Journal of child language*, *35*(3), 515–31.

Goodman, N. D., & Stuhlmüller, A. (2013). Knowledge and implicature: modeling language understanding as social cognition. *Topics in Cognitive Science*, *5*(1), 173–84.

Goodman, N. D., Ullman, T. D., & Tenenbaum, J. B. (2011). Learning a theory of causality. *Psychological Review*, *118*(1), 110–9.

Green, M. S. (1995). Quantity, Volubility, and Some Varieties of Discourse. *Linguistics and Philosophy*, *18*(1), 83–112.

Grice, H. P. (1969). Utterer's Meaning and Intentions. *Philosophical Review*, *78*(2), 147–177.

Grice, H. P. (1975). Logic and Conversation. *Syntax and Semantics*, *3*, 41–58.

Grice, H. P. (1978). Further Notes on Logic and Conversation. *Syntax and Semantics*, *9*, 183–197.

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*(1), 42–55.

Grodner, D. J., & Sedivy, J. C. (2011). The Effect of Speaker-Specific Information on Pragmatic Inferences. In N. Pearlmutter & E. Gibson (Eds.), *The Processing and Acquisition of Reference* (Vol. 2327, pp. 239–272). Cambridge, MA: MIT Press.

Gundel, J. K., Ntelitheos, D., & Kowalsky, M. (2007). Children's Use of Referring Expressions: Some Implications for Theory of Mind. In D. Bittner & N. Gagarina (Eds.), *ZAS Papers in Linguistics: Intersentential Pronominal Reference in Child and Adults Langauge. Proceedings of the Conference on Intersentential Pronominal Reference in Child and Adult Language, 48.*

Hale, J. (2001). A Probabilistic Earley Parser as a Psycholinguistic Model. In *Proceedings of the Second Meeting of the North American Chapter of the Asssociation for Computational Linguistics.*

Hanna, J. E., & Tanenhaus, M. K. (2004). Pragmatic effects on reference resolution in a collaborative task: evidence from eye movements. *Cognitive Science*, *28*(1), 105–115.

Hanna, J. E., Tanenhaus, M. K., & Trueswell, J. C. (2003). The effects of common ground and perspective on domains of referential interpretation. *Journal of Memory and Language*, *49*(1), 43–61.

Heller, D., Grodner, D., & Tanenhaus, M. K. (2008). The role of perspective in identifying domains of reference. *Cognition*, *108*(3), 831–6.

Hirschberg, J. (1985). *A Theory of Scalar Implicature.* Unpublished doctoral dissertation, University of Pennsylvania.

Horn, L. (1972). *On the Semantic Properties of the Logical Operators in English.* Unpublished doctoral dissertation, UCLA.

Horn, L. (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* (pp. 11–42). Washington: Georgetown University Press.

Horn, L. (1997). All John's children are as bald as the King of France: Existential import and the geometry of opposition. In *CLS 33* (pp. 155 – 179).

Horn, L. (2004). Implicature. In L. Horn & G. Ward (Eds.), *Handbook of Pragmatics.* Blackwell.

Huang, Y. T., Hahn, N., & Snedeker, J. (2010). Some inferences still take time: Prosody, predictability, and the speed of scalar implicatures. *Poster presented at the 23rd annual CUNY conference on Human Sentence Processing*.

Huang, Y. T., & Snedeker, J. (2009). Online interpretation of scalar quantifiers: insight into the semantics-pragmatics interface. *Cognitive Psychology*, *58*(3), 376–415.

Huang, Y. T., & 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.

Huttenlocher, J., Haight, W., Bryk, A., Seltzer, M., & Lyons, T. (1991). Early Vcabulary Growth: Relation to Language Input and Gender. *Developmental Psychology*, *27*(2), 236–248.

Ionin, T., Matushansky, O., & Ruys, E. (2006). Parts of speech: Toward a unified

semantics for partitives. In C. Davis, A. Deal, & Y. Zabbal (Eds.), *Proceedings of NELS 36.* Amherst, MA: University of Massachusetts, GLSA.

Israel, M. (1999). 'Some' and the pragmatics of indefinite construal. *Proceedings of the Berkeley Linguistics Society*, *25*, 169 – 182.

Jackendoff, R. (1977). *X-Bar Syntax: A Study of Phrase Structure.* Cambridge, MA: MIT Press.

