UNIVERSITY OF CALIFORNIA
SANTA CRUZ

ON THE TIMING OF DECISIONS ABOUT MEANING DURING INCREMENTAL COMPREHENSION

A dissertation submitted in partial satisfaction of the
requirements for the degree of

DOCTOR OF PHILOSOPHY

in

LINGUISTICS

by

John Duff

December 2023

The Dissertation of John Duff
is approved:

Professor Pranav Anand, Co-Chair

Assistant Professor Amanda Rysling, Co-Chair

Professor Matthew Wagers

Associate Professor Jesse A. Harris

Peter Biehl
Vice Provost and Dean of Graduate Studies

Table of Contents

List of Figures

List of Tables

Abstract

On the timing of decisions about meaning during incremental comprehension

John Duff

Language comprehension requires a complex series of decisions under uncertainty. This is especially obvious when one string may have multiple different interpretations, whether due to lexical ambiguity, or the potential for an inference beyond literal content. This dissertation profiles how the human system for language comprehension times those decisions, specifically when and why it sometimes postpones them. Evidence comes from nine reading experiments in English probing variation across a range of different types of uncertain meaning (homonymy and polysemy, predicate distributivity, scalar implicatures from *some*, and causal inferences from discourse coherence) and across two tasks (self-paced-reading and the Maze task of Forster et al., 2009). Diagnosing the presence of decisions by testing for garden-path effects and costs associated with selection against a bias, I highlight two key patterns. First, some decisions, e.g. sense specification for polysemous nouns, are delayed in normal reading, but occur immediately when a rapid decision would be more useful; I conclude that decisions to postpone are flexible and sensitive to a comprehender's goals. Second, possible scalar implicatures and causal inferences rapidly affect comprehension, but do not receive any typical, decisive commitment until much later; I conclude that comprehenders may develop expectations gradiently based on multiple possible interpretations before they make a firm decision. Throughout the dissertation, I explore how these and related facts might be explained as a consequence of the ways humans attempt to rationally allocate cognitive resources under uncertainty.

Acknowledgments

There is a typical belief in academia that as you progress towards your dissertation, you will know more and more about your subject area, so that by the time you defend, you feel like you are the world's foremost expert on your small patch of intellectual ground. Something else happened to me as my defense approached, to my delight: I became more and more aware of how much this work is a product of everyone who has inspired and supported me over the years. Sometimes that is very direct, like remembering the suggestion from a friend for a piece of terminology; sometimes it is more sweeping, like becoming aware of how much I owe my approaches and interests to particular instructors, mentors, and role models; and sometimes it is personal, like reminiscing about the way I have written some experimental items or analysis scripts happily working next to a friend or family member. I will try and fail to do some justice to all of these kinds of support and influence here.

I'll start by thanking the members of my committee, and Amanda Rysling first among them. This is appropriate, since she lies at the very beginning of the story of my career as a researcher. In spring 2015, she gave me my first hands-on experience with real scientific inquiry when I served as a research assistant under her supervision running participants in the UMass Phonetics Lab, and I was inspired from the beginning by her enthusiasm, dogged curiosity, and attention to detail. As time went on, she continued to have an outsized impact, introducing me to Lyn Frazier and Chuck Clifton (see below), coaching me through my first CUNY poster session, and offering tips on the PhD applications that ended up bringing me to Santa Cruz, just as she started her first quarter as an Assistant Professor. I had not necessarily expected to work with her closely here: after all, her principal interests are often in the domains of phonetics and speech processing, whereas I was by that point more interested in other components of comprehension. And yet, her advice and advising from that distance quickly revealed itself to be all the more valuable, and her encouragement to take inspiration from computationally-symmetric problems and solutions in other subfields has brought me to many of the projects that have excited me most. She may joke about delusions of nails owed to a hammer in the hand, but the truth is that you can make a lot of progress once you know all of the uses of a single tool. So, I thank her for every tool she's given me over the past nine years, and for the spirit to swing them with precision and intention.

Now, to Pranav Anand. With remarkable consistency, if you read the dissertations of any of the other students Pranav has advised, they will tell you that they entered Santa Cruz with other interests, and he turned them towards the light of formal semantics and pragmatics. This pattern has absolutely nothing to do with intentional evangelization on Pranav's part. The way he slowly curses you into caring about things is through an unceasing and infectious spirit of inquiry, guaranteed whenever you enter his office, even if you are just coming to return a book. I'm proud to be, as far as I know, the first exception to this rule: having seen him turn so many others, I put myself directly in the path of that curse, and have greatly enjoyed the ensuing ride. I know I will never rival his encyclopedic command of most every relevant literature, or his inexhaustible stamina for overflowing advising meetings. But I will still think of him whenever I take a second to run an example on a chalkboard, and whenever I remember to look at a few hundred naturally-occurring examples at the outset of a project. His support, in short, has made me a much better linguist, and for that I am endlessly grateful.

If my meetings with Pranav tended to run over, my meetings with Matt Wagers have been concise and exact, as a rule. This efficiency was always valuable, because I generally found myself walking away with three or four already-crystallized ideas that I want to set in motion immediately. In this way, his meetings have the same density and potency as his classes and seminars, which are responsible for most of the experiments in my "to-run" stack, and a large deal of my own goals for teaching, not to mention all of my preferred slideshow aesthetics. It has been a pleasure to look up to and collaborate with Matt over the past years, and I hope I'm lucky enough to continue working with him in the future.

Jesse Harris joined this committee much later than the local members, but has had a tremendous impact on the dissertation as it is written, providing a very necessary role as a trusted outsider. Over a few conversations, he helped me clarify greatly how this work fits into its greater context in the field, and raised many questions that I may never be able to fully get to the bottom of. I'm very grateful to do whatever work I can on those questions knowing I am in good company. Having his name on this title page is a pleasing synchronicity, since I am positive that his own 2012 dissertation was the first dissertation in linguistics that I ever laid eyes on. As much as any work, his willingness there to bridge the gap between formal theories of meaning and rigorous hypotheses about sentence processing has been a guiding star for my research in the years that have followed. I'm very happy to have benefitted from his insights, and hope that I do them some justice in my following.

Santa Cruz students do not really have a bounded committee, in the same way as we never have a single advisor. As such, I am lucky to have received mentorship and guidance from across the department over the past five and a half years. Adrian Brasoveanu's contributions in that regard have been invaluable, serving as an early and ceaseless cheerleader, a co-conspirator for the work that became the seed for this dissertation, and another of the many voices that has pushed me to seek connections and insight beyond the edges of our discipline. Although they have not been part of this particular work, Maziar Toosarvandani and Ivy Sichel have similarly had a remarkable influence on the way I approach research and teaching, always encouraging me to understand more completely the contours of any set of linguistic data and what we can learn from it. The work I have done with them, Matt, and the other members of r/lab and the Zapotec Project has been a joy, and I look forward to getting to continue it. In working on meaning, I have been grateful also for the questions, advice, and insight of Donka Farkas, Jess Law, Roumi Pancheva, and Sandy Chung, who each in their way have taught me a piece of what it means to look at language like a Santa Cruz semanticist (even if no single person can claim to encompass such an identity). On the opposite side of the gradient of abstraction, early courses and seminars in phonology from Ryan Bennett and Junko Ito have had far more influence than I would have expected, and I cherish the lessons they taught me about arguing for a proposal in a formal system. Similar thanks are due to all the other members of this department, each of whom has had a thumbprint on my work, and each of whom I wish I had longer and more frequent conversations with: Armin Mester, Grant McGuire, Haoze Li, Jim McCloskey, Jorge Hankamer, Judith Aissen, Mia Gong, and Rachel Walker. And in the tireless support of such a wonderful place to be a linguist, and the sometimes uncooperative linguists who are so happy there, I have to express gratitude to the members of the departmental staff, past and present, Ashley Hardisty, Gwyn Vandevere, James Funk, Logan Roberts, Maria Zimmer, and Sarah Amador.

But I would not have written the Santa Cruz dissertation that I have written had I not come first from a similarly remarkable community at UMass Amherst. One could not pick better roots. (I have to thank early and compelling advice from Nathan Sanders and Kyle Johnson for essentially picking these ones for me.) If some strange reader happened to tackle this document from back to front, it will come as no surprise that Lyn Frazier and Chuck Clifton were early and incredibly important mentors. Their generosity and wisdom has colored every step I've made since. To Lyn, in particular, I owe all of my favorite research topics, a pressure to always create testable and specific hypotheses and, (I admit) a penchant for wearing striped shirts to conferences. I can only hope to do these things some justice, and, if possible, pass them on to others some day. I also benefited from early guidance and mentorship from Alice Harris and John Kingston, some of the first times I had the privilege of getting my hands dirty with linguistic data. I'm grateful to have been able to do so in such a variety of domains so early. I also thank Shayne Sloggett and Caroline Andrews for always being willing to lend an ear and some advice to an eager undergraduate.

Others from outside of Santa Cruz and Amherst have provided support and feedback along the way. Chapter 4 had its roots in a qualifying exam and promissory note written in spring 2022, for which Hannah Rohde served as a valuable external member, pointing me in the right direction to sharpen my initial thoughts on explanatory relative clauses, in ways that I'm still looking forward to building on in the future. The same chapter can trace some of its core questions to a great ESSLLI class on discourse coherence from Daniel Altshuler and Julian Schlöder. An intervening but distantly related line of research on the notions of causality involved in building complex event descriptions helped drive me back to being interested on causality in event comprehension, for which work I also have to thank my co-authors Carolyn Andrews, Felipe Lopez, and Fe Silva-Robles (see below), and Bridget Copley for her warm welcome into the COCOA workshop. A broader thank you to many fellow-travelers at the intersection of sentence processing and formal semantics and pragmatics for conversations and inspiration, including chiefly Elsi Kaiser and Judith Degen for their particular involvement and encouragement in the CAMP community, but also Alex Göbel, Andrea Beltrama, Chao Sun, Noa Attali, Sarah Hye-yeon Lee, and Sebastian Schuster, among others. Dean McHugh may not be in that intersection, but he's good company, and I'm always happy to have an excuse to talk to him, so he's entirely unbefitting of a label of "normal semanticist". Mandy Cartner is not in that intersection either, but indeed I could never call her a "normal psycholinguist". She has a discerning eye for a well-made argument, and my work has often benefited from a conversation with her.

I'm very excited to be joining Vera Demberg and her lab in Saarland University. Many thanks to them for welcoming me so warmly, having all of the right questions, and giving me an excellent reason to finish things up this year.

On the other hand, the hardest part of finishing this degree has been having to say goodbye to an invigorating community of graduate students at Santa Cruz. None more invigorating than my cohort, Max Kaplan, Morwenna Hoeks, and Stephanie Rich, to whom I owe my sanity and too much joy to summarize here (it would indeed be like attempting to summarize an ocean for men). Who studying in Santa Cruz could ask for better people to do it alongside? We entered a lively and welcoming department, thanks to, among others, Kelsey Kraus and Deniz Rudin; although they finished their degrees before I entered, as the oldest of my academic siblings, I am deeply aware of the ways in which they cleared a path for me (and continued to be excellent company from the world beyond). Among those that began their time here before me, I have to also thank Jed Pizarro-Guevara, for his immediate and enthusiastic friendship; Margaret Kroll, for exemplifying how to do careful experiments on difficult phenomena; Steven Foley, for outstanding taste in all things (theoretical questions, wordplay...); Jake Vincent, for always being there in a pinch with practical advice and a level head; Tom Roberts, for being the Jadwa to my Halax Arn, the Gladys to my Big Chuck, and another raccoon co-piloting this car; Kelsey Sasaki, for not only laying the foundations for exactly the kind of work I wanted to do in the department, and not only also carving the path for engagement in meaningful partnerships with Senderos, but also also being a generous friend; Nick Van Handel and Netta Ben-Meir, for a warm welcome into the psycholinguistics corner of the Cave and much subsequent camaraderie; Lisa Hofmann, for her inspiring ability to deliver a complex point concisely and convincingly, in the conference hall and beyond; Andrew Angeles, for his never-ending good cheer; Andrew Hedding, for his bike and his willingness to share a genuine and thoughtful appreciation for life; Ben Eischens, for being such an excellent friend that our remarkably overlapping taste in 2000's-2010's indie folk barely comes up amidst all the other things to talk about; Jérémie Beauchamp, for his infectious dancing and humble presidency over the karaoke mic; Richard Bibbs, for his willingness to share a genuine and thoughtful appreciation for quality board games and food; and Allison Nguyen, for her evangelism regarding Star Trek, but also her enthusiasm for what people do with language. Over the years, that set of Thesean planks have been replaced with an equally fantastic set, among whom I have to thank Lalitha Balachandran, for her willingness to pursue questions about the proper nature of discourse and its interfaces with sentence processing to a sometimes COVID-inducing level of exhaustion; Maya Wax Cavallaro and Myke Brinkerhoff, for being such wonderful and patient collaborators that

I sometimes forgot I wasn't a tonal phonologist; Vishal Arvindam for being a superb keystone of the important Santa Cruz linguistics disco scene; Yaqing Cao, for always having the best question after a semantics talk; Dan Brodkin, for his boundless enthusiasm and serious knack for compelling theory-building; Niko Webster, for his careful attention to both the syntax of argument structure and the health of the graduate student body; Claire Miller Willahan, for her partnership in a curious summer dive into relative clauses; Delaney Gomez-Jackson, for her partnership in a curious winter dive into degree constructions, not to mention shared appreciation of pan de Laxopa and Electrolit; Matthew Kogan, for his impeccable trivia; Eli Sharf, for enthusiastic conversations spanning all topics en route to conferences, field trips, and concerts; Elifnur Ulusoy, for being a dedicated member of my first teaching team in a chaotic quarter; and David Tuffs, for being my final co-TA. A more general thank you is owed to everyone who I've been able to cross paths with during my time here, including Erik Zyman, Jason Ostrove, Nick Kalivoda, Hitomi Hirayama, Jenny Bellik, Maho Morimoto, Taijing Xiao, Jonathan Paramore, Jun Tamura, Sophia Stremel, Duygu Demiray, Ian Karpik, Larry Lyu, Richard Wang, Yagmur Kiper, Aidan Katson, Emily Knick, Hanyoung Byun, and Ruoqing Yao. Thanks also to the non-linguists: Anjelica, Dori, Kara, Nick, Rachel, Sienna, your friendship and willingness to endure much talk of adjuncts and lambdas was greatly appreciated.

I've also been grateful to have help from a small handful of undergraduate and high school students during my time at Santa Cruz, in particular, thank you to undergraduates Jennifer Wheelock, Dasha Komissarchik, and David Tuffs (again!) for their help collecting a corpus of naturalistic subjective conversations in the early pandemic; Science Internship Program interns Robert Liu, Stephane Palomares, Vindhya Shigehalli, Jacqueline Wen, and Naomi Wong for helping annotate and explore that corpus; and undergraduates Sebastian Bissiri, Pamela Casipe, Jason Kushner, Kasey La, and Kavi Mookherjee Amodt for their more recent help in the eyetracking lab—particularly Sebastian and Kasey, who assisted fearlessly in the preparation of endless Maze foils for some of the experiments in Chapters 3 and 4.

One of the highlights of my time in Santa Cruz was getting to meet and enjoy getting to know Maestra Fe Silva-Robles and her family. I have learned so much from Fe, and that time has been a tremendous gift. I'm grateful to her, Raul Diaz Robles, and many residents of Santiago Laxopa for the generosity of their time and interest in collaborating to work on their language. Another, sometimes-overlapping highlight has been the great privilege of playing music with others. Thank you to all the musicians of the Senderos musical ensemble, including especially Raul, Kristen, Chanel, David, and Beto, and also to the mighty Complications of Los Gatos: Angel, Brian, Chris, Gina, Jurgen, Jondell, Kia, Kincaid, Nathan, Ollie, Paul, Terry, and Tim. Relatedly, in lieu of releasing my entire writing playlist, for providing the ideal conditions for extracting this dissertation from my head I will give special thanks to the music of Mary Lattimore, The Field, Nils Frahm, and the Grateful Dead.

Over the years, I've been grateful to have much support from close friends and family. From Scituate, Gabe, James, Max, Mitchell, and Remy; from Amherst, Aiden, Case, Caleb, Emily, Ethan, Greg, Henry; and through much, Frank, Molly, and the other patrons of the Killenyarda; thanks to each of you and everyone else for the hours passed in good company. For foundational inspiration that it was a worthwhile thing to learn and to teach, I'm more grateful every year for Theresa Raymond and Richard Stower. And for making the bricks that preceded those foundations, I thank my family, especially Marguerite, Vincent, Eleanor, Lynne, Sylvia, Lois, Nick, Kathe, Janet, Bonnie, and of course Jim, Sue, and Lily.

"To be sure," said Canby; "you're on the Island of Conclusions. Make yourself at home. You're apt to be here for some time."

"But how did we get here?" asked Milo, who was still a bit puzzled by being there at all.

"You jumped, of course," explained Canby. "That's the way most everyone gets here. It's really quite simple: every time you decide something without having a good reason, you jump to Conclusions whether you like it or not. It's such an easy trip to make that I've been here hundreds of times."

…

"Well, I'm going to jump right back," announced the Humbug, who took two or three practice bends, leaped as far as he could, and landed in a heap two feet away.

"That won't do at all," scolded Canby, helping him to his feet. "You can never jump away from Conclusions. Getting back is not so easy. That's why we're so terribly crowded."

Norton Juster, *The Phantom Tollbooth*

Chapter 1

Introduction

Suppose you walk into a party to hear an acquaintance say the relatively prosaic utterance in (1).

(1) My mother donated an organ last year.

You probably have a very good idea of what your acquaintance means, even though it's a surprising fact: their mother must have been one of the roughly 6,000 Americans who made a living donation of a kidney or part of their liver in 2022.[1] But despite your likely certainty in this conclusion, perhaps you shouldn't be too hasty. After all, when the Juneau Empire wrote in a 2019 article that "Tim Fullam donated an organ," they were describing a very different donation, of a musical instrument.[2] Annual statistics on this kind of organ donation are harder to find, but it happens with some regularity. So, you have a decision on your hands: which type of *organ* does this acquaintance mean? Moreover, you have a problem of decision timing. Is it safe to make a decision about the meaning of their sentence now, or is it worth maintaining uncertainty and listening further? How long will you hold out before making up your mind?

This scenario, cartoonish though it may be, is familiar and frequent for the mental processes that achieve language comprehension. Linguistic input is ambiguous as a rule, especially as it is initially comprehended, without the benefit of hindsight from later information. Consequently, we have many interpretive decisions on our hands, and just as

[1] "2022 organ transplants again set annual records," United Network for Organ Sharing, 10 January 2023 (https://unos.org/news/2022-organ-transplants-again-set-annual-records/).

[2] "Donations fill church with joyful noise," 19 November 2019 (https://www.juneauempire.com/news/donations-fill-church-with-joyful-noise/).

many cases of this problem of decision timing. This dissertation is devoted to the question of how the comprehension mechanism handles this problem; in particular, the question of when and how we move from awareness of multiple potential interpretations, to something like a firm decision.

The chapters that follow aim to answer this question for a variety of different types of uncertainty in language comprehension, specifically in the reading of English. Formal theories of linguistic meaning, driven largely by intuitions about the structure and meaning of specimen sentences or discourses, identify many places where an observed piece of linguistic input might receive several different interpretations. I'll adopt the criterion in (2) to identify instances of uncertain meaning for the purposes of the present work. Notice that the relevant notion of meaning here references reasonable interpretations rather than necessary interpretations. The uncertainty under consideration will thus encompass both "semantic" uncertainty, cases where linguistic knowledge can map input to two different structures with distinct sets of truth conditions,[3] and "pragmatic" uncertainty, cases where reasoning about the goals and structure of communication may license an optional enriching inference beyond a fixed semantic meaning. Of course, this distinction is theory-internal; many cases argued to fall into the latter category can be modeled as cases in the former category featuring an optional implicit contentful element in the structure. It is an open question whether a division between the two is relevant for the process of comprehension, one which I will attempt to wrestle with here.

(2) Uncertainty Criterion

A string has an uncertain meaning if it contributes meaning M_1 when embedded in a context C_1, but meaning M_2 when embedded in a context C_2.

(A difference in meaning must mean a difference in the facts which a comprehender would reasonably conclude given the containing discourse, which difference is not attributable to merely the changes to the context itself.)

The specific cases of uncertainty which I will investigate under this umbrella are homonymy (cf. *organ*), polysemy, and ambiguities of distributivity for predicates with

[3] This way of carving up the world would include many well-known syntactic ambiguities, e.g. modifier attachment, as cases of semantic uncertainty, as they correspond to structures to which the semantics assigns distinct truth conditions. Nevertheless, my focus here will remain generally on semantic uncertainty that is not related to syntactic constituency decisions. Parenthetically: not all syntactic ambiguities have the above property; those which are semantically vacuous are presumed to be of limited importance to the comprehender.

plural subjects (all in §2), but also optional upper-bound meanings for *some* attributed to scalar implicature (§3) and optional causal inferences attributed to discourse coherence (§4). Each case has already been investigated substantially by researchers in sentence processing: for helpful exemplars with up-to-date perspectives on the literature see e.g. Brocher et al. (2016) on homonymy and polysemy, Dotlačil and Brasoveanu (2021) on distributivity ambiguities (the least investigated of this collection), Breheny (2019) on scalar implicature, and Hoek et al. (2021a) on causal inferences from coherence. However, these literatures have been mostly independent and idiosyncratic. As a consequence of investigating incremental comprehension of all of these constructions, I have observed patterns of similarity, and have begun to see the resolution of meaning uncertainty as a singular problem with a restricted typology of solutions. These studies were not planned with some global hypothesis in mind, but rather, I aim to demonstrate here how we might induce one, relating the reviewed literature and novel contributions from each chapter to the same core questions. The tentative theory which results generates testable predictions and raises substantial questions for the nature of the skilled human behavior that is language comprehension. I hope these predictions and questions will be pursued in future work.

1.1 Empirical and theoretical scene-setting

The prevailing view of decision timing in incremental comprehension against which my contributions are situated has approximately two main tenets, what we can call Rapid Incremental Decision-Making (3) and Deferred Decision (4).

(3) Rapid Incremental Decision-Making

Comprehenders analyze the linguistic signal incrementally, often making implicit decisions about the (likely) structure and meaning of a sentence before they have seen all its component parts.

(4) Deferred Decision

Comprehenders put off certain decisions about meaning for longer than others.

The first observes that comprehenders generally make decisions, or something like decisions, quite rapidly as they analyze partial input. This is a remarkably general observation, and foundational to the field of sentence processing, best studied in work on syntactic parsing (e.g. Frazier & Rayner, 1982; Stowe, 1986), but also shown for lexical selection (e.g. Duffy et al., 1988) and implicit ambiguities of semantic interpretation (e.g. Frazier et al., 1999). One classic piece of evidence for the existence of rapid analysis of partial input comes from the observation of "garden-path" effects, where comprehenders exhibit difficulty when late-arriving input cannot be given a coherent analysis based on a preferred analysis of an earlier ambiguity. If no analysis had been made for the early ambiguity, it would be hard to explain why this particular cost for dispreferred late disambiguation should arise, therefore psycholinguists generally take this to reflect some degree of analysis has occurred. Explanations of the classic syntactic garden-path effects are varied, and rely on different models of exactly what sort of analysis has occurred. Classic models of serial parsing (Frazier, 1978, *et sequens*) hold that the parser has rapidly selected a single analysis of the ambiguous input, and the observed cost for late disambiguation is a procedure of costly reanalysis, returning to and revising that decision. See also Frazier and Rayner (1982) for early evidence that this reanalysis occurs selectively, that is, comprehenders' targeted regressive eye movements during reanalysis suggest that they exploit some awareness of what decision they need to revise. The most popular class of alternative approaches (e.g. MacDonald et al., 1994) holds that comprehenders may consider multiple competing analyses of the input in parallel, in an interactive system sensitive to information from various sources (e.g. discourse context, lexical content of the ambiguous structure). Such models can attribute garden-path costs to costly re-ranking of potential analyses. Modern exemplars of parallel models have had success formalizing these re-ranking costs in terms of information theory (e.g. Hale, 2006; Levy, 2008). No matter whether a serial or parallel model is adopted, the field agrees that garden-path costs are diagnostic of incompatibility between the current input and an analysis that the comprehender has given some degree of credence.

Nevertheless, turning to (4), sentence processing has sometimes failed to observe garden-path effects where they would be expected given linguistic theory. For instance, despite evidence for a preferred and dispreferred sense for some polysemous nouns, late disambiguation to the dispreferred sense is not associated with any observable cognitive difficulty (Frazier & Rayner, 1990; Frisson & Pickering, 1999; Pickering & Frisson, 2001; McElree et al., 2006; Foraker & Murphy, 2012; Brocher et al., 2016, 2018). The typical conclusion is that specification among the possible senses of the polyseme has been entirely deferred in these cases (Frisson, 2009). Another influential line of work has suggested that decisions which are made incrementally in some places can sometimes be deferred when a determinate analysis may not be required (e.g. Swets et al., 2008, but see critical discussion in Logačev and Vasishth, 2016 and discussion in Chapter 5).

The major previous work to attempt a theory of when and why the comprehender will engage in Rapid Incremental Decision-Making vs. Deferred Decision is Frazier (1999). Building on ideas laid out in Frazier and Rayner (1990), Frazier depicts the comprehender as engaging in a problem of weighing the utility and risk of immediate input. On the one hand, analyzing partial input demonstrably opens the comprehender up to costs when they make an incorrect decision; the safest way to avoid these costs is to heuristically postpone all analysis until all relevant input has been observed. On the other hand, comprehenders are engaged in a difficult task of interdependent decisions, fighting against a limited capacity to represent input in memory without linguistic analysis.[4] She suggests that these pressures leave comprehenders with different strategies, determined by the nature of the decision they must make. For cases of uncertain analysis which are consequential for grammatical representation, i.e. where a grammatical representation of the input cannot be made without settling on one analysis, comprehenders generally make decisions immediately. Here, deferring interpretation risks substantial difficulty representing the input at all. But for all other comprehension decisions, i.e. where there is a grammatical representation possible that does not distinguish between the two analyses, the comprehender should generally defer decisions until further information is available.

1.2 The contributions of this dissertation

Against this backdrop, the present work has two main empirical contributions, obtained principally by looking for the hallmarks of rapid comprehension decisions across a variety of task environments and linguistic phenomena. In Chapter 2, I show that sense specification for polysemous nouns, the marquee example for deferred decision-making, is not deferred for comprehenders in a task which has been argued to encourage highly-incremental comprehension, the Maze task of Forster et al. (2009). Together with new evidence for task-dependent postponement of distributivity ambiguities, I relate this to existing proposals that the risk and utility of certain online decisions is dependent on the

[4] See Christiansen and Chater (2015) and its commentaries, especially those by Potter, Chacón et al. and S. C. Levinson, for discussion of this pressure and its sources in modern theories of memory and language processing.

task (Pickering et al., 2006; Logačev & Vasishth, 2016), the language (Cutler et al., 1986; O. Bott & Gattnar, 2015), and the cognitive resources available to the comprehender (Stine-Morrow et al., 2006). I argue that this evidence pushes us to extrapolate Frazier's (1999) functional approach to an active and flexible optimization over risk and utility in a given context.

In Chapters 3 and 4, I turn attention to scalar implicature and causal inferences from discourse coherence, two phenomena which (some) formal models treat as optional enrichment of a determinate linguistic representation. For both phenomena, work in sentence processing has found evidence that the purported enriched meanings are rapidly available during comprehension. Nevertheless, across five experiments, I present novel evidence that these enrichments do not provoke garden-path-like costs in cases where they must later be abandoned. I argue from this lack of garden paths and other aspects of behavior that comprehenders do not typically make firm decisions on enriched meanings during incremental comprehension, even in circumstances where they are anticipating such meanings.

Attempting to reckon with these observations leaves us with some questions about important features of our general theory of sentence processing, which I will engage with principally in the concluding Chapter 5. For one, the observations in Chapters 3 and 4 seem to require a dissociation between generating and considering an analysis of ambiguous input in general, and whatever degree of consideration drives garden-path effects. The model that I find intuitively provides the best fit for this two-stage process is one where comprehenders at one stage generate and entertain multiple analyses of input in parallel, and may exploit their confidence in an analysis to drive expectations for later input, before in a second stage, comprehenders finally select one analysis and inhibit the others.[5] In contrast, it seems difficult, if not impossible, to model these distinctions within a fully parallel, gradient-activation framework. One would need to suppose some meaningful difference between the initial, strictly facilitatory effects of consideration, and a later stage where consideration has progressed to a point where it entails revision costs. Certainly more evidence and a more fully-specified model would be necessary to make fur-

[5] To be specific, I will use "selection" to refer to the process of arriving at just one interpretation from many competing interpretations. To refer to the output of that process I might say that one analysis has been "selected" or "specified," or that the comprehender has "committed" to that analysis. This will generally be contrasted with "consideration", by which I mean a process where a potential analysis is generated, explored, and perhaps given some non-exclusive credence.

ther progress here, but the argument that these two stages must be distinguished serves as a surprising challenge against a fully parallel comprehension mechanism.

The second question of larger significance then concerns the relative timing of the latter stage, which I will refer to as selection, across the domains surveyed here. Looking across the several case studies, I note that the timing of selection is the only important parameter we need to vary in order to capture the diversity of behavior. On the one hand, the finding that selection is delayed for scalar implicature and causal inference from discourse coherence falls nicely in line with Frazier's division into grammatical ambiguity vs. post-grammatical meaning uncertainty, at least so long as one follows a post-grammatical account of the latter two phenomena. But in light of the contributions of Chapter 2, we seem to want a more flexible typology than the original two-way split. Across the dissertation, I have evidence for strictly grammatical decisions which always obligate immediate selection (homonymy) and more complex but still strictly grammatical decisions which are at least sometimes flexibly postponed (distributivity), plus post-grammatical decisions which are flexibly postponed until the sentence boundary (polysemy), and more complex post-grammatical decisions which are uniformly postponed indefinitely (scalar implicature and causal inferences from discourse coherence). I will imagine that we could explain this exploded typology as a spectrum based on the functional risk and utility of each type of decision given the goals of comprehension. Although I do not yet here advance a testable version of this net-value calculus, I argue that in order to capture observed patterns of variation, it must be sensitive to at least (a) the task-specific utility of fine-grained expectations about upcoming input, (b) the resources required for selection vs. continued deferment, and (c) the quality of evidence available for the decision so far.

If a plausible overarching model of this possibility space can be developed in this manner, it provides a path of insight into a larger research program, to what extent adult sentence processing behaviors can be modeled as acquired, optimal execution of a particular computational problem.[6] Sentence processing frequently forms generalizations about the behavior of adult comprehenders, but more rarely imagines where those "native comprehension" behaviors come from—more serious consideration here fueled by a specific theoretical mechanism for investigation would be very valuable.

[6] I should be clear here: when I turn to this domain-general way of understanding language processing behavior, I nevertheless remain convinced that the inputs and outputs of this computational problem are determined by concrete components of linguistic competence of the kind that are theorized in theoretical linguistics.

1.3 A few words on methods

The novel data which will motivate the discussion to come will come from the observation of participant reading behavior. The standard linking hypothesis for the analysis of such data is that the way readers allocate their time and effort to various portions of a sentence or discourse is sensitive to the difficulty of the ongoing process of comprehension. The gold standard for measuring this difficulty is comparing the natural behavior of readers across different passages, using a task like eyetracking while reading, where participants' eye movements are recorded during self-directed reading. Natural reading behavior is complex, and examination of eye movement records across many trials has yielded valid measures of several different kinds of difficulty. For instance, as readers struggle with initial recognition of a word, researchers generally observe prolonged durations of the first fixation made on that word; whereas as readers struggle to integrate a word with preceding context, researchers generally observe increased probability of a regressive eye movement to earlier in the discourse, and more time spent there before moving onward. Although there can be some vagueness of which measures ought to reflect certain hypothetical difficulty, eye movement data is generally supported by relatively clear linking hypotheses, and has the appreciable quality of leaving participants to read somewhat like they would normally.

Nevertheless, eyetracking experiments can be intensive, requiring single-participant, in-person data collection with careful supervision, and specialized equipment. Self-paced reading (SPR), where comprehenders move themselves forward through a sentence in chunks by pressing a button, has stood as a historical alternative for more flexible examination of reading behavior; these experiments can be carried out remotely, without special equipment or one-on-one supervision, and have been a popular method for psycholinguists recruiting participants on crowd-sourcing platforms like Mechanical Turk or Prolific. Here, the only measure of difficulty available is the latency of button presses, and analysis relies on the assumption that these latencies correlate with the difficulty of the comprehension processes of interest at that point in the sentence; that participants will be slower to press the button and move on when they are engaged in more difficulty processing, just as eye movements slow down. Reading in an SPR experiment is unarguably less natural than in an eyetracking experiment—for one, readers are not free to re-examine earlier stages of the text as they are used to doing.[7] But perhaps more troublingly, SPR has been critiqued for encouraging strategies which differ from typical reading, most notably the option of moving through many portions of a sentence at speed before pausing to attempt post-hoc integration, sometimes called "buffering" (Witzel et al., 2012). This leads to a common observation that behavioral effects observed for a region of interest in eyetracking may be spread over multiple following spillover regions in an SPR study. Perhaps for all of these reasons, labs anecdotally often observe that SPR studies often yield data with high proportions of measurement noise, often exhibit striking effects of participant fatigue, and sometimes struggle to replicate well-attested effects at all.

The studies I report here take seriously the idea that comprehenders adjust their reading behavior strategically to meet the demands and obstacles of a given task. From this perspective, although it is plagued by high degrees of measurement noise, SPR can offer an interesting barometer for what participants will do in one particular, somewhat unnatural, comprehension environment. They will serve as a key comparison point here for studies using another unquestionably unnatural method, the Maze task of Forster et al. (2009). Like (word-by-word) SPR, participants in the Maze move through a sentence with word-by-word button presses, but in the Maze, these come from a binary forced-choice decision between two visually-presented alternative words. Target continuations are paired with somehow-implausible foils, and the latency at a given word is taken to reflect the time taken to examine both choices, select the target, and integrate it before moving on. As will be discussed at more length in Chapter 2, Maze latencies, despite the unnaturalness of the added decision task, have generally been observed to correlate more closely than SPR with eyetracking measures, and are in particular less prone to spillover effects (Witzel et al., 2012). Maze studies are currently being adopted more widely, but I will demonstrate in Chapter 2 that they too, maybe unsurprisingly, promote atypical processing strategies, which I connect with the high utility of a fully specified partial analysis for optimal performance on the decision task. I thus carry a comparison between SPR and Maze experiments throughout the dissertation as a way of examining the ways readers are sometimes willing to change key elements of their approach to incremental comprehension. It is a regret that there are no eyetracking studies contained here, to provide a third comparison and

[7] In typical implementation, that is. See Paape and Vasishth (2021) for a recent exception, though self-paced regression remains an unnatural alternative to readers' typical regressive sweeps.

critical benchmark for behavior in a more natural task, but I hope the existing comparison highlights areas where such follow-up work would be most informative.

1.4 A roadmap

The dissertation progresses from here in a series of three case studies. Chapter 2 begins closest to the work that has already been done on the timing of commitment, with a series of SPR and Maze studies on the selection of lexical meaning and compositional semantic meaning. Exploiting comparisons across the tasks, I demonstrate that the delayed selection of a sense for a polysemous noun first observed by Frazier and Rayner (1990) can be replaced by immediate interpretation driving unexpected garden-path effects (Experiments 1–2). A follow-up on the processing of distributivity ambiguities shows that here too the Maze encourages early selection of a single analysis (Experiment 4), as compared to a failure to replicate the garden-path effects of Frazier et al. (1999) in SPR (Experiment 3). Discussion here explores how and why comprehenders' likelihood of immediate selection might be sensitive to the environment.

Chapter 3 brings questions of selection timing to the processing of *some*, as an exemplar of scalar implicature, where garden-path effects have been surprisingly absent in the existing literature. I argue that the available generalizations about the status of enriched meaning for *some* over time are best explained by a hypothesis where comprehenders are aware of multiple meanings, and sensitive to their likelihood in context, but do not select one or the other in typical reading. (Note that this depends on an architecture where more precise expectations can facilitate likely continuations without penalizing unlikely continuations; I will demonstrate one way this could be achieved.) Consistent with this hypothesis, and contrary to the predictions of an alternative hypothesis where enrichment happens decisively during reading, I look for but do not find garden-path effects in SPR (Experiment 5) or in the Maze (Experiment 6).

Chapter 4 raises a similar question for the status of explanatory causal inferences driven by discourse coherence. While online effects driven by explanation inferences are well-attested, they appear to fit the pattern of consideration without selection advanced in the previous chapter. Replicating evidence that these inferences are considered rapidly in one Maze experiment (Experiment 7), I nevertheless find again a surprising absence of garden-path effects there, and across further SPR and Maze experiments examining the reading of longer discourses (Experiments 8–9). I conclude this to be a final example of prolonged uncertainty, although I discuss how that conclusion depends on certain questions about the nature of the incremental interpretation of even explicit causal statements.

In Chapter 5, I present concluding discussions, focusing on the apparent flexibility of selection timing, and the existence of partial analysis effects preceding selection. I then bring into focus a few other relevant cases where uncertain meaning has been studied (ambiguities of quantifier scope, relative temporal order, verbal aspect, modifier attachment, and discourse anaphora), and show that they fall neatly within the same typology. This motivates various desiderata for a theory of rational decision timing during incremental comprehension that could capture the spectrum of behavior we observe. I discuss the nature of this theory, how it compares to more familiar two-stage underspecification models, and the ways in which I do and do not see it as related to grammatical representations of language, before concluding by highlighting avenues for future inquiry.

Chapter 2

Strategic variation in incremental semantic processing

Human comprehenders analyze the linguistic signal incrementally, often making implicit decisions about the (likely) structure and meaning of a sentence before they have seen all its component parts.[1] This is a remarkably general observation, true not only for syntactic parsing (e.g. Frazier & Rayner, 1982; Stowe, 1986), but also for lexical selection (e.g. Duffy et al., 1988) and implicit ambiguities of semantic interpretation (e.g. Frazier et al., 1999). And yet, incremental interim decision-making is not universal across structures: comprehenders seem to delay certain decisions about structure or meaning until the end of the sentence (e.g., Frazier & Rayner, 1990). One approach to these instances of decisions that are made far after the stimulus is observed, following Frazier (1999), is to take them as the processor's default, unmarked behavior. On this approach, decision-making in comprehension is governed by a heuristic of MINIMAL EFFORT, which favors deferred commitments, and which is violated only when constraints from the grammar force an earlier decision for the sake of representing structure.

This minimal effort hypothesis can explain many of the patterns we see in the sentence processing literature. On the one hand, there is evidence that, in order to build a representation of the structure that they are parsing, comprehenders make mid-sentence decisions on the meaning of homonymous nouns (Duffy et al., 1988; Rayner & Frazier, 1989) and verbs (Pickering & Frisson, 2001), for certain ambiguities of constituency (Frazier

[1] A version of this chapter has been prepared for journal publication as Duff et al. (2023). All materials, data, and analysis scripts are available for review in this OSF repository.