Jacobs, R. a., Jordan, M. I., Nowlan, S. J., & Hinton, G. E. (1991). Adaptive Mixtures of Local Experts. *Neural Computation*, *3*(1), 79–87.

Jaeger, T. F. (2006). *Redundancy and Reduction in Spontaneous Speech.* Unpublished doctoral dissertation, Stanford University.

Jaeger, T. F. (2008). Categorical Data Analysis: Away from ANOVAs (transformation or not) and towards Logit Mixed Models. *Journal of memory and language*, *59*(4), 434–446.

Jaeger, T. F. (2010). Redundancy and reduction: speakers manage syntactic information density. *Cognitive Psychology*, *61*(1), 23–62.

Jäger, G. (2013). *Rationalizable signaling.* to appear in Erkenntnis.

Karttunen, L., & Peters, S. (1979). Conventional implicature. In C.-K. Oh & D. A. Dinneen (Eds.), *Syntax and Semantics* (pp. 1–56). Academic Press.

Katsos, N., & Bishop, D. V. M. (2011). Pragmatic tolerance: implications for the acquisition of informativeness and implicature. *Cognition*, *120*(1), 67–81.

Katsos, N., Breheny, R., & Williams, J. (2005). The Interaction of Structural and Contextual Constraints During the On-line Generation of Scalar Inferences. In B. Bara, L. Barsalou, & M. Bucciarelli (Eds.), *Proceedings of the 27th Annual Conference of the Cognitive Science Society* (pp. 1108–1113). Mahwah, NJ: Erlbaum.

Kaufman, E., Lord, M., Reese, T., & Volkman, J. (1949). The Discrimination of Visual Number. *The American Journal of Psychology*, *62*(4), 498–525.

Keenan, E. O. (1976). The Universality of Conversational Postulates. *Language*

in Society, *5*(1), 67–80.

Keysar, B., Barr, D. J., Balin, J. a., & Brauner, J. S. (2000). Taking Perspective in Conversation: The Role of Mutual Knowledge in Comprehension. *Psychological Science*, *11*(1), 32–38.

Kronmüller, E., & Barr, D. J. (2007). Perspective-free pragmatics: Broken precedents and the recovery-from-preemption hypothesis. *Journal of Memory and Language*, *56*(3), 436–455.

Kutas, M., & Hillyard, S. A. (1980). Reading Senseless Sentences: Brain Potentials Reflect Semantic Incongruity. *Science*, *207*(4427), 203–205.

Ladusaw, W. A. (1979). *Polarity Sensitivity as Inherent Scope Relations*. Unpublished doctoral dissertation, University of Texas at Austin.

Ladusaw, W. A. (1982). Semantic constraints on the English partitive construction. In D. Flickinger, M. Macken, & N. Wiegand (Eds.), *Proceedings of the First West Coast Conference on Formal Linguistics* (pp. 231 – 242). Stanford, CA: CSLI Publications.

Ladusaw, W. A. (1994). Thetic and categorical, stage and individual, weak and strong. In M. Harvey & L. Santelmann (Eds.), *Proceedings of SALT IV* (pp. 220 – 229). Cornell U. DMLL, Ithaca, NY.

Levinson, S. C. (2000). *Presumptive Meanings - The Theory of Generalized Conversational Implicature*. MIT Press.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–77.

Levy, R., & Jaeger, T. F. (2007). Speakers optimize information density through syntactic reduction. In B. Schlökopf, J. Platt, & T. Hoffman (Eds.), *Advances in Neural Information Processing Systems* (Vol. 19, pp. 849–856). Cambridge, MA: MIT Press.

Lewis, D. (1969). *Convention. A Philosophical Study.* Harvard University Press.

Lumsden, M. (1988). *Existential Sentences: Their Structure and Meaning*. Lon-

don: Croom Helm.

MacDonald, M., Pearlmutter, N., & Seidenberg, M. (1994). The lexical nature of syntactic ambiguity resolution. *Psychological Review*, *101*, 676–703.

Mandler, G., Shebo, B. J., & Vol, I. (1982). Subitizing: An Analysis of Its Component Processes. *Journal of Experimental Psychology: General*, *111*(1), 1 – 22.