& Rayner, 1982), for the location of a gap for a displaced constituent (Stowe, 1986), for the relationship between a verb and members of its plural subject (Frazier et al., 1999; Dotlačil & Brasoveanu, 2021), and for the internal aspect of a predicate (Piñango et al., 1999, 2006; Todorova et al., 2000). The classic evidence for these effects is the observation of difficulty for the comprehender when disambiguation occurs downstream of a locally ambiguous target, often referred to as a "garden-path" effect. Garden-path difficulty is thought to come from the cost of reanalysis, that is, the cost of ruling out or disfavoring a structure that had initially been chosen when the target was first encountered. Following Frazier (1999), all of these patterns of early commitment to one analysis are cases where grammatical necessity—such as the need to determine the lexical category of a word in order to build the appropriate constituents over it—overrides minimal effort: the parser dictates that a choice must be made immediately wherever there is a choice point with consequences for the grammatical representation of the sentence.

On the other hand, there is evidence that, for polysemous lexical items—those hypothesized to have a singular lexical entry but multiple senses— comprehenders are free to delay selection of a sense until the end of the sentence, and are thus not prone to garden-path effects (Frazier & Rayner, 1990; Frisson & Pickering, 1999; Pickering & Frisson, 2001; McElree et al., 2006; Foraker & Murphy, 2012; Brocher et al., 2016, 2018). Because distinctions between senses are, by hypothesis, extra-grammatical (Frisson, 2009), these are the cases for Frazier (1999) where comprehenders fall back on minimal effort heuristics.

Of course, such divisions do not prove to be this clean upon further investigation. Evidence for the presence of incremental decisions on, e.g., a predicate's internal aspect is highly dependent on the task employed. While lexical decision times (Piñango et al., 1999) and word-by-word sensicality judgments (Todorova et al., 2000) diagnose reanalysis costs for unexpected disambiguations after a verb, attempts to measure this aspectual garden-path effect in self-paced reading and eyetracking have sometimes found less success (Pickering et al., 2006; O. Bott, 2010; though cf. Brennan and Pylkkänen, 2008; Townsend, 2013). Pickering et al. (2006) suggested, following Pylkkänen and McElree (2006), that distinctions of internal aspect for a given predicate are strictly speaking extra-grammatical in the same way as sense distinctions in polysemy, and can thus be postponed. On their account, early decisions and reanalysis costs still obtain for extra-grammatically represented sense distinctions in some tasks only because participants are violating usual minimal effort heuristics in those tasks to suit goals that are not present in "normal reading."

Task effects like the one hypothesized by Pickering et al. (2006) are a growing area of interest in sentence processing. They offer a way to examine possible variation in the way humans can deploy mechanisms for comprehension. For instance, Logačev and Vasishth (2016) suggested that comprehenders, when faced with particular comprehension questions, will maintain multiple syntactic parses for a sentence that usually receives only a single parse. But research has also shown that task demands may have a limited footprint in online processing: in an eyetracking investigation, Weiss et al. (2018) found that difficult comprehension questions could increase the probability of late re-reading, but had little effect on earlier reading measures. It remains a largely open question how much the difficulty of a task can affect low-level comprehension procedures in the way suggested by Pickering et al. (2006).

In this chapter, I present four experiments that examine the influence of task on the timecourse of decision-making for two types of comprehension decisions. Examining the marquee case for minimal effort, polysemous nouns, I replicate delayed commitment patterns in a less demanding task, but find that comprehenders in a difficult task will select a particular sense uncharacteristically early, and fall prone to garden-path effects as a result. I take this to be the first case of a task effect modulating behavior at the level of lexical selection. I then examine the interpretation of verbs with plural subjects, finding both more variability than previous studies and again that comprehenders in a difficult task make earlier commitments than in an easier task. On the whole, the results reported here support the idea that the time course of commitment in comprehension is the outcome of situation-specific strategic decision making. I argue that this can be understood as compatible with an alternative effort calculus, under which decision timing preferences are context- and construction-specific and derived from a derived optimization of cost and benefit within a given environment.

2.1 The Maze task

The main tool I use to examine task effects is the Maze task (Forster et al., 2009). In the Maze task, participants are asked at each word to make a forced choice between a word that makes a sensical continuation of the sentence and a foil, which does not. Participants thus advance through target sentences never seeing more than one word at a time, much like in word-by-word versions of self-paced reading or incremental sensicality judgments (‘Stops-Making-Sense’ tasks, Boland et al., 1995). Foils, the incorrect choices which would not sensically continue the sentence, are generated by the experimenter for each position, and can include non-words (an “L-maze”), hand-picked ungrammatical continuations (a “G-maze”), or in the implementation of Boyce et al. (2020), high-surprisal continuations generated by a language model (an “A-maze”). A-maze foils are typically ungrammatical, or else have meanings which are far less compatible with context than their respective targets. When a participant chooses a foil, they receive feedback informing them of their incorrect choice, and the trial concludes.

Previous research has found that the differences in response latencies observed in the Maze largely correlate with the differences in response latencies observed in self-paced reading or differences in reading time measures observed in eye movements. For instance, Maze studies have observed expected slowdowns in response latencies both when parsing object relative clauses and when reading infrequent words (Forster et al., 2009), and when reanalyzing relative clause and adverb attachment (Witzel et al., 2012; Boyce et al., 2020). In one early use, the task provided evidence for agreement attraction illusions when parsing ungrammatical plural agreement in the presence of a plural attractor (Nicol et al., 1997).

Forster et al. (2009), introducing the Maze, highlighted its intuitive potential to require “full structural commitment… at each point in the sentence” (p. 164). For the G-maze (ungrammatical foils) and A-maze (high-surprisal foils), selection of each correct continuation is contingent on the structure that has been built for the sum of all previous decisions. For this reason, not only are participants obligated to remember the previous choices they made, they are also encouraged to maximally interpret the string resulting from those choices, since without a determinate representation of a partial structure, it would be much harder to identify whether and how candidate continuations might be added.

There is indeed some evidence that participants construct determinate interpretations at each position in a manner different from participants in other kinds of reading experiments. For instance, Witzel et al. (2012) found that effects that are observable only later in a stimulus sentence in self-paced reading were found directly on critical regions in the Maze. These delays in self-paced reading have been thought to diagnose a measurement problem with that task, whereby participant behavior disobeys the linking hypothesis that button presses should be tightly locked to interpretation at each position. It seems clear that the Maze is not as prone to these kind of artifactual displacements.

Similarly, Forster et al. (2009) found small differences from previous results in a Maze replication of an eyetracking study on homonym disambiguation by Dopkins et al. (1992). In eyetracking, Dopkins et al. (1992) originally found costs of reanalysis downstream from late disambiguating material, suggesting that participants recognized the need for reanalysis only after eye movements continued past the critical material. In contrast, in the Maze replication, Forster et al. (2009) found the effect with no delay, such that participants were apparently engaging in reanalysis as soon as they encountered disambiguating input. This would suggest that the motivation that the Maze provides for incremental interpretation not only obviates undesirable experimental artifacts, it may even surpass the typical pressures active in normal reading.

In the wake of these studies, and particularly with the advent of the language-model tools designed by Boyce et al. (2020), the Maze seems to be a convenient alternative to eyetracking, with greater fidelity to the incremental time course of processing than self-paced reading, and which can be run during pandemic isolation conditions. In this paper I join previous studies in demonstrating its power to find clear effects of reanalysis. But I argue that the Maze task's pressures for incremental interpretation can force commitment and reanalysis far in advance of where it is expected in natural reading, providing suggestive evidence for Pickering et al.'s (2006) argument and perhaps providing a cautionary illustration for those who want to rely on any one task as a measure of typical sentence processing behavior.

2.2 Lexical meaning and underspecification

Our first area of investigation is the selection of lexical meaning. Since Frazier and Rayner (1990), the sentence processing literature has made a crucial divide between how participants comprehend (i) homonyms versus (ii) polysemes. Homonyms are distinct lexical items which share a phonological representation by (synchronic) accident, for instance *jam*, as in the fruit preserve, and *jam*, as in the traffic obstruction. In contrast, polysemes are typically thought to involve only a single lexical entry, but can be used with multiple related senses, for instance *newspaper*, as in the printed object, and *newspaper*, as in the corporate entity or cultural institution that can, e.g., endorse a political candidate. Polyseme senses can in some cases be analyzed as derived from a core 'literal' meaning through productive rules, including place-for-institution and place-for-event metonymy (Frisson & Pickering, 1999; Frisson, 2009). In other cases, as for *wire*, the metal material, and *wire*, the listening device, these senses might have a more idiosyncratic relationship (Brocher et al., 2016). Empirical evidence that the different senses of a polyseme share a core meaning, and so a single word representation, comes from priming effects. Different senses of a polyseme prime each other more than could be explained by similarities in meaning and form alone (Pylkkänen et al., 2006). Still, different senses are at least to some extent discrete, as priming within senses is stronger than across senses, modulated by degree of meaning overlap (Klein and Murphy, 2001; Klepousniotou, 2002; Klepousniotou et al., 2008; see also the extensive review in Brocher et al., 2016). Thus, for both homonyms and polysemes we might refer to 'dominant' and 'subordinate' meanings and senses, and determination of which sense is dominant versus subordinate is usually carried out by examining meaning and sense frequencies.

The strongest empirical generalization about the processing of homonyms and polysemes concerns their comprehension in sentences where they are preceded by a neutral context and then followed by disambiguation to a subordinate meaning. In an eye-tracking study, Frazier and Rayner (1990) compared the two types of words, and observed that while subordinate meaning disambiguations of homonyms later in a sentence were associated with increased reading times and regressions out of the disambiguating region, disambiguations of polysemes to a subordinate meaning later in a sentence showed no such costs. While the original results were mostly limited to regular polysemes, this asymmetry has proven robust, including for irregular polysemes (Brocher et al., 2016), even those without a dominant meaning (Brocher et al., 2018). Later work has clarified that the absence of reanalysis costs for polysemes only extends so far: disambiguation after a sentence boundary becomes costly much in the same way as homonyms (Frisson and Frazier, 2004 cited in Frisson, 2009; Foraker and Murphy, 2012).

Frazier and Rayner (1990) took this difference between homonyms and polysemes in later disambiguation costs to be evidence of a distinction in the commitments necessary during incremental processing. Polysemes do not require an immediate commitment to any particular sense, because to construct a determinate grammatical representation of the input, comprehenders need only identify a lexical entry—not determine the exact one of its many related senses that is required. One might say that, at this point during processing, the meaning of the polyseme has been momentarily 'underspecified', because no one sense has been selected (Frisson, 2009). Under this account, distinctions between senses are a type of meaning enrichment that is necessary only for complete sentence interpretation, and plausibly form part of the ‘wrap-up effects’ observed at sentence boundaries (Frazier, 1999; see e.g. Warren et al., 2009). This proposal parallels the summary in the introduction of this paper: comprehenders decide on a meaning for homonyms immediately upon encountering them, because they are compelled by the requirement to construct a determinate linguistic representation, while polyseme sense selection is delayed because it represents a type of extra-grammatical meaning refinement.

However, another empirical generalization complicates the picture somewhat. Another key measure for the time course of meaning selection has been the presence of what have been called ‘subordinate bias’ effects. Duffy et al. (1988) found that comprehenders exhibited difficulty in eye movements when they encountered a homonym in a context that was consistent with its subordinate meaning. That is, even when a context should have been supportive of a subordinate meaning, the reading of that word with that meaning was nonetheless relatively slow (see also Rayner and Frazier, 1989; Rayner et al., 1994; Folk and Morris, 1995; Binder and Rayner, 1998; Kambe et al., 2001). The slowdown has been argued to come from difficulties during a process of selection among all of a word’s possible meanings; key evidence for a fleeting stage of exhaustive activation even in biased contexts comes from cross-modal priming tasks (Swinney, 1979; Onifer and Swinney, 1981; see Morris, 2006 for review). Proposals differ in exactly how they model this difficult selection, but they generally agree that it is a consequence of a conflict between frequency and context, making the selection process somewhat more difficult than when both types of evidence point to the same meaning (see e.g. Duffy et al., 2001). Going forward, I will call this effect a ‘subordinate selection’ cost, under the expectation that we expect to observe it whenever the comprehender executes selection of a subordinate meaning.

Under the first model laid out above, in which polysemes are always underspecified until the end of a sentence, subordinate selection effects are not expected to occur during the reading of a polyseme itself, because no selection among its senses is required during initial processing. Indeed, some eyetracking investigations found no such effect (Frisson & Pickering, 1999; McElree et al., 2006), and Pickering and Frisson (2001) even found some evidence for subordinate selection costs at the ends of sentences in their investigation of verbal polysemy, which would support the claim that no selection occurs at the polyseme itself. But, still other studies find evidence that is not compatible with this proposal that underspecification is the initial, default response to encountering polysemes. Frazier and Rayner (1990) reported costly selection effects on pre-disambiguated polysemes, as did Lowder and Gordon (2013). L. Bott et al. (2016), investigating these subordinate selection costs in a speed-accuracy tradeoff study, found that the cost could not be attributed to delayed availability of a subordinate sense, but instead seemed to arise from a more difficult process of selection. This is compatible with models of selection for homonymy, if they were augmented to further hypothesize that the senses of a polyseme can be exhaustively accessed and will compete in the same way as homonym meanings given a specifying context. In the same vein, eyetracking and self-paced reading follow-ups by Brocher and colleagues (2016, 2018) have found that these selection costs are highest for polysemes with closely-related senses, which should also entail difficult selection due to high levels of mutual priming among the senses.

These studies on the cost of sense selection during the reading of polysemes in pre-disambiguating contexts thus suggest that comprehenders are sometimes willing to select a sense immediately upon encountering a polyseme, contra the original interpretation of Frazier and Rayner's differential findings for homonyms and polysemes. Still other findings suggest that sense selection may be initiated even by the presence of structural cues: in another eyetracking study, Fishbein and Harris (2014) showed that heuristic expectations about the features of subjects can fuel sense selection and lead to reanalysis. Polyseme specification thus does not seem to be *obligatorily* delayed; it may be delayed in the absence of evidence, but if comprehenders are given some indication of the appropriate meaning, it appears that they may be willing to commit early.

To revise the general approach to the distribution of effort during sentence processing that was sketched in the introduction, one might say that comprehenders violate simple heuristics of minimal effort in two scenarios: (1) when decisions are required in order for the parser to represent the input, which by hypothesis does not apply to polysemes, or (2) when otherwise-optional decisions can be settled based on the current evidence. At-issue for the current study is whether this list is now complete, or whether minimal effort in the processing of polysemy may be violated even in the absence of evidence for appropriate sense selection, such as if task demands alone encourage it.

In the first two experiments of the present paper, I investigated this question of whether early selection for polysemes can also be motivated by task demands. To do so, I studied the comprehension of sentences like those used by Frazier and Rayner (1990) first in a self-paced-reading task, and then in a Maze task. If lexical sense selection does not vary in the face of task demands, we would expect to find no evidence of difficulty at late subordinate disambiguating regions in sentences containing polysemes in both self-paced-reading and in the Maze. But, given the reasoning above about the task pressures introduced by the Maze, if task-specific demands can control the time course of sense selection, we would expect to find evidence of difficulty at late subordinate disambiguating regions in sentences containing polysemes in the Maze, diagnosing early sense selection. My results support the latter hypothesis.

2.3 Experiment 1

In my first experiment, I aimed to conduct a conceptual replication of Frazier and Rayner (1990) using fixed-window word-by-word self-paced reading, in order to set a baseline for further investigation in the Maze. I focused on two of Frazier and Rayner's conditions: those involving homonym targets in which both meanings were of the same degree of animacy and those involving polyseme targets. I predicted minimally that I would observe the same contrast that Frazier and Rayner did between these conditions when target words were presented in a neutral context and then followed by subordinate disambiguation. According to this prediction, items with polyseme targets should be associated with no particular cost when followed by material that disambiguated to a subordinate sense, because neither their grammatical representation nor the demands of the task require early commitment to a single sense, while the later subordinate disambiguation condition with homonyms should show signs of reanalysis, because their grammatical representations depend on specifying one meaning.

2.3.1 Method

2.3.1.1 Participants

48 native English speakers participated in the experiment online in early 2020. They were recruited from two pools: 24 from a pool of undergraduate students taking a linguistics class at an university in the United States, and 24 from the online experiment platform Prolific. Students were compensated with course credit, and Prolific participants were compensated according to a $12 hourly wage. Prolific participants were required to

have US nationality, at least the equivalent of a high school degree, and a minimum of 20 prior submissions with an acceptance rate of 90% on the platform. Student participants also predominantly were raised in the US, with three exceptions raised in Poland, India, and Hong Kong.

Initial analysis of these 48 participants and comparison to data from Experiment 2 revealed a higher degree of noise in this experiment, potentially due to the overall lower sensitivity of self-paced reading compared to the Maze (Forster et al., 2009). In an effort to offset this sensitivity difference and more easily compare behavior across the two tasks, another 48 participants were recruited from Prolific in early 2023, using the same criteria. Below, I report the results from the entire pool of 96 participants. Findings were largely unaffected by the increase in sample size.

2.3.1.2 Materials

The experiment featured 32 test items with a polyseme target and 32 test items with a homonym target. 16 items from each set were minimal variants of items from Frazier and Rayner (1990), and the remainder were constructed to the same template. All items contained an initial adverbial so that regions of interest were never sentence-initial. The target word always served as the subject of the sentence's matrix clause, and was always disambiguated by a modifying *after*- or *when*-clause. The position and the content of the disambiguator was manipulated across four within-item conditions in a 2×2 design, either coming **early**, before the target, or **late**, after the matrix predicate, and disambiguating towards either the **dominant** or the **subordinate** meaning of the target. Sample items containing a polyseme and a homonym are provided in Table 2.1, and a list of all targets used in Table 2.2.

Meaning dominance for all critical target words was assessed in a separate forced-choice relative judgment task with 32 participants recruited from Prolific according to the same restrictions described above. The procedure of this task was taken directly from the norming procedure used by Frazier and Rayner (1990). Participants saw pairs of sentences featuring the same target word but different disambiguations, and were asked to select which sentence "best expresses (or is most consistent with) the meaning" of the target word. The pairs always contained disambiguating regions in the same position for both sentences, but this position varied across items and participants, so that judgments were gathered for each sentence with early and late disambiguation.

Table 2.1: Sample items from Experiments 1 and 2.

	Polysemy	
Meaning	Early Disambiguation	Late Disambiguation
Dominant	Unfortunately, after it was soaked with rain the newspaper was destroyed.	Unfortunately, the newspaper was destroyed after it was soaked with rain.
Subordinate	Unfortunately, after it lost its advertising profits the newspaper was destroyed.	Unfortunately, the newspaper was destroyed after it lost its advertising profits.
	Homonymy	
Meaning	Early Disambiguation	Late Disambiguation
Dominant	Reportedly, after it made his toast soggy the jam displeased Tom.	Reportedly, the jam displeased Tom after it made his toast soggy.
Subordinate	Reportedly, after it doubled his morning commute the jam displeased Tom.	Reportedly, the jam displeased Tom after it doubled his morning commute.

Table 2.2: All targets used in Experiments 1 and 2.

			Polysemy			
newspaper	book	library	city	notice	dinner	article
play	firm	dollar	novel	poem	letter	message
pamphlet	lunch	phone	zoo	aquarium	museum	state house
laptop	school	embassy	store	bakery	court	university
palace	hospital	announcement	breakfast			
			Homonymy			
match	ring	fall	ball	jam	shade	cabinet
records	suit	tie	racket	change	bar	drive
deed	poker	tip	organ	spring	file	plant
bridge	tap	punch	deck	gas	port	chest
pipe	straw	pen	mold			

Results from this norming task indicated moderate preferences between the senses of the words in the majority of items: preferred meanings of homonyms were chosen on 74% of trials, while preferred meanings of polysemes were chosen on 71% (cf. 72% and 74%, respectively, in Frazier and Rayner, 1990). The strength of these preferences was evaluated by computing 95% confidence intervals using a non-parametric bootstrap around the proportion of responses in line with the overall preference. These confidence intervals excluded chance (50%) for 18 of the 32 polysemy targets and 21 of the 32 homonymy targets. Responses also indicated that these preferences are different than the preferences of the participants in the original studies: half of the polysemes repeated from Frazier and Rayner (1990) received opposite patterns of dominance in my sample (*library*, *notice*, *dinner*, *dollar*, *novel*, *message*, *pamphlet*, and *lunch*), as well as more than half of the repeated homonyms (*match*, *ring*, *fall*, *shade*, *cabinet*, *records*, *tie*, *racket*, *change*, and *deed*).

2.3.1.3 Procedure

The experiment was prepared in Ibex (Drummond, 2010), and deployed on IbexFarm and PCIbexFarm. Participants proceeded through all items using fixed-window, word-by-word self-paced reading, such that each press of their space bar displayed the next word in the center of the screen. Presentation was non-cumulative; these presentation choices were adopted so as to allow the most minimal comparison with the Maze task of Experiment 2.

Items were followed by binary forced-choice comprehension questions. For half of the test items, these were shallow questions that tested the participants' ability to recall surface-level aspects of the sentence. For instance, after the *newspaper* item in Table 2.1, participants were asked whether the newspaper was "destroyed" or "eaten". For the other half of the test items, these questions probed the interpretation of the target word in particular; for instance, after another polysemy item involving the target *dinner*, participants were asked whether the sentence referred to "an event" or "some food". These more detailed questions were adopted here to ensure that participants had a certain baseline motivation to completely comprehend the critical items.

Test items were presented in pseudo-randomized order across four Latin-squared forms, balanced within each of my two samples of participants. They were mixed with filler items from unrelated experiments, 60 featuring anaphoric and cataphoric dependencies across multi-clausal sentences, and 32 featuring quantifier scope ambiguities. To ensure these fillers served as suitable camouflage for the critical items, all featured similar sentence-initial adverbs, and the quantifier scope fillers featured similar *when* and *after* modifiers in various positions. A further 36 fillers with unambiguous senses were designed to exactly parallel the structure of the critical items. All fillers were also followed by comprehension questions. Four of the unambiguous fillers were presented to participants as practice items, and another 14 sampled from all filler sets were reserved as "burn-in" items, which were presented at the beginning of the main body of the experiment, but sequenced before participants were shown the first critical item.

After completing all 192 trials, participants completed short exit questionnaires on their experience in the study, and demographic information regarding their language history, before receiving compensation. The procedure in entirety was estimated to take about 40 minutes.

2.3.1.4 Regions and analysis

715 test trials in which a participant responded incorrectly to a comprehension question or a software error prevented accurate latency collection were excluded from analysis. The remaining sample includes data from 5429 critical trials.

Three measures of word-by-word response latencies were computed for the analysis of test items, aiming to observe the effects noted in Frazier and Rayner (1990). All measures relied on residual log response latencies, derived from a linear mixed-effects model fit using the `lme4` package in `R` (R Core Team, 2016; Bates et al., 2015) to log response latencies for words in all unexcluded trials, with fixed slopes for number of characters and position in the sentence, and random participant intercepts. Critical measures were (i) the summed residual log response latencies within the disambiguating temporal adjunct, (ii) the residual log response latencies on the target, (iii) the summed residual log response latencies on the matrix predicate following the target.[2] These regions are indicated for an

[2] Said another way, latencies were subjected to (natural) log-transform, the residualizing model was fit to those log latencies, and the residuals from that model were summed across each region. Note that final summarizing across the region was thus carried out in log-space. As a reviewer points out, this is meaningfully different from an alternative procedure where residuals were exponentiated back into real-space and summed in real-space before being subjected again to log-transform. I have not seen any suggestion for best practices in this case in the literature. On the one hand, the proposed alternative method of summarizing in real-space permits a more natural interpretation of the resulting measurement. On the other hand, it has been argued that it is mathematically preferable to perform summaries in log-space on (roughly) log-normal variables like response latency (e.g. the common advice in many other fields to prefer geometric means as a measure

Table 2.3: Regioning used for analysis of items with early and late disambiguation in Experiments 1 and 2. Note that the spillover region (the main predicate) and the disambiguating adjunct clause varied in length across items, but not conditions.

	Disambiguator		Target	Spillover
Unfortunately,	[after it lost its advertising profits]	the	[newspaper]	[was destroyed].

	Target	Spillover	Disambiguator
Unfortunately, the	[newspaper]	[was destroyed]	[after it lost its advertising profits].

Table 2.4: Model-fitting specifications used in brms. All other parameters (e.g. priors for random effects) received their default value.

Prior for intercept	$\mathcal{N}(0, 0.05)$	Chains	6
Prior for fixed effects	$\mathcal{N}(0, 1)$	Iterations	10000 (incl. 2000 warmup)

example stimulus in Table 2.3. Total log response latencies for the entire sentence were also analyzed, but did not show any effects of interest, and I will not discuss them further.

For analysis, Bayesian linear mixed-effects models were fitted to these three measures with Stan (Stan Development Team, 2019) using the brms package in R (Bürkner, 2017, 2018) with principled weakly-informative priors, maximal random effects structures, and sum-coded predictors. Subordinate meanings, late disambiguation, and homonym targets were coded as positive. Particular model-fitting specifications are summarized in Table 2.4. I take parameters whose 95% credible intervals (CRIs) does not contain 0 to indicate noteworthy effects. All models reported feature $\hat{R} = 1.00$ for the parameters of interest.

2.3.2 Results

All 96 participants retained for analysis answered comprehension questions with accuracy of greater than 80%. Other participants who had been recruited were excluded from analysis to ensure a base level of attentive comprehension. Mean accuracy in the final sample was 92.2%. In this section, I report the response latencies in the various regions of interest. Before doing so, I will highlight the tangible predictions made by a few different

of central tendency for data which comes from a hypothesized log-normal distribution, see discussion in Vogel, 2020). Because both options seem *a priori* reasonable, supplementary analyses were performed on the disambiguating region in Experiments 1 and 2 using this alternative dependent variable—the log sum of residual latencies. Models yielded comparable effects. In lieu of any strong argument for one procedure or the other, I retain the original method, here and elsewhere in the dissertation, of calculating summary measures for a region in log-space.

possible generalizations about the interpretation of polysemy and homonymy.

If comprehenders made an immediate commitment to a single interpretation of both polysemes and homonyms, we would expect two general effects. In neutral contexts, this would engender costly reanalysis in the case of late disambiguation to a subordinate meaning, which would surface as a consistent positive Position × Meaning interaction within the disambiguating adjunct for both target types, and no credibly non-zero three-way interaction. In subordinate-biased contexts, this would engender costly access on both kinds of targets, which would surface as a consistent negative Position × Meaning interaction in the target or spillover regions for both target types, and no credibly non-zero three-way interaction.

If the comprehender always delays incremental commitments for polysemy, while it always makes them rapidly for homonymy, target type should matter more. In the disambiguating adjunct, the positive Position × Meaning interaction diagnosing reanalysis cost should occur only for homonyms, driving a credibly positive three-way interaction. Likewise, on the target or subsequent spillover, the negative Position × Meaning interaction diagnosing costly subordinate selection should occur only for homonyms, driving a credibly negative three-way interaction.

I also discuss above a variant of this generalization, where the comprehender delays incremental commitments for polysemy only in neutral contexts. This would predict again that only homonyms will exhibit the positive Position × Meaning interaction diagnosing reanalysis cost in the disambiguating adjunct, and thus also expects a positive three-way interaction there. But because polysemes should receive a rapid interpretation in biasing contexts, this generalization predicts that both polysemes and homonyms will exhibit the negative Position × Meaning interaction diagnosing costly subordinate selection on the target/spillover, and thus no three-way interaction there.

Finally, if comprehenders postpone the interpretation of all polysemes and homonyms, we should expect none of these Position × Meaning interactions for any target in any region, as no distinctions between dominant or subordinate sense should be at all active. The predicted results under each of these generalizations are summarized in Table 2.5.

Table 2.5: Predictions for Experiments 1 and 2 regarding marginal Position × Meaning interactions and three-way Position × Meaning × Target interactions under four possible generalizations about the processing of polysemy and homonymy. Note that a marginal positive Position × Meaning interaction diagnoses particular cost in the Late, Subordinate condition, while a marginal negative Position × Meaning interaction diagnoses particular cost in the Early, Subordinate condition.

	Disambiguator			**Target/Spillover**		
Account	Pos × Mng		P × M × T	Pos × Mng		P × M × T
	Pol	*Hom*		*Pol*	*Hom*	
Commitment for both	+	+	0	−	−	0
Always delay polysemy	0	+	+	0	−	−
Delay polysemy w/o context	0	+	+	−	−	0
No commitment for either	0	0	0	0	0	0

2.3.2.1 Disambiguator

Summed residual log response latencies in the disambiguating region are presented in Figure 2.1 and Table 2.6. Posterior values for $\hat{\beta}$ and σ_β from the model along with 95% CRIs are provided for fixed parameters of interest in Table 2.7. A credibly negative intercept indicates that this region was generally read faster than would be predictable on the basis of word length and position alone, $\hat{\beta}$ = -0.15, 95% CRI = (-0.20, -0.10). I observe no main effects or interactions which are credibly non-zero, including the interaction between disambiguator position and meaning dominance, $\hat{\beta}$ = 0.02, 95% CRI = (-0.02, 0.06), and the predicted three-way interaction between disambiguator position, meaning dominance, and target type, $\hat{\beta}$ = -0.01, 95% CRI = (-0.05, 0.03).

Given the absence of a credibly non-zero three-way interaction, I extract marginal comparisons from the model to better understand the nature of the effects of late disambiguation across the different conditions. These comparisons reveal some evidence that late disambiguation of homonyms was associated with slower response latencies, $\hat{\delta}$ = 0.09, $P(\delta > 0)$ = 0.86. A corresponding difference was not found for the disambiguation of polysemes, $\hat{\delta}$ = -0.01, $P(\delta > 0)$ = 0.43. This is weak evidence for the presence of added difficulty in reading post-homonym disambiguators in general, consistent with a greater frequency of costly reanalysis after homonyms. Nevertheless, I see no particular evidence that this pattern is mediated as predicted by distinctions in meaning dominance for homonyms, where the marginal Position × Meaning interaction is estimated close to zero, $\hat{\delta}$ = 0.04, $P(\delta > 0)$ = 0.64. In fact, there is an unexpected

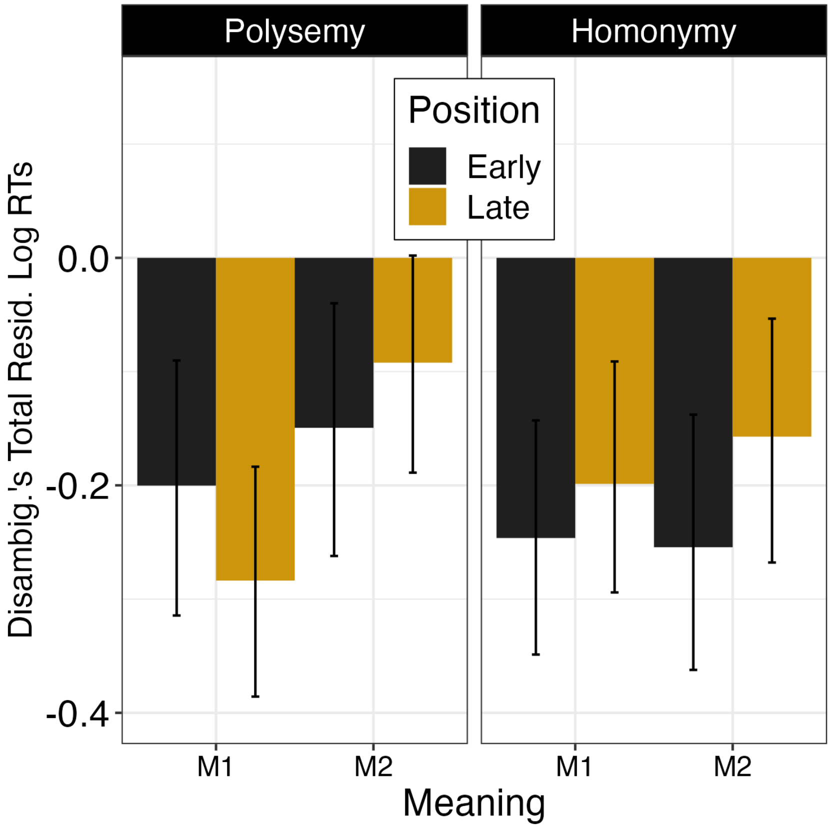

Figure 2.1: Summed residual log response latencies in the disambiguation region in Experiment 1, by condition. Error bars represent bootstrapped 95% confidence intervals around the mean.

trend towards a marginal Position × Meaning interaction for the polysemes, $\hat{\delta}$ = 0.13, $P(\delta > 0)$ = 0.87. Disambiguation to the dominant sense was weakly associated with unexpected faster reading times when late, $\hat{\delta}$ = -0.08, $P(\delta < 0)$ = 0.79, consistent with some baseline difficulty processing the early adjuncts, perhaps due to some cost associated with processing them in their non-canonical fronted position. In contrast, late disambiguation to the subordinate sense was weakly associated with slower reading times, $\hat{\delta}$ = 0.05, $P(\delta > 0)$ = 0.71, although this marginal difference was smaller than that observed for homonyms with late dominant disambiguation, $\hat{\delta}$ = 0.07, $P(\delta > 0)$ = 0.75 or late subordinate disambiguation, $\hat{\delta}$ = 0.11, $P(\delta > 0)$ = 0.87. Of the accounts considered in Table 2.5, this overall pattern of slight difficulty with late disambiguation unmediated by meaning dominance is most compatible with those which expect no incremental commitment to either polysemes or homonyms.

Table 2.6: Conditional means and measures of spread for the disambiguation region in Experiment 1. Standard errors are reported over the sum of raw response latencies in the region, and bootstrapped 95% confidence intervals are reported over the critical measure, summed residualized log response latencies.

Target	Meaning	Position	Sum RT	SE	Sum Resid. Log RT	95% CI
Polysemy	M1	Early	2482	54	-0.20	(-0.32, -0.09)
Polysemy	M1	Late	2464	76	-0.28	(-0.38, -0.18)
Polysemy	M2	Early	2514	55	-0.15	(-0.25, -0.05)
Polysemy	M2	Late	2490	58	-0.09	(-0.19, 0.01)
Homonymy	M1	Early	2944	78	-0.25	(-0.35, -0.14)
Homonymy	M1	Late	2792	48	-0.20	(-0.30, -0.10)
Homonymy	M2	Early	2936	91	-0.25	(-0.37, -0.13)
Homonymy	M2	Late	2842	50	-0.16	(-0.25, -0.06)

Table 2.7: Bayesian linear mixed-effects model fit to summed residual log response latencies in the disambiguation region in Experiment 1.

Effect	Posterior $\hat{\beta}$	Posterior σ_β	95% CRI
Intercept	-0.15	0.03	(-0.20, -0.10)
Disambiguator (Late)	0.02	0.03	(-0.05, 0.08)
Meaning (Subord.)	0.04	0.02	(-0.01, 0.08)
Target (Homonym)	-0.01	0.03	(-0.06, 0.04)
Disambig. × Meaning	0.02	0.02	(-0.02, 0.06)
Disambig. × Target	0.02	0.02	(-0.02, 0.07)
Meaning × Target	-0.02	0.02	(-0.06, 0.02)
Dis. × Mng. × Tgt.	-0.01	0.02	(-0.05, 0.03)

To assess the informativity of these findings with regard to the predictions discussed above, I computed Bayes factors between models which contained a term which corresponded to a particular hypothesized effect, and reduced models without that term, BF_{10}. Following state-of-the-art recommendations for Bayes factor computation (Schad et al., 2022; Nicenboim et al., 2022), I employed bridge-sampling tools in `brms` (i.e. the function `bayes_factor`), and used informative priors for the models I compared.

Best practices dictate that these priors should be extracted from a meta-analysis of studies on this phenomenon using the same methodology, but I lack an existing self-paced reading replication of Frazier and Rayner (1990), and other self-paced reading studies of homonymy do not examine the critical costs of late disambiguation. In lieu of direct empirical priors, I composed priors based on likely values under the current best-supported hypotheses, which expect that commitment to the meaning of a polyseme should be delayed here. I drew on the effect sizes reported in other studies on reanalysis of various temporary ambiguities in self-paced reading (Ferreira & Henderson, 1990; O. Bott & Hamm, 2014; Dotlačil & Brasoveanu, 2021) to characterize expectations for the homonymy conditions, while polysemy conditions were expected to feature no reanalysis costs. The resulting priors are listed in Table 2.8, and correspond to the weak expectation for a penalty of about 0.01 in residualized log latencies for late disambiguation regions which support a subordinate meaning of a homonym.

Using these as priors for a fully-specified model, I compared in sequence the evidence for inclusion of the predicted three-way interaction, an interaction of disambiguator position and meaning dominance, and an interaction of disambiguator position and target type. In the taxonomy of Lee and Wagenmakers (2013), results of this analysis indicate "anecdotal evidence" against the presence of the predicted three-way interaction (BF_{10} = 0.43), "extreme evidence" against a dependency between disambiguator position and meaning in the absence of a three-way interaction (BF_{10} = 0.001), and moderate evidence against a dependency between disambiguator position and target type in the absence of both interaction terms above (BF_{10} = 0.14). That is, even when the noisiness of the data and the small effect sizes expected by the literature are taken into account, the data are most compatible with the absence of any predicted effects in this region.

Table 2.8: Informative priors used for Bayes factor analysis of the disambiguator region in Experiment 1, derived from self-paced reading studies of other reanalysis effects.

Effect	Distribution
Intercept	$\mathcal{N}(0.00, 0.20)$
Disambiguator (Late)	$\mathcal{N}(0.01, 0.01)$
Meaning (Subord.)	$\mathcal{N}(0.01, 0.01)$
Target (Homonym)	$\mathcal{N}(0.00, 0.20)$
Disambig. × Meaning	$\mathcal{N}(0.01, 0.01)$
Disambig. × Target	$\mathcal{N}(0.01, 0.01)$
Meaning × Target	$\mathcal{N}(0.01, 0.01)$
Dis. × Mng. × Tgt.	$\mathcal{N}(0.01, 0.01)$

2.3.2.2 Target

Residual log response latencies on the target are presented in Figure 2.2 and Table 2.9. The model fitted to these latencies is reported in Table 2.10. I observe a main effect of disambiguator position, such that targets receive faster latencies in a neutral context than after disambiguation, $\hat{\beta}$ = -0.02, -0.03, -0.01, 95% CRI = (., I) also observe a small just-credible main effect of target type, such that homonyms were generally read slower, $\hat{\beta}$ = 0.01, 0.00, 0.02, 95% CRI = (., A)ll interactions are estimated close to zero, suggesting that these main effects are not meaningfully different across conditions.

Indeed, marginal comparisons reveal that neutral contexts were credibly faster than pre-disambiguation to much the same extent for dominant senses of polysemes, $\hat{\delta}$ = -0.04, $P(\delta < 0)$ = 0.97, subordinate senses of polysemes, $\hat{\delta}$ = -0.04, $P(\delta < 0)$ = 0.96, dominant meanings of homonyms, $\hat{\delta}$ = -0.04, $P(\delta < 0)$ = 0.98, and subordinate meanings of homonyms, $\hat{\delta}$ = -0.05, $P(\delta < 0)$ = 0.99. Of the accounts considered in Table 2.5, this overall pattern of slight difficulty on pre-disambiguated targets unmediated by meaning dominance is, again, most compatible with those which expect no incremental commitment to either polysemes or homonyms.