Marcus, M., Santorini, B., Marcinkiewicz, M., & Taylor, A. (1999). *Treebank-3.*

Marcus, M. P., Santorini, B., & Marcinkiewicz, M. (1993). Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, *19*(2), 313–330.

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information.* San Francisco: W.H. Freeman.

Marslen-Wilson, W. D. (1987). Functional parallelism in spoken word-recognition. *Cognition*, *25*(1-2), 71–102.

Matsumoto, Y. (1995). The Conversational Condition on Horn scales. *Linguistics and Philosophy*, *18*(1), 21–60.

Mcdonald, S. A., & Shillcock, R. C. (2003). Eye movements reveal the on-line computation of lexical probabilities during reading. *Psychological Science*, *14*(6), 648 – 652.

McNally, L., & Geenhoven, V. V. (1998). *Redefining the weak/strong distinction.* Paper presented at the 1997 Paris Syntax and Semantics Colloquium.

McRae, K., Spivey-Knowlton, M. J., & Tanenhaus, M. K. (1998). Modeling the Influence of Thematic Fit (and Other Constraints) in On-line Sentence Comprehension. *Journal of Memory and Language*, *38*(3), 283–312.

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience*, *24*, 167–202.

Milsark, G. (1974). *Existential Sentences in English.* Unpublished doctoral dis-

sertation, MIT.

Milsark, G. (1977). Toward an Explanation of Certain Peculiarities of the Existential Construction in English. *Linguistic Analysis*, *3*, 1 – 30.

Nadig, A. S., & Sedivy, J. C. (2002). Evidence of Perspective-Taking Constraints in Children's On-Line Reference Resolution. *Psychological Science*, *13*(4), 329–336.

Neely, J. H. (1977). Semantic priming and retrieval from lexical memory: Roles of inhibitionless spreading activation and limited-capacity attention. *Journal of Experimental Psychology: General*, *106*(3), 226–254.

Nieuwland, M. S., & Van Berkum, J. J. a. (2006). When peanuts fall in love: N400 evidence for the power of discourse. *Journal of Cognitive Neuroscience*, *18*(7), 1098–111.

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

Noveck, I., & Posada, A. (2003). Characterizing the Time Course of an Implicature: an Evoked Potentials Study. *Brain and Language*, *85*(2), 203–210.

Noveck, I., & Reboul, A. (2008). Experimental pragmatics: a Gricean turn in the study of language. *Trends in Cognitive Sciences*, *12*(11), 425–431.

Novick, J. M., Trueswell, J. C., & Thompson-Schill, S. L. (2005). Cognitive control and parsing: Reexamining the role of Broca's area in sentence comprehension. *Cognitive, Affective, & Behavioral Neuroscience*, *5*(3), 263–281.

Papafragou, A., & Musolino, J. (2003). Scalar implicatures: experiments at the semantics-pragmatics interface. *Cognition*, *86*(3), 253–82.

Parikh, P. (1991). Communication and strategic inference. *Linguistics and Philosophy*, *14*, 473–514.

Piantadosi, S. T., Tily, H., & Gibson, E. (2012). The communicative function of ambiguity in language. *Cognition*, *122*(3), 280–91.

Posner, M. I., & Snyder, C. R. (1975). Facilitation and inhibition in the processing

of signals. In P. Rabbitt & S. Dornic (Eds.), *Attention and Performance* (pp. 669 – 682). New York: Academic Press.

Potts, C. (2005). *The Logic of Conventional Implicatures*. Oxford: Oxford University Press.

Prince, E. F. (1981). Toward a taxonomy of given/new information. In P. Cole (Ed.), *Radical Pragmatics* (pp. 223 – 254). New York: Academic Press.

Qian, T., & Jaeger, T. F. (2012). Cue effectiveness in communicatively efficient discourse production. *Cognitive Science*, *36*(7), 1312–36.

Recanati, F. (2004). Embedded implicatures. *Philosophical Perspectives*, *17*, 1299 –1332.

Reed, A. M. (1991). On interpreting partitives. In D. Napoli & J. Kegl (Eds.), *Bridges between psychology and linguistics: A Swarthmore Festschrift for Lila Gleitman* (pp. 207 – 223). Hillsdale, NJ: Erlbaum.

Richardson, S., & Green, P. J. (1997). On Bayesian Analysis of Mixtures with an Unknown Number of Components. *Journal of the Royal Statistical Society*, *59*(4), 731–792.