As above, to quantify the evidence these findings provide for or against my critical hypotheses, I computed various BF_{10} Bayes factors. In this case, to set informative priors, I drew on the effect sizes for specification costs reported in existing self-paced reading studies for homonyms (Binder & Rayner, 1998) and polysemes (Foraker & Murphy,

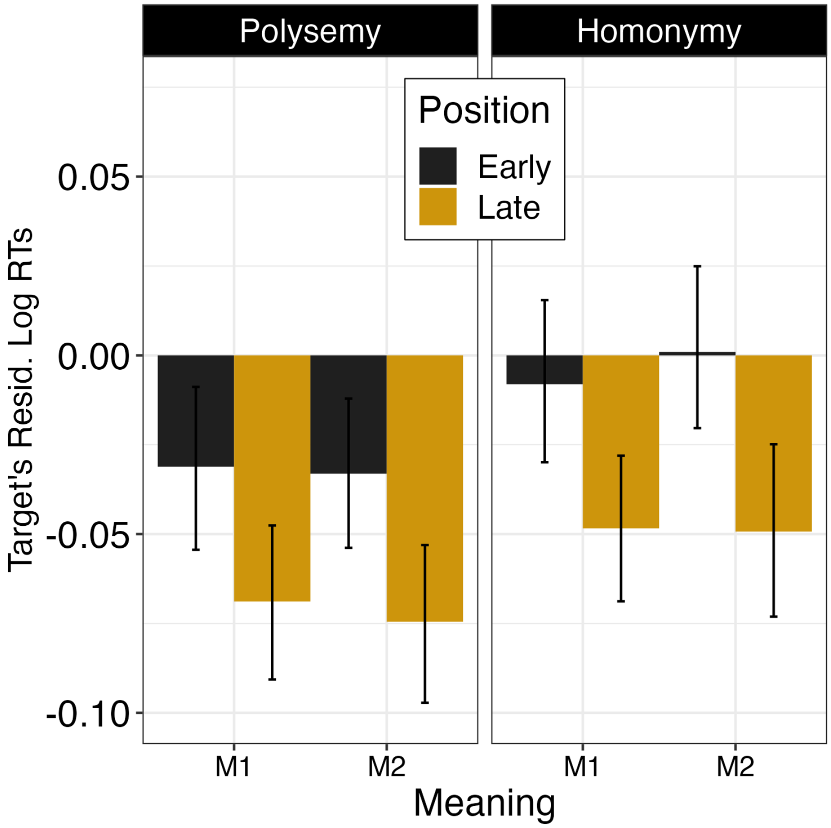

Figure 2.2: Residual log response latencies on the target in Experiment 1, by condition. Error bars represent bootstrapped 95% confidence intervals around the mean.

Table 2.9: Conditional means and measures of spread for the target in Experiment 1. Standard errors are reported over raw response latencies, and bootstrapped 95% confidence intervals are reported over the critical measure, residualized log response latencies.

Target	Meaning	Position	RT	SE	Resid. Log RT	95% CI
Polysemy	M1	Early	386	9	-0.03	(-0.05, -0.01)
Polysemy	M1	Late	382	11	-0.07	(-0.09, -0.05)
Polysemy	M2	Early	383	8	-0.03	(-0.06, -0.01)
Polysemy	M2	Late	373	9	-0.07	(-0.10, -0.05)
Homonymy	M1	Early	400	12	-0.01	(-0.03, 0.01)
Homonymy	M1	Late	376	8	-0.05	(-0.07, -0.03)
Homonymy	M2	Early	407	22	0.00	(-0.02, 0.03)
Homonymy	M2	Late	386	11	-0.05	(-0.07, -0.02)

Table 2.10: Bayesian linear mixed-effects model fit to residual log response latencies on the target in Experiment 1.

Effect	Posterior $\hat{\beta}$	Posterior σ_β	95% CRI Lower	95% CRI Upper
Intercept	-0.04	0.01	(-0.05, -0.03)	
Disambiguator (Late)	-0.02	0.01	(-0.03, -0.01)	
Meaning (Subord.)	-0.00	0.00	(-0.01, 0.01)	
Target (Homonym)	0.01	0.00	(0.00, 0.02)	
Disambig. $\times$ Meaning	-0.00	0.00	(-0.01, 0.01)	
Disambig. $\times$ Target	-0.00	0.00	(-0.01, 0.01)	
Meaning $\times$ Target	0.00	0.00	(-0.01, 0.01)	
Dis. $\times$ Mng. $\times$ Tgt.	-0.00	0.00	(-0.01, 0.01)	

2012). The resulting priors are listed in Table 2.11, and correspond to weak expectations for a general penalty of about 0.005 in residualized log latencies for all pre-disambiguated polysemes, and about 0.01 for pre-disambiguated homonyms just in subordinate meaning conditions.

The resulting analysis indicates extreme evidence for the absence of the predicted three-way interaction ($BF_{10} <$ 0.001).

2.3.2.3 Spillover

Summed residual log response latencies in the spillover region following the target are presented in Figure 2.3 and Table 2.12. The model fit to these latencies is reported in Table 2.13. I observe again a main effect of disambiguator position, larger in this region, such that targets receive faster latencies in a neutral context than after disambiguation, $\hat{\beta}$ = -0.09, 95% CRI = (-0.13, -0.05). This again seems to reflect some cost for reading some targets in a specifying context. I also observe again a main effect of target type, such that spillovers after homonyms receive in general slower latencies than spillovers after polysemes, $\hat{\beta}$ = 0.06, 95% CRI = (0.02, 0.09). Here, notably, the three-way interaction approaches a credibly non-zero estimate, $\hat{\beta}$ = -0.01, 95% CRI = (-0.04, 0.01).

Marginal comparisons suggest that this interaction reflects asymmetry in the strength of the penalty for targets in a specifying context. Spillovers after pre-disambiguated homonyms are associated with a large cost particularly when homonyms

Table 2.11: Informative priors used for Bayes factor analysis of the target in Experiment 1, derived from self-paced reading studies on lexical access for homonymy and polysemy.

Effect	Distribution
Intercept	$\mathcal{N}(0.00, 0.20)$
Disambiguator (Late)	$\mathcal{N}(-0.02, 0.01)$
Meaning (Subord.)	$\mathcal{N}(0.01, 0.01)$
Target (Homonym)	$\mathcal{N}(0.00, 0.20)$
Disambig. × Meaning	$\mathcal{N}(-0.01, 0.01)$
Disambig. × Target	$\mathcal{N}(0.00, 0.01)$
Meaning × Target	$\mathcal{N}(0.01, 0.01)$
Dis. × Mng. × Tgt.	$\mathcal{N}(-0.01, 0.01)$

were pre-disambiguated to their subordinate meaning, $\hat{\delta}$ = -0.20, $P(\delta > 0)$ = 0.99, and somewhat less when to their dominant, $\hat{\delta}$ = -0.13, $P(\delta < 0)$ = 0.99. This difference drives a trending marginal interaction of position and meaning for the homonyms, $\hat{\delta}$ = -0.07, $P(\delta < 0)$ = 0.85, a canonical subordinate selection effect. In contrast, spillovers after polysemes are associated with a consistent cost when the polysemes were pre-disambiguated, independent of whether they are disambiguated to dominant, $\hat{\delta}$ = -0.21, $P(\delta < 0)$ = 0.99 or subordinate meanings, $\hat{\delta}$ = -0.18, $P(\delta < 0)$ = 0.99, leaving the marginal interaction of position and meaning estimated near zero, $\hat{\delta}$ = 0.03, $P(\delta < 0)$ = 0.30. These differences are suggestive of the presence of a canonical subordinate selection penalty in homonyms but not polysemes, as would be predicted by accounts which expect that homonyms, but not polysemes, receive incremental commitment (see Table 2.5).

Bayes factor analysis was performed with priors based on spillover effect sizes in Binder and Rayner (1998) and Foraker and Murphy (2012) (Table 2.14), suggesting anecdotal evidence against the predicted three-way interaction (BF_{10} = 0.58).

2.3.3 Discussion

Results from self-paced reading reflect only some of the critical results of Frazier and Rayner (1990). Critical reanalysis effects were not observed in the disambiguation region. The best-supported hypotheses expect that homonyms will be specified sentence-medially, and thus that late disambiguation should sometimes be associated with reanaly-

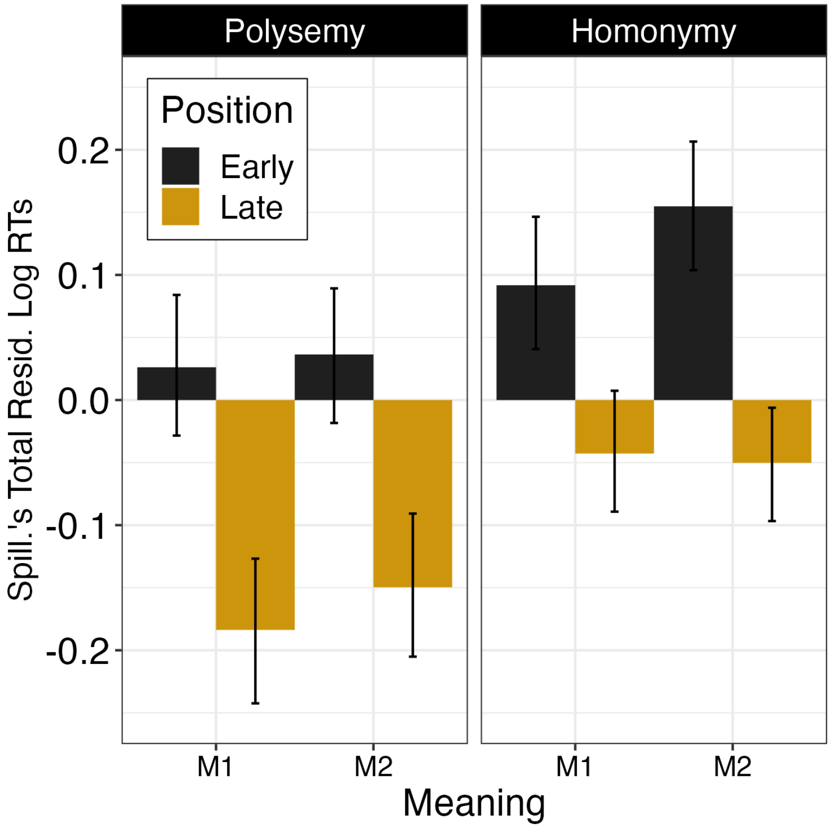

Figure 2.3: Summed residual log response latencies in the spillover region following the target in Experiment 1, by condition. Error bars represent bootstrapped 95% confidence intervals around the mean.

Table 2.12: Conditional means and measures of spread for the the spillover region following the target in Experiment 1. Standard errors are reported over the sum of raw response latencies in the region, and bootstrapped 95% confidence intervals are reported over the critical measure, summed residualized log response latencies.

Target	Meaning	Position	Sum RT	SE	Sum Resid. Log RT	95% CI
Polysemy	M1	Early	1264	31	0.03	(-0.03, 0.09)
Polysemy	M1	Late	1213	44	-0.18	(-0.24, -0.13)
Polysemy	M2	Early	1246	30	0.04	(-0.02, 0.09)
Polysemy	M2	Late	1158	30	-0.15	(-0.21, -0.10)
Homonymy	M1	Early	1129	28	0.09	(0.05, 0.14)
Homonymy	M1	Late	1067	27	-0.04	(-0.09, 0.01)
Homonymy	M2	Early	1134	26	0.15	(0.11, 0.21)
Homonymy	M2	Late	1045	25	-0.05	(-0.10, -0.00)

Table 2.13: Bayesian linear mixed-effects model fit to summed residual log response latencies in the spillover region in Experiment 1.

Effect	Posterior $\hat{\beta}$	Posterior σ_β	95% CRI Lower	95% CRI Upper
Intercept	-0.02	0.02	(-0.05, 0.02)	
Disambiguator (Late)	-0.09	0.02	(-0.13, -0.05)	
Meaning (Subord.)	0.01	0.01	(-0.01, 0.03)	
Target (Homonym)	0.06	0.02	(0.02, 0.09)	
Disambig. × Meaning	-0.00	0.01	(-0.03, 0.02)	
Disambig. × Target	0.01	0.01	(-0.02, 0.03)	
Meaning × Target	0.00	0.01	(-0.02, 0.02)	
Dis. × Mng. × Tgt.	-0.01	0.01	(-0.04, 0.01)	

Table 2.14: Informative priors used for Bayes factor analysis of the spillover in Experiment 1, derived from self-paced reading studies on lexical access for homonymy and polysemy.

Effect	Distribution
Intercept	$\mathcal{N}(0.00, 0.20)$
Disambiguator (Late)	$\mathcal{N}(-0.03, 0.01)$
Meaning (Subord.)	$\mathcal{N}(-0.01, 0.01)$
Target (Homonym)	$\mathcal{N}(0.00, 0.20)$
Disambig. × Meaning	$\mathcal{N}(0.01, 0.01)$
Disambig. × Target	$\mathcal{N}(0.03, 0.01)$
Meaning × Target	$\mathcal{N}(0.01, 0.01)$
Dis. × Mng. × Tgt.	$\mathcal{N}(-0.01, 0.01)$

sis costs. In turn, those reanalysis costs are not expected for polysemes, which by hypothesis remain underspecified and never require reanalysis. However, I observed only weak evidence of any sort of cost for late disambiguation, and although this was numerically more robust for homonyms, Bayes factor analysis concluded evidence was on the whole in favor of uniformity across target types and meaning dominance.

Evidence from latencies on the target and spillover was more in line with previous studies. Although no differences across target types or meanings were observed on the target itself, at the spillover I found some evidence for subordinate selection costs for homonyms only, driving a small near-credible three-way interaction, as expected if homonyms but not polysemes received a committed interpretation online.

Some studies have observed similar subordinate selection costs on polysemes, concluding that polysemes may be specified upon first encounter when context makes a particular sense salient (Lowder & Gordon, 2013; Brocher et al., 2016, 2018). Unlike those studies, I observed these costs only for homonyms. This could be a product of differing degrees of bias in the items: as noted in the discussion of norming above, this set of polysemes were on the whole less biased than the homonyms, and previous investigations have found these kind of general competition effects for polysemes without a bias (Brocher et al., 2018). It could also reflect a task difference: perhaps these SPR readers were less likely to engage in this hypothetically-optional specification than readers in those eyetracking studies.

The effect here which is more radically inconsistent with the eyetracking literature is the lack of costs on late material disambiguating to a subordinate meaning for homonyms, i.e. the lack of a canonical garden-path effect (cf. Duffy et al., 1988; Rayner and Frazier, 1989; Frazier and Rayner, 1990, etc.). This is especially unexpected given that evidence on the spillover following the target suggests that incremental commitment was taking place. In an effort to more directly compare my results to previous observations, I conducted a post-hoc analysis which examined participants' behavior in just the first half of the study—by later exposures, participants' reading may have been driven by familiarity with the stimuli in a way which previous studies had avoided by collecting fewer observations. Summed residual log response latencies in the disambiguating adjunct, split by experiment half, are presented in Figure 2.4. Analysis of behavior in the first half suffers from lack of power, and indeed a linear mixed-effects model finds no credible effects. However, marginal comparisons reveal that late disambiguation was about twice as costly for homonyms—whether dominant, $\hat{\delta}$ = 0.13, $P(\delta > 0)$ = 0.84, or subordinate, $\hat{\delta}$ = 0.11, $P(\delta > 0)$ = 0.81—than for polysemes—whether dominant, $\hat{\delta}$ = 0.06, $P(\delta > 0)$ = 0.66, or subordinate, $\hat{\delta}$ = 0.05, $P(\delta > 0)$ = 0.66. This fueled a positive but not credibly non-zero interaction of disambiguator position and target type, $\hat{\beta}$ = 0.02, 95% CRI = (-0.04, 0.08); Bayes factor analysis following the procedure described above yields "extreme" evidence for the presence of this interaction term in the model (BF_{10} > 1000). This sample more closely resembles the patterns reported in the literature, seeming to diagnose a higher likelihood of reanalysis following homonyms, although I still fail to observe an asymmetry by meaning dominance. The failure to find asymmetry here could be the result of inaccurate dominance norms, or measurement noise induced by the self-paced reading task, or else perhaps these participants selected lexical entries for the target more stochastically than in previous samples. Whatever the explanation for the lack of definition in the effect with homonyms, the presence of these costs in late disambiguation regions suggests reanalysis costs particular to homonyms, and so is in line with delayed specification for polysemes.

The fact that this distinction collapsed as the experiment continued into the second half suggests that comprehension patterns regarding commitment and reanalysis were subject to additional variation as comprehenders began to be familiar with the stimuli and procedure. Although strategic performance is a central focus of this study, without particular hypotheses or expectations about these kinds of effects of experience, I decline to interpret this effect in any detail.

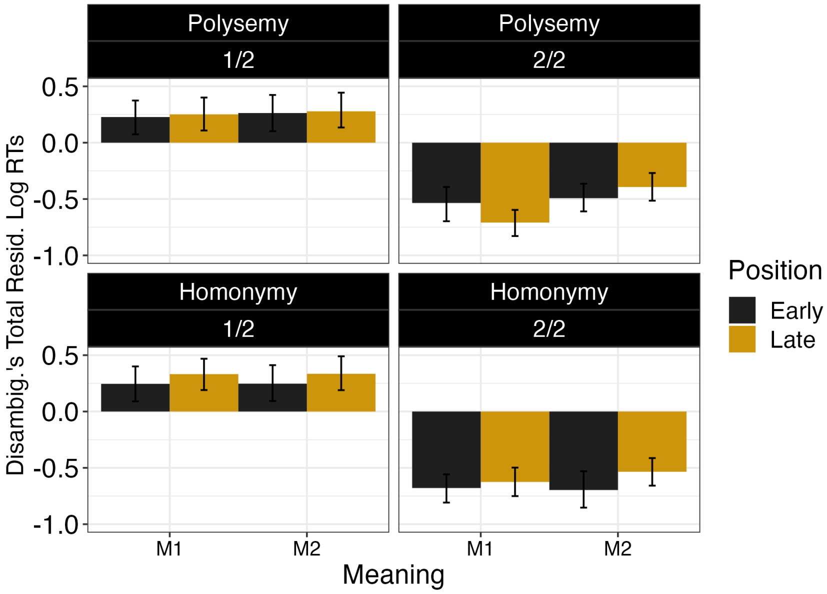

Figure 2.4: Summed residual log response latencies in the disambiguation region in Experiment 1, by condition. Error bars represent bootstrapped 95% confidence intervals around the mean.

On the whole, the self-paced reading data collected in Experiment 1 is weakly consistent with patterns observed in previous reading studies on polysemy, and the accounts they have argued for. Polysemes in neutral contexts are less likely to receive a complete interpretation than homonyms in neutral contexts, as evinced by a relative lack of difficulty during later disambiguation, although here this emerges only in the first half of participants' exposure to the test items, and was not subject to asymmetry conditioned on meaning dominance. I have also found partial evidence that when contextual support for a given sense is available, comprehenders are more likely to specify a homonym than a polyseme, with associated selection costs surfacing on reading behavior in the spillover. On the whole, these data exemplify the noise and temporal delays that have often been associated with self-paced reading experiments compared to eyetracking (Witzel et al., 2012). In the next experiment, I will probe for similar effects in a Maze task, to determine to what extent the timeline of polysemy resolution can be affected by task demands.

2.4 Experiment 2

In my second experiment, I investigated reading of this same set of stimuli in an A-Maze task, where incremental representations of sentential meaning are by hypothesis more useful than in normal day-to-day reading. In Experiment 1, I partially replicated a typical finding from the literature, that comprehenders appeared more likely to postpone decisions about the sense of a polyseme, reflected in the relative absence of costly reanalysis effects and subordinate selection effects for polysemes compared to homonyms. If the time course of decision-making in lexical comprehension is sensitive to task pressures, we might expect to find that participants in the Maze task treat polysemes more like homonyms, resolving to a full interpretation at the first opportunity. In particular, items with polysemes would be subject to reanalysis penalties when they receive late subordinate disambiguation, just like homonyms. Otherwise, if lexical processing decisions are insulated from this kind of non-linguistic contextual effect, we would expect to continue to observe those penalties only for homonyms.

2.4.1 Method

2.4.1.1 Participants

Another 48 native English speakers were recruited from the same Prolific and student pools as Experiment 1 in spring 2020, using the same criteria for participation. All participants again were of US nationality, except for two student participants raised in Canada.

2.4.1.2 Materials

The same 64 test items used in Experiment 1 served as the target sentences in the Maze task. Foils, which would have made nonsensical continuations of the sentences at each word, were prepared using the methods outlined in Boyce et al. (2020), by generating high-surprisal continuations using the Gulordava et al. (2018) language model and replacing repetitive or too-plausible foils by hand. Foils were matched across dominant and subordinate meanings of target words, and order was transposed in early and late disambiguation conditions so that, e.g., across all four conditions, participants would encounter a polyseme target paired with the same foil. Example foil strings for the target sentences in Table 2.1 are given in 2.15.

Table 2.15: Foil strings from Experiment 2 corresponding to the target sentences in Table 2.1.

Polysemy	
Early Disambiguation	Late Disambiguation
x-x-x intend in job lips discover obtain kid conducted add extension.	x-x-x kid conducted add extension intend in job lips discover obtain.
Homonymy	
Early Disambiguation	Late Disambiguation
x-x-x come fit detail sir thinks begin kept ours indecision Need.	x-x-x kept ours indecision Need come fit detail sir thinks begin.

About 60% of foils were syntactically ill-formed. In the remaining 40% of cases, foils were syntactically possible continuations which were nevertheless less plausible than their corresponding targets due to semantic context alone. For instance, in one practice item, participants encountered the choice between *dropped* and *welfare* after the context *The referee had*. Correctly choosing target *dropped* should not have posed much difficulty, but it did require participants to draw on something more than a syntactic representation of the context.

2.4.1.3 Procedure

The experiment was also prepared in Ibex, and deployed on IbexFarm. Participants proceeded through all Maze sentences word-by-word, by providing binary forced-choice responses on their keyboards. To motivate attentive participation in the difficult task, the interface displayed a counter of how many words in a row the participant had answered correctly. Participants who navigated through the sentence correctly were shown the same binary forced-choice comprehension questions used in Experiment 1. In the event that a participant picked a foil, the trial would end prematurely, the participant would be alerted to their mistake, their counter would reset, and they would still be shown this comprehension question.

Other than the mechanics of the Maze task, presentation, form assignment, and randomization was carried out as in Experiment 1. The same fillers, practice, and burn-in items were used. This procedure was estimated to take about 60 minutes.

2.4.1.4 Regions and analysis

1337 test trials in which a participant responded incorrectly to a Maze decision or a comprehension question were excluded from analysis of response latencies. The remaining sample includes data from 1735 test trials. Maze decision errors were analyzed separately as a secondary measure of incremental difficulty, but revealed no patterns of interest. Critical response latency measures and their analysis were computed using the same procedures as in Experiment 1.

2.4.2 Results

2.4.2.1 Disambiguator

Summed residual log response latencies in the disambiguating region are presented in Figure 2.5 and Table 2.16. The model fit to these latencies is reported in Table 2.17. Unlike Experiment 1, I observe several credibly non-zero fixed effects. First, I observe a main effect of disambiguator position, such that late disambiguating adjuncts elicited faster latencies, $\hat{\beta}$ = -0.24, 95% CRI = (-0.31, -0.17). This effect is consistent with some baseline difficulty processing the early adjuncts, perhaps due to some cost associated with encountering a pronoun before any plausible antecedent. I also observed a main effect of meaning, such that disambiguating adjuncts are generally read slower when consistent with a subordinate interpretation of the target, $\hat{\beta}$ = 0.08, 95% CRI = (0.01, 0.16).

Crucially, the model also estimated an interaction between disambiguator position and meaning, such that late disambiguation to the subordinate meaning of a target is associated with particularly slow latencies, $\hat{\beta}$ = 0.08, 95% CRI = (0.02, 0.13). Marginal comparisons confirm that this interaction is credibly non-zero for polysemes ($\hat{\delta}$ = 0.38, $P(\delta > 0)$ = 0.99), although it is only approaching credibility for homonyms ($\hat{\delta}$ = 0.23, $P(\delta > 0)$ = 0.93). This is the kind of pattern typically observed for only homonyms in previous literature. Referring back to Table 2.5, it is most compatible with an account that suggests that polysemes and homonyms both receive rapid incremental commitments. Even after a neutral context, in the Maze participants were apparently willing to specify a sense for a polyseme, and later subordinate disambiguation often required reanalysis.

As above, to quantify the evidence these findings provide for or against my crit-

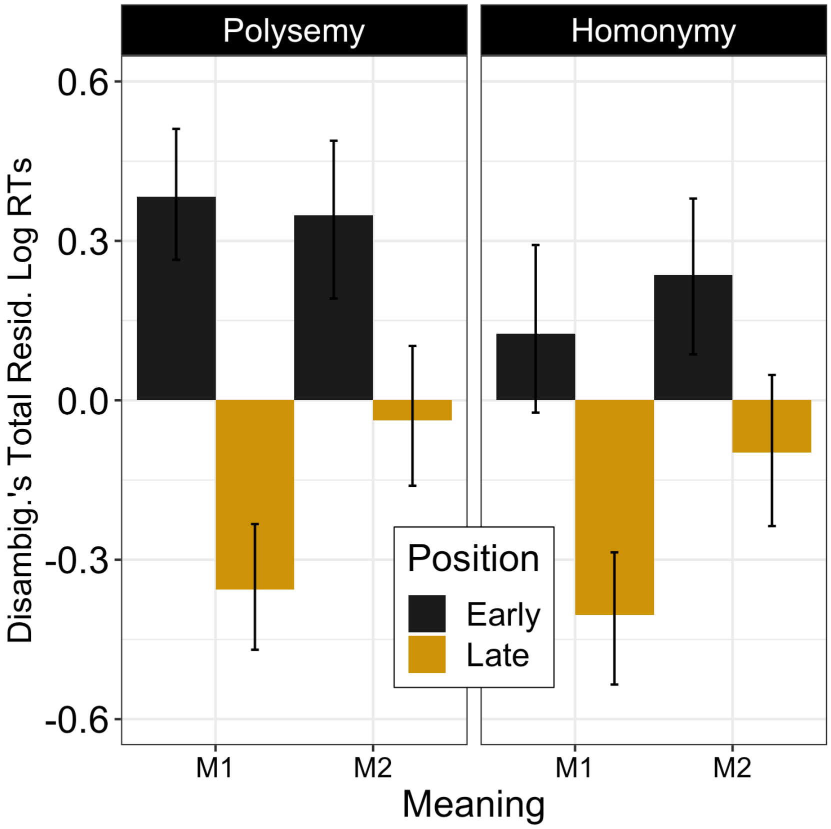

Figure 2.5: Summed residual log response latencies in the disambiguation region in Experiment 2, by condition. Error bars represent bootstrapped 95% confidence intervals around the mean.

Table 2.16: Conditional means and measures of spread for the disambiguation region in Experiment 2. Standard errors are reported over the sum of raw response latencies in the region, and bootstrapped 95% confidence intervals are reported over the critical measure, summed residualized log response latencies.

Target	Meaning	Position	Sum RT	SE	Sum Resid. Log RT	95% CI
Polysemy	M1	Early	6199	165	0.38	(0.25, 0.51)
Polysemy	M1	Late	5433	130	-0.36	(-0.47, -0.24)
Polysemy	M2	Early	6268	168	0.35	(0.21, 0.50)
Polysemy	M2	Late	5861	171	-0.04	(-0.16, 0.09)
Homonymy	M1	Early	7142	350	0.13	(-0.03, 0.29)
Homonymy	M1	Late	6259	132	-0.40	(-0.53, -0.28)
Homonymy	M2	Early	7067	150	0.24	(0.10, 0.39)
Homonymy	M2	Late	6850	164	-0.10	(-0.25, 0.05)

Table 2.17: Bayesian linear mixed-effects model fit to summed residual log response latencies in the disambiguation region in Experiment 2.

Effect	Posterior $\hat{\beta}$	Posterior σ_{β}	95% CRI Lower	95% CRI Upper
Intercept	0.03	0.03	-0.04	0.10
Disambiguator (Late)	-0.24	0.04	-0.31	-0.17
Meaning (Subord.)	0.08	0.04	0.01	0.16
Target (Homonym)	-0.05	0.05	-0.14	0.05
Disambig. × Meaning	0.08	0.03	0.02	0.13
Disambig. × Target	0.03	0.03	-0.03	0.10
Meaning × Target	0.02	0.04	-0.06	0.10
Dis. × Mng. × Tgt.	-0.02	0.03	-0.08	0.04

ical hypotheses, I computed various BF_{10} Bayes factors. In all region analyses for Experiment 2, informative priors were adopted from the same sources as reported in the analysis of Experiment 1, with expected parameter weights tripled given previous observations of how effect sizes in Maze results compare to effect sizes in self-paced reading (Witzel et al., 2012). The resulting analysis indicates anecdotal evidence against the presence of the three-way interaction predicted by some accounts (BF_{10} = 0.55), but strong evidence for a general interaction between position and meaning in the absence of the three-way interaction term (BF_{10} = 22.15).

2.4.2.2 Target

Residual log response latencies on the target are presented in Figure 2.6 and Table 2.18. The model fitted to these latencies is reported in Table 2.19. In addition to a credibly positive intercept, indicating more difficulty at this position than predicted by length and sentence position alone, I observed two credibly non-zero main effects. First, a small positive effect of disambiguator position, such that targets received slower latencies in a neutral context than after disambiguation to a dominant meaning, $\hat{\beta}$ = 0.04, 95% CRI = (0.02, 0.06). This is of interest given the opposite pattern observed in Experiment 1. Where that pattern in the self-paced reading task may suggest the presence of occasional costly specification, this effect in the Maze task is compatible with the possibility that a costly specification process was launched regardless of the context. This small additional cost in neutral contexts may reflect more difficulty in the absence of contextual information, as has been found in previous work on homonyms (Duffy et al., 1988; Rayner & Frazier, 1989), and as would be expected in a model of selection among competing meanings, because less evidence from context would provide less of an advantage to either candidate sense.

Second, I observed a larger main effect of target type, such that homonyms generally received slower latencies than polysemes, $\hat{\beta}$ = 0.09, 95% CRI = (0.06, 0.12). As this is a comparison across items, I refrain from over-interpreting it; there are many reasons why one set of words may have prompted a higher baseline degree of difficulty than another.

If homonyms and polysemes are behaving alike in the Maze, we also expect to find subordinate selection costs for both targets with early disambiguation. This would manifest here as a general interaction of disambiguator position and meaning, but that effect is estimated at zero in this region, $\hat{\beta}$ = 0.00, 95% CRI = (-0.02, 0.03). Marginal comparisons suggest the predicted subordinate selection effect does not hold for either polysemes

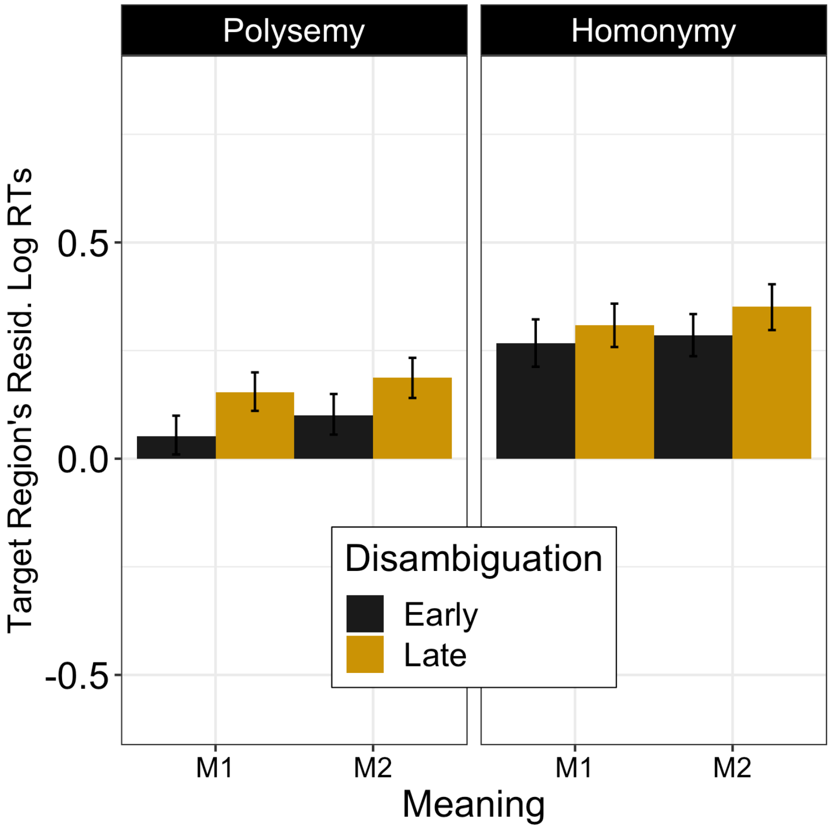

Figure 2.6: Residual log response latencies on the target in Experiment 2, by condition. Error bars represent bootstrapped 95% confidence intervals around the mean.

($\hat{\delta}$ = 0.00, $P(\delta < 0)$ = 0.49) or homonyms ($\hat{\delta}$ = 0.03, $P(\delta < 0)$ = 0.33).

Bayes factor analyses indicate strong evidence against the presence of the predicted three-way interaction (BF_{10}=0.07), and strong evidence against a general interaction between position and meaning in the absence of the three-way interaction term (BF_{10}=0.07).

2.4.2.3 Spillover

Summed residual log response latencies in the spillover region following the target are presented in Figure 2.7 and Table 2.20. The model fitted to these latencies is reported in Table 2.21. Here, I observed a main effect of meaning dominance, such that spillover response latencies are slower in sentences following a target in a subordinate specifying context, $\hat{\beta}$ = 0.04, 95% CRI = (0.00, 0.08). An additional interaction between disambiguator

Table 2.18: Conditional means and measures of spread for the target in Experiment 2. Standard errors are reported over raw response latencies, and bootstrapped 95% confidence intervals are reported over the critical measure, residualized log response latencies.

Target	Meaning	Position	RT	SE	Resid. Log RT	95% CI
Polysemy	M1	Early	999	30	0.05	(0.01, 0.10)
Polysemy	M1	Late	1103	39	0.15	(0.11, 0.20)
Polysemy	M2	Early	1074	39	0.10	(0.05, 0.15)
Polysemy	M2	Late	1150	38	0.19	(0.14, 0.23)
Homonymy	M1	Early	1256	53	0.27	(0.22, 0.32)
Homonymy	M1	Late	1253	41	0.31	(0.26, 0.35)
Homonymy	M2	Early	1234	35	0.29	(0.24, 0.34)
Homonymy	M2	Late	1332	52	0.35	(0.30, 0.41)

Table 2.19: Bayesian linear mixed-effects model fit to residual log response latencies on the target in Experiment 2.

Effect	Posterior $\hat{\beta}$	Posterior σ_{β}	95% CRI Lower	95% CRI Upper
Intercept	0.29	0.03	0.15	0.22
Disambiguator (Late)	0.04	0.01	0.02	0.06
Meaning (Subord.)	0.01	0.01	-0.01	0.03
Target (Homonym)	0.09	0.02	0.06	0.12
Disambig. × Meaning	0.00	0.01	-0.02	0.03
Disambig. × Target	-0.01	0.01	-0.03	0.01
Meaning × Target	-0.00	0.01	-0.03	0.02
Dis. × Mng. × Tgt.	0.00	0.01	-0.02	0.02

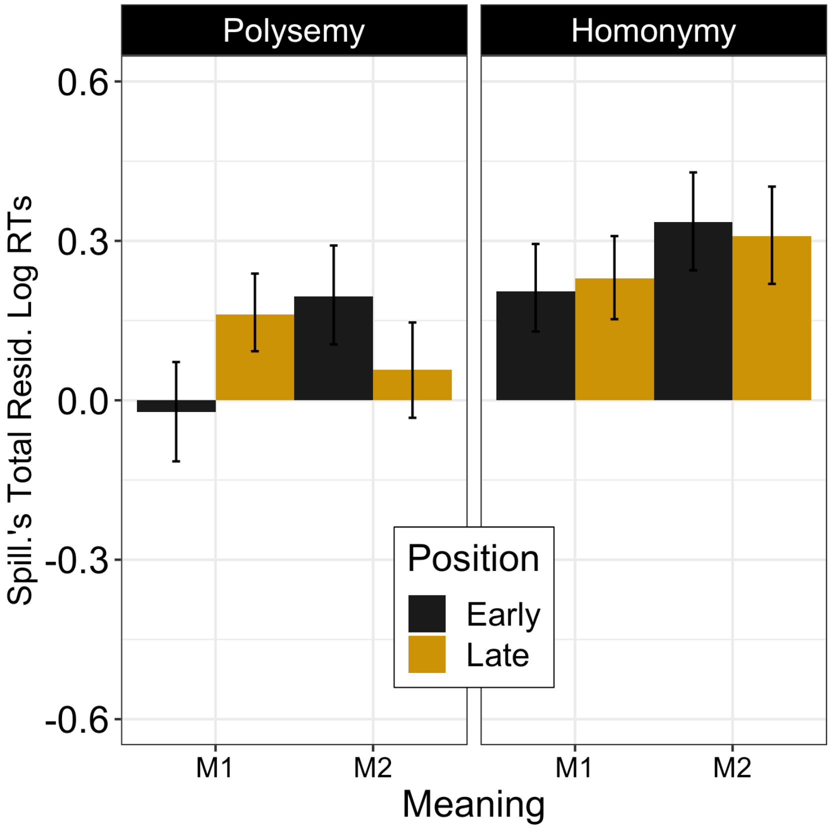

Figure 2.7: Summed residual log response latencies in the spillover region in Experiment 2, by condition. Error bars represent bootstrapped 95% confidence intervals around the mean.

position and meaning, $\hat{\beta}$ = -0.04, 95% CRI = (-0.08, -0.01), reflects the fact that the difference between dominant and subordinate meanings did not arise in late disambiguation conditions, where there were no meaning-driven differences at the spillover. This pattern may indicate the presence of subordinate selection effects at a short delay in the Maze. These delayed subordinate selection effects are present in the spillover for polyseme targets ($\hat{\delta}$ = 0.22, $P(\delta < 0)$ = 1.00) and for homonym targets ($\hat{\delta}$ = 0.12, $P(\delta < 0)$ = 0.95). This pattern is broadly in line with the predictions of either a maximal commitment account or an account where polysemes are delayed only in neutral contexts, per Table 2.5. Finally, the three-way interaction approaches a credibly positive estimate, $\hat{\beta}$ = 0.03, 95% CRI = (-0.00, 0.06), apparently driven by the fact that polysemes display a more prominent subordinate selection effect than homonyms in this region, counter expectations about lexical access for polysemes.

Table 2.20: Conditional means and measures of spread for the the spillover region following the target in Experiment 2. Standard errors are reported over the sum of raw response latencies in the region, and bootstrapped 95% confidence intervals are reported over the critical measure, summed residualized log response latencies.

Target	Meaning	Position	Sum RT	SE	Sum Resid. Log RT	95% CI
Polysemy	M1	Early	2972	95	-0.02	(-0.11, 0.07)
Polysemy	M1	Late	3081	99	0.16	(0.09, 0.24)
Polysemy	M2	Early	3149	111	0.20	(0.10, 0.29)
Polysemy	M2	Late	2921	96	0.06	(-0.03, 0.14)
Homonymy	M1	Early	2808	81	0.21	(0.12, 0.30)
Homonymy	M1	Late	2781	79	0.23	(0.15, 0.31)
Homonymy	M2	Early	2982	89	0.34	(0.24, 0.43)
Homonymy	M2	Late	2803	89	0.31	(0.22, 0.40)

Table 2.21: Bayesian linear mixed-effects model fit to summed residual log response latencies in the spillover region in Experiment 2.