Roberts, C. (1996). Information Structure in Discourse: Towards an Integrated Formal Theory of Pragmatics. In J. H. Yoon & A. Kathol (Eds.), *OSU Working Papers in Linguistics 49: Papers in Semantics* (pp. 91–136). Columbus, The Ohio State University.

Roberts, C. (2004). Information structure in discourse. *Semantics and Pragmatics*, *5*, 1 – 69.

Rohde, D. (2005). *TGrep2 User Manual*. Unpublished manuscript.

Rohde, H., Seyfarth, S., Clark, B., Jäger, G., & Kaufmann, S. (2012). Communicating with Cost-based Implicature: a Game-Theoretic Approach to Ambiguity. In *Proceedings of the 16th Workshop on the Semantics and Pragmatics of Dialogue* (pp. 107 – 116).

Rooy, R. V. (2003). Conversational implicatures and communication theory.

In J. van Kuppevelt & R. Smith (Eds.), *Current and New Directions in Discourse and Dialogue* (pp. 283–303). Dordrecht: Kluwer.

Ross, I. (2006). *Games Interlocutors Play: New Adventures in Compositionality and Conversational Implicature.* Phd thesis, University of Pennsylvania.

Russell, B. (2006). Against grammatical computation of scalar implicatures. *Journal of Semantics*, *23*(4), 361.

Russell, B. (2012). *Probabilistic Reasoning and the Computation of Scalar Implicatures.* Unpublished doctoral dissertation, Brown University.

Sauerland, U. (2004). Scalar Implicatures in Complex Sentences. *Linguistics and Philosophy*, *27*(3), 367–391.

Schwartz, R., & Terrell, B. (1983). The role of input frequency in lexical acquisition. *Journal of Child Language*, *10*, 57–64.

Sedivy, J. C., Tanenhaus, M. K., Chambers, C. G., & Carlson, G. N. (1999). Achieving incremental semantic interpretation through contextual representation. *Cognition*, *71*(2), 109–47.

Seidenberg, M. S., & McClelland, J. L. (1990). A distributed, developmental model of word recognition and naming. *Psychological Review*, *96*(4), 523 – 568.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell Systems Technical Journal*, *27*(4), 623–656.

Sharvit, Y., & Gajewski, J. (2008). On the Calculation of Local Implicatures. In *Proceedings of the 26thWest Coast Conference on Formal Linguistics* (pp. 411–419). Cascadilla Press.

Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing. *Psychological Review*, *84*, 127–190.

Sperber, D., & Wilson, D. (1995). *Relevance: Communication and Cognition.* Oxford: Blackwell.

Spivey, M. J., Tanenhaus, M. K., Eberhard, K. M., & Sedivy, J. C. (2002). Eye

movements and spoken language comprehension: effects of visual context on syntactic ambiguity resolution. *Cognitive Psychology*, *45*(4), 447–81.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement Learning*. MIT Press.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, *268*, 1632 – 1634.

Tanenhaus, M. K., & Trueswell, J. C. (1995). Sentence comprehension. In J. L. Miller & P. D. Elmas (Eds.), *Speech, Language, and Communication. Handbook of Perception and Cognition.* (pp. 217 – 262). San Diego, CA: Academic Press.

Tenenbaum, J. B., Griffiths, T. L., & Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences*, *10*(7), 309–18.

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.

Trueswell, J. C., Sekerina, I., Hill, N. M., & Logrip, L. (1999). The kindergarten-path effect: studying on-line sentence processing in young children. *Cognition*, *73*, 898–911.

Trueswell, J. C., Tanenhaus, M. K., & Kello, C. (1993). Verb-specific constraints in sentence processing: separating effects of lexical preference from garden-paths. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *19*(3), 528–53.

Ullman, T. D., Goodman, N. D., & Tenenbaum, J. B. (2012). Theory learning as stochastic search in the language of thought. *Cognitive Development*, *27*(4), 455–480.

van Kuppevelt, J. (1996). Inferring from topics. *Linguistics and Philosophy*, *19*(4), 393–443.

van Rooij, R., & Schulz, K. (2004). Exhaustive Interpretation of Complex Sentences. *Journal of Logic, Language and Information*, *13*(4), 491–519.