Effect	Posterior $\hat{\beta}$	Posterior σ_β	95% CRI Lower	95% CRI Upper
Intercept	0.10	0.04	0.03	0.17
Disambiguator (Late)	0.01	0.02	-0.03	0.05
Meaning (Subord.)	0.04	0.02	0.00	0.08
Target (Homonym)	0.09	0.05	-0.00	0.18
Disambig. × Meaning	-0.04	0.02	-0.08	-0.01
Disambig. × Target	-0.01	0.02	-0.04	0.02
Meaning × Target	-0.00	0.02	-0.03	0.04
Dis. × Mng. × Tgt.	0.03	0.02	-0.00	0.06

Bayes factor analyses indicate strong evidence against the presence of the predicted three-way interaction (BF_{10}=0.06), and strong evidence against a general interaction between position and meaning in the absence of the three-way interaction term (BF_{10}=0.02).

2.4.3 Discussion

Results from the Maze task differed in several ways from the more familiar patterns observed in the eyetracking literature and partially reflected in self-paced reading. Most relevant to my predictions, in the Maze I observed a cost for late subordinate disambiguation of polysemy, comparable to what I observed for homonyms. This diagnoses the presence of mid-sentence commitments to a particular polyseme sense in neutral contexts. Indeed, latencies on the target word are consistent with the onset of a costly specification process upon encountering a polyseme, regardless of its context. It would thus appear that the tendency to delay the specification of a polyseme in typical reading can be bypassed in an unnatural reading task where earlier specification would be useful.

Moreover, when this early specification occurs, it closely resembles the profile of more canonical lexical ambiguity resolution. Where the results of Experiment 1 were suggestive of stochastic homonym reanalysis effects during disambiguation to dominant and subordinate meanings, in Experiment 2 both homonyms and polysemes showed costs specific to subordinate meaning, as expected under models of lexical ambiguity following Rayner and Frazier (1989). In Experiment 2, it would appear that upon encountering either a homonym or a polyseme in a neutral context, comprehenders engaged in full specification of its meaning, in which item-specific biases guided comprehenders more often towards a dominant meaning, so that later disambiguation to a subordinate meaning required reanalysis more frequently than later disambiguation to the dominant meaning.

These results fall in line with other recent proposals that task effects may modulate how comprehenders deploy processing mechanisms in real time (Pickering et al., 2006; Logačev & Vasishth, 2016). They are, to my knowledge, the first evidence that this goal-oriented modulation can adjust the process of lexical access. As such, they provide strong support that what has been previously characterized as default conservativity in processing may be less of a pre-determined heuristic, and more of a consequence of strategic allocation of resources depending on a task-specific notion of benefit. I extend this reasoning in the general discussion.

In requiring the choice of a sense-making continuation at each word in a target sentence, this experiment also provided precise localization of the effects during the processing of polysemy and homonymy in a specifying context with an unnatural reading task. I take these data to support the validity of my linking assumptions for interpreting response latencies in the Maze task. Two effects that were contingent on exposure or merely trending for homonyms in SPR emerged as more robust here: costly late subordinate disambiguation in a neutral context and costly subordinate selection in a biasing context. Nevertheless, the slowdown during subordinate selection, in both studies, is somewhat delayed compared to what has been observed in those eyetracking studies and in other investigations of homonyms in the Maze (Forster et al., 2009), emerging not on the target word but in the predicate which followed. These results thus broadly add to previous claims that the Maze allows for more precise estimation of difficulty in incremental reading than self-paced reading (Witzel et al., 2012), but perhaps cast doubt on the idea that they can be expected to compare directly to results from monitoring eye movements.

In order to demonstrate that the effects of interest here are the consequence of general differential resource allocations across tasks, and not merely the by-product of some anomalous properties of homonymy and polysemy, I extended the current approach to another area of linguistic meaning in which comprehenders are expected to eventually specify the sense of a word in order to comprehend its containing sentence, but may not need to do so in order to construct an incremental parse—distributivity.

2.5 Distributivity in incremental comprehension

We turn now to another case study using the Maze to study task effects in incremental comprehension, moving beyond lexical processing to the resolution of systematic implicit ambiguities in the relation of verbs to plural subjects. In a sentence like *Bernadette and Jackie washed two cars*, work in formal semantics has distinguished at least three possible readings (Roberts, 1987; Landman, 2000; Nouwen, 2012; Brasoveanu, 2013; Champollion, 2020). On a **distributive** interpretation, Bernadette and Jackie washed two cars *each*, so that four cars were washed in total. On a **collective** interpretation, Bernadette and Jackie washed two cars *together*, so that both were involved in washing each car. Finally, on a **cumulative** interpretation, Bernadette and Jackie merely must have done some, less specified, combination of car-washing so that two cars were washed in the end. I will focus on the first two here, which are by assumption cleanly disambiguated by the presence of adverbs *each* and *together*. Within the theoretical semantics literature, the dominant account of the difference between these two interpretations is a distinction in structure, whereby distributive sentences contain a phonetically-silent operator D in their syntactic and semantic representation which mediates the relationship between the individuals in a plural subject and the events described by a predicate. In the absence of the distributive operator, it is assumed that the sentence entails the subjects' collective participation. Whether or how comprehenders may resolve the ambiguity between these two readings in real time language comprehension is not predicted by this theoretical semantic account, but perhaps a minimal augmentation of it to form a processing prediction would posit that comprehenders may, but do not necessarily have to, posit a more complex structural relationship between individuals and events when confronted with a verb which describes a potentially distributive event.

Existing studies on the incremental comprehension of such sentences have sought to relate the processing of distributivity ambiguities to general models of ambiguity resolution in sentence processing. Under the hypothesis that decisions of linguistic representation must be resolved when they are encountered, Frazier et al. (1999) posited that, if distributivity ambiguities are indeed structural, a comprehender must commit to either a distributive or a collective reading upon initial representation of the verb (e.g., at *washed*). Assuming that simpler, collective readings will be favored as the initial interpretation of events predicts costly reanalysis when a late disambiguating adverb *each* is read after the object of the verb, as compared to late collective disambiguation with *together*. In an eyetracking experiment, Frazier et al. found evidence for exactly this: extra difficulty at the spillover region following *each* in first pass and total reading times, and greater probability of regressions out.

In recent work, Dotlačil and Brasoveanu (2021) built on this result in two ways in a pair of self-paced reading experiments. First, they conducted a conceptual replication of Frazier et al.'s study and found comparable effects with different items and critical adverbs (e.g., *individually*). Second, they confirmed that the reanalysis effect is limited to the ambiguities of **phrasal distributivity** arising for verbs with explicit objects (*The girls (each) slept on a narrow bed.*), rather than distributive readings which may arise without structural distinctions for simpler predicates (*The girls slept.*).

Despite the consistencies between these two studies, more fine-grained questions about the time course of distributive versus collective representations remain ill-addressed by the existing data. First, it is unclear that online evidence supports the assumption that distributivity requires a more complex representation than collectivity: predicates which were pre-disambiguated by *each* trended towards faster reading times for Frazier et al. (1999), and residual response latencies for *together* were longer than for *individually* in Dotlačil and Brasoveanu (2021). Second, with higher power than Frazier et al. (1999), the Dotlačil and Brasoveanu (2021) study nevertheless found only a very small reanalysis effect on the critical target, and somewhat stronger effect on an immediately following spillover. There may thus be more variability in the timing and bias of commitment to a collective or distributive reading than extant theories have predicted.

We might also particularly be curious about these effects given the hypothesis from Pickering et al. (2006) that certain aspects of the particular interpretation of a verb phrase are not mandatorily resolved before the end of a sentence. It would be surprising, pre-theoretically, if distinctions of distributivity were mandatory representational decisions but distinctions of verbal aspect were not: both involve the same kinds of interpretive consequences for, e.g., the number of events that occurred, and the relationship between those events and their arguments. As a result, we might expect to see the same kind of flexibility of decision-making for both in normal reading. If there is flexibility, this offers another chance to examine the ways in which task pressures modulate decision-making.

I take up the investigation of distributive ambiguity across self-paced reading and the Maze in Experiments 3 and 4. Under the hypothesis that distinctions in linguistic representation are resolved immediately, and assuming that distributive ambiguities are indeed true structural ambiguities, I expected to replicate previous studies in both self-paced reading and the Maze. In particular, I expected to observe costs for late disambiguation to a distributive reading. Under the hypothesis that some higher-order representational distinctions may be optionally postponed, and that task demands can control this decision-making process as observed in Experiments 1 and 2, we expect to see variation across the two tasks. In particular, I expected to observe costs for late disambiguation to a distributive reading, and other evidence for immediate interpretation, more strongly in the Maze than in self-paced reading. My results support the latter hypothesis.

2.6 Experiment 3

To begin, I aimed to conduct a conceptual replication of Frazier et al. (1999) and Dotlačil and Brasoveanu (2021) using fixed-window word-by-word self-paced reading, to set a baseline for further investigation in the Maze. I based my materials on the test item sets from Frazier et al. (1999), expanding and systematizing them to increase my power to detect the effects of interest. I predicted minimally that I would observe their same evidence for comprehension difficulty when possibly-distributive predicates were presented in a neutral context followed by distributive disambiguation.

2.6.1 Method

2.6.1.1 Participants

As in Experiment 1, 48 native English speakers were recruited equally from Prolific and a student participation pool in early 2021, and an additional sample of 48 were added in 2023, using the same criteria for participation. All participants again were of US nationality, except for one student participant raised in Poland.

2.6.1.2 Materials

The experiment contained 32 critical items with conjoined proper name subjects. 16 items were minimal variants of items from Frazier et al. (1999), and the remainder were constructed to the same template. All items contained an initial adverbial so that regions of interest were never sentence-initial. Conjoined proper name subjects were always followed by a predicate containing some countable quantity, most often a direct object introduced with *a(n)*, *one*, or a larger number. Predicates were always disambiguated by an adverb *together* or *each*. The position and the identity of the disambiguating adverb were manipulated across four within-item conditions in a 2×2 design, either coming early (directly before the verb) or late (after the verb phrase). A sample item is provided in Table 2.22.

Meaning dominance for all predicates in the materials was assessed in a separate forced-choice interpretation task with 36 participants recruited from a student pool according to the same restrictions described above. Participants saw un-disambiguated versions of the critical predicates, and were asked to select the "most likely meaning" of the sentence. For instance, participants saw the string *Bernadette and Jackie washed two*

Table 2.22: A sample item from Experiments 3 and 4.

Meaning	Early Disambiguation	Late Disambiguation
together	Luckily, Bernadette and Jackie together washed two cars before the hose broke.	Luckily, Bernadette and Jackie washed two cars together before the hose broke.
each	Luckily, Bernadette and Jackie each washed two cars before the hose broke.	Luckily, Bernadette and Jackie washed two cars each before the hose broke.

cars, and were asked to choose between the responses *They washed two cars each* and *They washed two cars together.*

Results from this norming task indicated reliable preferences for collective interpretations for the majority of items: collective readings were selected on on 73% of trials. The strengths of these preferences were evaluated by computing 95% confidence intervals using a non-parametric bootstrap around the proportion of responses in line with a collective preference. Intervals which excluded chance (50%) were taken to diagnose a reliable preference. 24 of the 32 predicates evinced a reliable collective preference, 7 predicates evinced no reliable preference, and 1 predicate (*weighed 220 pounds*) evinced a reliable distributive preference. Because not all predicates were reliably collective-biased, secondary analyses for Experiments 3 and 4 were computed using only the reliable collective items. Unless otherwise noted, these analyses provided comparable results.

2.6.1.3 Procedure

The experiment was prepared in Ibex, and deployed on PCIbexFarm (Zehr & Schwarz, 2018). Like Experiment 1, participants proceeded through all items using non-cumulative, fixed-window self-paced reading.

Items were followed by binary forced-choice comprehension questions. For half of the critical items, these were shallow questions that tested the participants' ability to recall surface-level aspects of the sentence. For instance, after an item with the predicate *baked several cakes*, participants were asked whether the characters were baking "cakes" or "bread". For the other half of the critical items, these questions probed the distributivity of the predicate; for instance, after the item in Table 2.22, participants were asked whether "two" or "four" cars got washed. As in Experiments 1 and 2, these more detailed questions served to ensure that participants had a baseline motivation to completely comprehend the critical items in both tasks.

Test items were presented in pseudo-randomized order across four Latin-square lists, balanced within each of my two samples of participants. They were mixed with filler items from unrelated experiments, 68 containing unambiguous transitive sentences, and 28 containing aspectual ambiguities resolved by a fronted or post-verbal temporal adverbial. To ensure these fillers served as suitable camouflage for the critical items, many featured similar sentence-initial adverbs. A further 32 unambiguous fillers were designed to exactly parallel the structure of the critical items. All fillers were also followed by comprehension questions. Eight of the unambiguous fillers were presented to participants as practice items, and another eight were reserved as "burn-in" items, presented in the main body of the experiment but sequenced before participants were shown the first test item.

After completing all 160 trials, participants completed the same questionnaires as used in Experiments 1 and 2, before receiving compensation. The procedure in entirety was estimated to take about 40 minutes.

2.6.1.4 Regions and analysis

221 critical trials in which a participant responded incorrectly to a comprehension question were excluded from analysis. The remaining sample included data from 2851 critical trials.

Four measures of word-by-word response latencies were computed for the analysis of critical items, aiming to observe the effects noted in Frazier et al. (1999) and Dotlačil and Brasoveanu (2021). All measures relied on residual log response latencies, derived from a linear mixed-effects model fit using the `lme4` package in `R` (R Core Team, 2016; Bates et al., 2015) to log response latencies for words in all unexcluded trials, with fixed slopes for number of characters and position in the sentence, and random participant intercepts. Critical measures were (i) the residual log response latencies on the critical disambiguating adverb, (ii) the summed residual log response latencies across the first three words of the predicate, (iii) residual log response latencies on the first word of the spillover region which was directly after the adverbs in their late position, and (iv) the summed residual log response latencies across the first three words in that spillover region. Regions are demonstrated for a sample item in Table 2.23.

For analysis, Bayesian linear mixed-effects models were fitted to these three

Table 2.23: Regioning used for analysis of items with early and late disambiguation in Experiments 3 and 4. Note that some items featured predicates longer than three words, in which case only the first three were analyzed.

	Adverb	Predicate	Spillover	
Luckily, Bernadette and Jackie	[each]	[washed two cars]	[before the hose]	broke.

	Predicate	Adverb	Spillover	
Luckily, Bernadette and Jackie	[washed two cars]	[each]	[before the hose]	broke.

Table 2.24: Predictions for Experiments 3 and 4 regarding critical Position × Meaning interactions under three possible generalizations about the processing of distributivity. Note that a marginal positive Position × Meaning interaction diagnoses particular cost in the Late *each* or Early *together* condition, while a marginal negative Position × Meaning interaction diagnoses particular cost in the Late *together* or Early *each* condition.

Account	**Adverb**	**Predicate**
Immediate commitment w/ collective bias	$+$	$-$
Immediate commitment w/ distributive bias	$-$	$+$
Commitment only when pre-specified	0	$+/-$
Never commitment or no bias	0	0

measures with STAN (Stan Development Team, 2019) using the `brms` package in `R` (Bürkner, 2017, 2018) using weakly-informative priors and other parameters set as in Table 2.4, with maximal random effects and treatment-coded predictors. Early disambiguation and *together* were coded as reference levels.

2.6.2 Results

All 96 participants retained for analysis answered comprehension questions with accuracy of greater than 80%. Other recruited participants were excluded from analysis to ensure a base level of attentive comprehension. Mean accuracy in the final sample was 97.1%. In this section, I report the response latencies in the various regions of interest. Before doing so, I will highlight the tangible predictions made by a few different possible generalizations about the interpretation of distributivity, summarized in Table 2.24.

If comprehenders made an immediate commitment to the distributivity of a predicate, predictions depend on the direction of their bias. If comprehenders are biased towards a collective interpretation, they will commit to it in neutral contexts, driving costly reanalysis in cases of late disambiguation to a distributive interpretation. In that case, this design would observe a positive interaction term at the disambiguating adverb such that late *each* drove particularly slow reading. Immediate commitment with a collective bias would also predict subordinate selection costs at the verb and following regions driven by difficulty selecting the subordinate reading; here that would be the distributive reading. This design would then observe a negative interaction term at the predicate/spillover such that early *each* drove particularly slow reading at the subsequent predicate

On the other hand, if comprehenders are biased towards a distributive interpretation, costly reanalysis would be observed on late disambiguation to a collective interpretation. This should drive a negative interaction term at the adverb such that late *together* drove particularly slow reading. Likewise, we would predict subordinate selection costs for selecting the collective reading, as reflected in a positive interaction term at the predicate and spillover such that early *together* drove particularly slow reading.

If we imagine that distributivity ambiguities are only resolved when unambiguous, we expect no particular garden-path costs associated with late disambiguation, but subordinate selection costs on predicates with pre-disambiguation according to whatever bias we imagine as above. E.g. under a collective bias, we would expect some difficulty with predicates pre-disambiguated by *each*, which would be associated with a negative interaction.

Finally, if there is no consistent bias for one meaning or another, or if distinctions between distributive and collective readings are not represented at all during incremental comprehension, even when disambiguated, we expect no asymmetries of context or disambiguation, and thus no credible interactions in either region.

The rest of this section will report the patterns of latencies observed in each region. Distributions of latencies in the various critical regions are displayed in Figures 2.8 and 2.9.

2.6.2.1 Adverb

Residual log response latencies on the critical adverb are presented in Table 2.25. The model fitted to those latencies is reported in Table 2.26. I observed a main effect of disambiguator position, such that post-verbal adverbs received slower latencies than pre-verbal adverbs, $\hat{\beta}$ = 0.02, 95% CRI = (0.01, 0.04). Marginal comparisons reveal that this slowdown was present and of a similar size for both *together* ($\hat{\delta}$ = 0.05, $P(\delta > 0)$ = 1.00) and *each* ($\hat{\delta}$ = 0.04, $P(\delta > 0)$ = 0.98). I also observed a main effect of meaning, such that

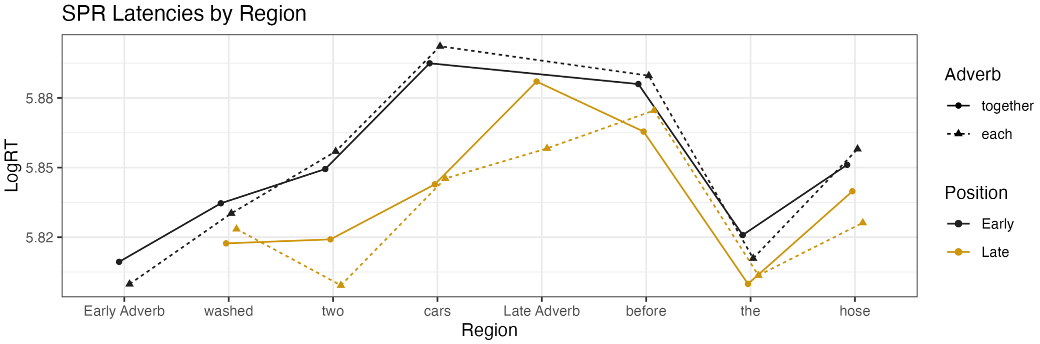

Figure 2.8: Log response latencies at various positions in Experiment 3, by condition.