Webber, B. L. (1983). So what can we talk about now? In M. Brady & R. Berwick (Eds.), *Computational Models of Discourse* (pp. 331 – 371). MIT Press.

Wilson, D., & Sperber, D. (1995). Relevance Theory. In G. Ward & L. Horn (Eds.), *Handbook of Pragmatics* (pp. 607–632). Oxford: Blackwell.

Wu, S., & Keysar, B. (2007). The effect of culture on perspective taking. *Psychological Science*, *18*(7), 600–6.

Xu, F., & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological review*, *114*(2), 245–72.

Zipf, G. K. (1949). *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*. New York: Addison-Wesley.

Zondervan, A. (2008). Experiments on QUD and focus as a contextual constraint on scalar implicature calculation. In U. Sauerland & K. Yatsushiro (Eds.), *From Experiment to Theory, Proceedings of Experimental Pragmatics 2007*.

Zondervan, A. (2010). *Scalar implicatures or focus: an experimental approach*. Unpublished doctoral dissertation, Universiteit Utrecht, Amsterdam.

Zweig, E. (2009). Number-Neutral Bare plurals and the Multiplicity Implicature. *Linguistics and Philosophy*, *32*(4), 353 – 407.

A Sampled set sizes in Exps. 1

Table A.1: Set sizes sampled by the twelve base lists used in Exps. 1a and 1b

List	Set sizes	List	Set sizes
1	1, 2, 5, 9	7	2, 3, 5, 9
2	1, 2, 6, 10	8	2, 3, 6, 10
3	1, 3, 5, 9	9	2, 4, 7, 11
4	1, 3, 7, 11	10	2, 4, 8, 12
5	1, 4, 6, 10	11	3, 4, 7, 11
6	1, 4, 8, 12	12	3, 4, 8, 12

B Full post hoc mixed effects linear regression model for Exp. 2a

Table B.1: Full post hoc mixed effects linear regression model predicting log-transformed response time from fixed effects of relevance, response, responder type, their interactions, and response inconsistency, as well as random by-participant intercepts. All fixed effects predictors were centered before entering the analysis with the exception of the 3-level relevance predictor, which was Helmert-coded.

	Coef β	SE(β)	**t**	p
Intercept	7.35	0.04	196.5	**<.0001**
Relevant.vs.Rest	−0.25	0.09	−2.8	**<.01**
Lessrelevant.vs.Unclear	0.24	0.10	2.5	**<.01**
Response	−0.04	0.05	−0.9	<.41
Responder type	0.06	0.06	0.9	<.35
Response inconsistency	0.01	0.02	0.4	<.39
Relevant.vs.Rest:Response	−0.01	0.17	−0.1	<.94
Lessrelevant.vs.Unclear:Response	−0.11	0.18	−0.6	<.59
Relevant.vs.Rest:Responder type	−0.58	0.20	−2.9	**<.01**
Lessrelevant.vs.Unclear:Responder type	0.44	0.21	2.1	<.06
Response:Responder type	−0.30	0.10	−3.0	**<.01**
Rel.vs.Rest:Response:Responder type	0.53	0.35	1.5	<.14
Lessrel.vs.Unclear:Response:Responder type	−0.83	0.35	−2.4	**<.05**

C Full mixed effects linear regression model for Exp. 6

Table C.1: Full mixed effects linear regression model predicting implicature ratings from fixed effects for cues of interest and log-transformed sentence length as well as random by-participant intercepts. All fixed effects predictors were centered before entering the analysis.

	Coef β	SE(β)	**t**	p
Intercept	4.02	0.05	77.7	**<.0001**
Partitive	1.01	0.05	22.0	**<.0001**
Strength	−0.54	0.03	−21.1	**<.0001**
Linguistic mention	0.33	0.04	8.6	**<.0001**
Topicality	0.43	0.05	8.2	**<.0001**
Modification	0.11	0.03	3.2	**<.01**
Sentence length	0.15	0.03	5.3	**<.0001**
Partitive:Strength	0.43	0.05	8.4	**<.0001**
Linguistic mention:Topicality	0.17	0.12	1.3	>0.18
Linguistic mention:Modification	0.33	0.07	4.5	**<.0001**
Topicality:Modification	0.25	0.10	2.5	**<.05**
Linguistic mention:Topicality:Modification	0.54	0.24	2.2	**<.05**