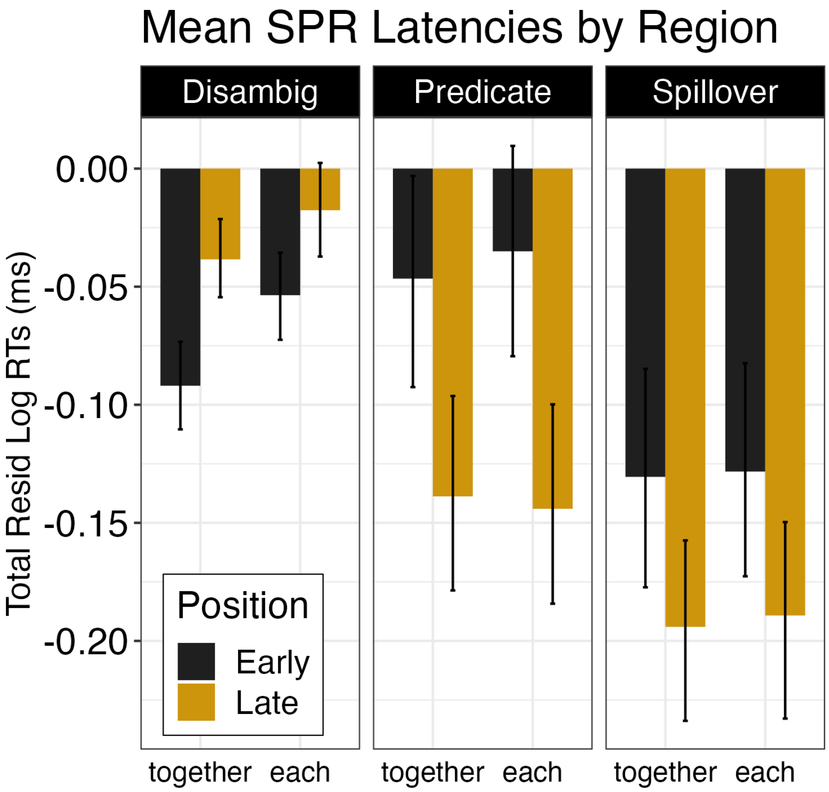

Figure 2.9: Summed residual log response latencies in the critical regions of Experiment 3, by condition.

Table 2.25: Conditional means and measures of spread for the disambiguating adverb in Experiment 3. Standard errors are reported over raw response latencies, and bootstrapped 95% confidence intervals are reported over the critical measure, residualized log response latencies.

Meaning	Position	RT	SE	Resid. Log RT	95% CI
together	Early	378	7	-0.04	(-0.06, -0.02)
together	Late	386	7	-0.03	(-0.05, -0.01)
each	Early	379	8	-0.05	(-0.07, -0.03)
each	Late	396	9	-0.02	(-0.04, 0.00)

each received faster latencies than *together*, $\hat{\beta}$ = 0.02, 95% CRI = (0.00, 0.03). The predicted interaction of disambiguator position and meaning, which would be consistent with particular cost for post-verbal *each*, was not credibly non-zero, $\hat{\beta}$ = -0.00, 95% CRI = (-0.01, 0.01).

As in the previous experiments, Bayes factor analyses were computed over alternative models with informative priors based on expectations from previous findings. In this case, priors came directly from the observed effects in Dotlačil and Brasoveanu (2021), Experiment 1. Comparisons between models with and without interaction parameters suggest strong evidence for the absence of the predicted interaction (BF_{10}=0.06).

The lack of a credible interaction can only be compatible with stochastic specification without bias, or else some delay of specification (Table 2.24). The main effect of disambiguator position could be compatible with either case. On the one hand, if participants chose to commit to a distributive meaning stochastically without any particular bias, they would have to engage in reanalysis on some even proportion of the trials for late *together* and late *each*. The slower average reading for late disambiguation could come from trials which featured this reanalysis. On the other hand, if participants have postponed a decision about the fine verbal structure, late adverbs could be more costly than early adverbs because they launch a specification process for the verb. In the latter case, we would expect to see a trade-off between observing symmetric costs of specification on the adverb (and spillover) in late disambiguation trials, and observing symmetric costs of specification on the predicate in early disambiguation trials.

Table 2.26: Bayesian linear mixed-effects model fit to residual log response latencies on the disambiguating adverb in Experiment 3.

Effect	Posterior $\hat{\beta}$	Posterior σ_β	95% CRI Lower	95% CRI Upper
Intercept	-0.05	0.01	-0.07	-0.03
Disambiguator (Late)	0.02	0.01	0.01	0.04
Meaning (*each*)	0.02	0.01	0.00	0.03
Disambig. × Meaning	-0.00	0.01	-0.01	0.01

Table 2.27: Conditional means and measures of spread for the predicate region in Experiment 3. Standard errors are reported over summed raw response latencies, and bootstrapped 95% confidence intervals are reported over the critical measure, summed residualized log response latencies.

Meaning	Position	Sum RT	SE	Sum Resid. Log RT	95% CI
together	Early	1197	21	-0.00	(-0.05, 0.04)
together	Late	1144	18	-0.11	(-0.15, -0.07)
each	Early	1193	21	0.02	(-0.03, 0.07)
each	Late	1131	22	-0.11	(-0.16, -0.07)

2.6.2.2 Predicate

Summed residual log response latencies in the predicate region are presented in Table 2.27. The model fitted to those latencies is reported in Table 2.28. I observe a negative main effect of disambiguator position, $\hat{\beta}$ = -0.05, 95% CRI = (-0.07, -0.03), such that predicates are read slower when they are pre-disambiguated. I observe no other effects of note, including the absence of a credible interaction with meaning, $\hat{\beta}$ = -0.00, 95% CRI = (-0.03, 0.02), and Bayes factor analysis suggests anecdotal evidence against the presence of any interaction (BF_{10}=0.49).

The cost for pre-disambiguated predicate is the mirror-image of the cost observed above for late disambiguation on the adverb itself. I entertained the idea that the late disambiguation cost was a result of late, adverb-triggered selection between apparently equibiased collective and distributive interpretations. This opposite pattern of costs on the predicate is compatible with that proposal, as a signature of the same costly selection process, running here on the predicate given the presence of a specifying context.

Table 2.28: Bayesian linear mixed-effects model fit to summed residual log response latencies in the predicate region in Experiment 3.

Effect	Posterior $\hat{\beta}$	Posterior σ_β	95% CRI Lower	95% CRI Upper
Intercept	-0.08	0.02	-0.12	-0.03
Disambiguator (Late)	-0.05	0.01	-0.07	-0.03
Meaning (*each*)	0.00	0.01	-0.03	0.03
Disambig. × Meaning	-0.00	0.01	-0.03	0.02

2.6.2.3 First Spillover

Residual log latencies at the first word of the spillover region are presented in Table 2.29. The model fitted to those latencies is reported in Table 2.30. I observed no credibly non-zero effects in this region. A main effect of position is approaching credibility as a negative estimate, $\hat{\beta}$ = -0.01, 95% CRI = (-0.02, 0.00), indicating faster processing in conditions with late disambiguation, where this region followed the disambiguating adverb, and slower processing in early disambiguation conditions, where this region followed a pre-disambiguated predicate. Again, I do not observe a credible interaction, $\hat{\beta}$ = 0.00, 95% CRI = (-0.01, 0.01), and Bayes factor analysis suggests moderate evidence for the absence of the predicted interaction (BF_{10}=0.13).

The above examination of the earlier regions in the sentence revealed that both predicates with early disambiguation and late adverbs were found to exhibit some processing costs. It may be the case that those costs are spilling over somewhat into this region, which makes it hard to interpret any effects here. If anything, at this position it appears that the slowdown associated with interpreting a predisambiguated predicate had a more influential spillover effect than the slowdown associated with interpreting a late adverb.

2.6.2.4 Full Spillover

Summed residual log latencies across the full spillover region are presented in Table 2.31. The model fitted to those latencies is reported in Table 2.32. I observe a credibly negative main effect of disambiguator position, $\hat{\beta}$ = -0.03, 95% CRI = (-0.05, -0.01), indicating faster comprehension following late disambiguating adverbs compared to following pre-disambiguated predicates, in line with the trend observed on the first position. As in all other regions, I observe an interaction whose estimate is not credibly non-zero, $\hat{\beta}$ = 0.01,

Table 2.29: Conditional means and measures of spread for the first word of the spillover region in Experiment 3. Standard errors are reported over raw response latencies, and bootstrapped 95% confidence intervals are reported over the critical measure, residualized log response latencies.

Meaning	Position	RT	SE	Resid. Log RT	95% CI
together	Early	369	7	-0.05	(-0.07, -0.03)
together	Late	360	6	-0.08	(-0.10, -0.06)
each	Early	374	11	-0.06	(-0.08, -0.04)
each	Late	358	6	-0.07	(-0.09, -0.05)

Table 2.30: Bayesian linear mixed-effects model fit to residual log response latencies on the first word of the spillover region in Experiment 3.

Effect	Posterior $\hat{\beta}$	Posterior σ_β	95% CRI Lower	95% CRI Upper
Intercept	-0.01	0.01	-0.03	0.01
Disambiguator (Late)	-0.01	0.01	-0.02	0.00
Meaning (*each*)	0.00	0.01	-0.01	0.02
Disambig. × Meaning	0.00	0.01	-0.01	0.01

Table 2.31: Conditional means and measures of spread for the full spillover region in Experiment 3. Standard errors are reported over summed raw response latencies, and bootstrapped 95% confidence intervals are reported over the critical measure, summed residualized log response latencies.

Meaning	Position	Sum RT	SE	Sum Resid. Log RT	95% CI
together	Early	1159	19	-0.16	(-0.20, -0.11)
together	Late	1138	18	-0.20	(-0.24, -0.16)
each	Early	1177	23	-0.13	(-0.18, -0.09)
each	Late	1119	16	-0.21	(-0.25, -0.17)

Table 2.32: Bayesian linear mixed-effects model fit to summed residual log response latencies in the full spillover region in Experiment 3.

Effect	Posterior $\hat{\beta}$	Posterior σ_{β}	95% CRI Lower	95% CRI Upper
Intercept	-0.13	0.02	-0.17	-0.08
Disambiguator (Late)	-0.03	0.01	-0.05	-0.01
Meaning (*each*)	-0.00	0.01	-0.03	0.02
Disambig. × Meaning	0.01	0.01	-0.02	0.03

95% CRI = (-0.02, 0.03). Bayes factor analyses suggest moderate evidence for the absence of the predicted interaction (BF_{10}=0.18).

As above, given the presence of independent costs which may spill over from the pre-spillover content regardless of the position of the disambiguator, it is difficult to interpret costs in this region. However, the effect seems to indicate that the costs observed above for pre-disambiguated predicates are more liable to spill over than the costs observed for late disambiguating adverbs.

2.6.3 Discussion

The results of Experiment 3 are not consistent with the findings of previous work, which has argued for rapid specification of distributivity ambiguities according to a preferential bias for collective interpretations. The latencies from self-paced reading show no evidence of any particular selection or late disambiguation cost for distributive interpretations, which is surprising given the results of Dotlačil and Brasoveanu (2021). It is possible that experiments with sample sizes in the neighborhood of this one and Dotlačil and Brasoveanu (2021) lack the power to reliably estimate what may be a small effect through a noisy measure like self-paced reading. It is also possible that there is simply more variance across participants in the comprehension of these sentences than previously considered.

Whatever the reason may be, having been unable to observe the systematic penalties we expect for late distributive disambiguation, I call into question whether the distributivity of a predicate is always mandatorily determined at the verb. Results from this experiment are consistent with a variable specification process, comparable to polysemy: specification costs that occur on pre-disambiguated verbs seem to trade off with specification costs triggered by late adverbs, without any evidence for reanalysis triggered by a more complicated disambiguation. A proposal of immediate specification without a collective bias would be closer to previous results, and similarly predicts no discernable interactions, but it would expect no particular costs for late adverbs, and thus cannot explain the observed trade-off. Nor could a more extreme proposal of complete lack of online specification, which would expect no particular effects of position.

The proposal of a temporary delay in specification leaves room for insight from the Maze. If comprehenders in less demanding reading tasks do not engage systematically in early decision-making for distributivity, we may expect that task demands can encourage them to do so to a greater degree.

2.7 Experiment 4

In my fourth experiment, I investigated reading of the same set of distributivity stimuli as Experiment 3 in an A-Maze task. In Experiment 3, I found that comprehenders may not generate incremental commitments about the distributivity of a predicate as systematically as previously observed. Based on my observations about the processing of polysemy in Experiment 2, participants in the Maze task may be more likely to adopt early commitments for distributive ambiguities as well. In particular, we should see reanalysis penalties emerge here for late disambiguation to a distributive meaning.

Table 2.33: Foil strings from Experiment 4 corresponding to the target sentences in Table 2.22.

Early Disambiguation	Late Disambiguation
x-x-x Congressed jobs Morals election bailey yeah pope health hear lieu proud.	x-x-x Congressed jobs Morals bailey yeah pope election health hear lieu proud.

2.7.1 Method

2.7.1.1 Participants

48 native English speakers were recruited from the same Prolific and student pools in early 2021, using the same criteria for participation as the previous experiments. All participants were of US nationality.

2.7.1.2 Materials

The same 32 test items used in Experiment 3 served as the targets in the Maze task. Corresponding foils were generated as in Experiment 2, matching foils across conditions and transposing order in early and late disambiguation cases so that critical adverbs within one item were always seen paired with the same foil. Example foil strings for the target sentences in Table 2.22 are given in 2.33.

As before, some foils were syntactically impossible continuations, while others were syntactically possible, but impossible in semantic context.

2.7.1.3 Procedure

The experiment was prepared in Ibex, and deployed on PCIbexFarm (Zehr & Schwarz, 2018). Maze trials proceeded as described in Experiment 2, followed by the comprehension questions described above for Experiment 3. Other than the mechanics of the Maze task, presentation, form assignment, and randomization was carried out as in Experiment 3. The same fillers, practice, and burn-in items were also used. This procedure was estimated to take about 60 minutes.

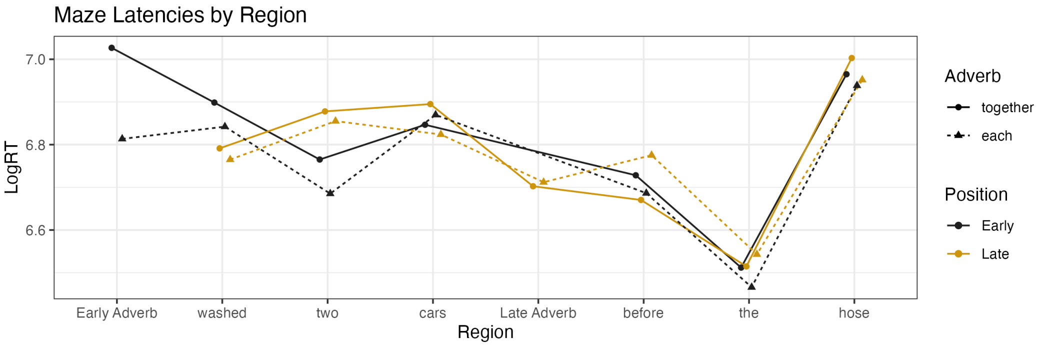

Figure 2.10: Log response latencies at various positions in Experiment 4, by condition.

2.7.1.4 Regions and analysis

427 critical trials in which a participant responded incorrectly to a Maze decision or a comprehension question were excluded from analysis of response latencies. The remaining sample included data from 1109 critical trials. Maze decision errors were analyzed separately as a secondary measure of incremental difficulty, but revealed no patterns of interest. Critical response latency measures and their analysis were computed using the same procedures as in Experiment 3.

2.7.2 Results

I now report the response latencies in the various regions of interest. Distributions of latencies in the various critical regions are displayed in Figures 2.10 and 2.11.

2.7.2.1 Adverb

Residual log response latencies on the critical adverb are presented in Table 2.34. The model fitted to those latencies is reported in Table 2.35. I observed a main effect of disambiguator position, such that post-verbal adverbs received faster latencies than pre-verbal adverbs, $\hat{\beta}$ = -0.11, 95% CRI = (-0.14, -0.09). This suggests some baseline cost for pre-verbal disambiguation, the reverse of what was observed in Experiment 3. Post-hoc investigation of marginal comparisons reveals that this holds for both *together*, $\hat{\delta}$ = -0.34, $P(\delta < 0)$ = 0.99, and *each*, $\hat{\delta}$ = -0.11, $P(\delta < 0)$ = 0.99.

The estimated effect of disambiguator meaning is not credibly non-zero, but trends towards *each* receiving slower latencies than *together*, $\hat{\beta}$ = 0.02, 95% CRI = (-0.01,

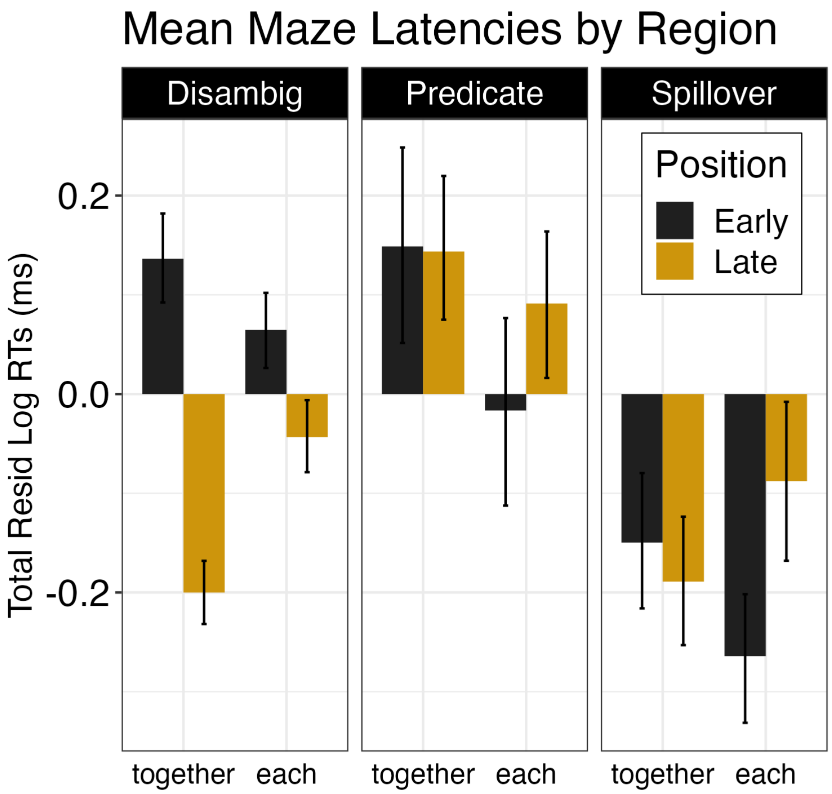

Figure 2.11: Summed residual log response latencies in the critical regions of Experiment 4, by condition.

Table 2.34: Conditional means and measures of spread for the disambiguating adverb in Experiment 4. Standard errors are reported over raw response latencies, and bootstrapped 95% confidence intervals are reported over the critical measure, residualized log response latencies.

Meaning	Position	RT	SE	Resid. Log RT	95% CI
together	Early	1263	54	0.14	(0.09, 0.18)
together	Late	864	23	-0.20	(-0.23, -0.17)
each	Early	970	24	0.06	(0.03, 0.10)
each	Late	877	23	-0.04	(-0.08, -0.01)

Table 2.35: Bayesian linear mixed-effects model fit to residual log response latencies on the disambiguating adverb in Experiment 4.

Effect	Posterior $\hat{\beta}$	Posterior σ_β	95% CRI Lower	95% CRI Upper
Intercept	-0.01	0.01	(-0.03, 0.02)	
Disambiguator (Late)	-0.11	0.01	(-0.14, -0.09)	
Meaning (*each*)	0.02	0.01	(-0.01, 0.05)	
Disambig. × Meaning	0.06	0.01	(0.04, 0.08)	

0.05). Post-hoc investigation of marginal comparisons reveals that this is driven by credible costs for post-verbal *each*, $\hat{\delta}$ = 0.16, $P(\delta > 0)$ = 0.99, while pre-verbally, *each* is read credibly faster than *together*, $\hat{\delta}$ = -0.08, $P(\delta < 0)$ = 0.98. This contrast is in line with expectations for greater difficulty for late *each* due to the presence of some reanalysis cost, and is reflected in a credible positive estimate for the interaction of disambiguator position and meaning, $\hat{\beta}$ = 0.06, 95% CRI = (0.04, 0.08). Bayes factor analysis with priors based on Dotlačil and Brasoveanu (2021) adjusted for typical Maze effect sizes concludes that these results provide extreme evidence for the presence of the expected interaction (BF_{10} > 1000). This apparent garden-path interaction is consistent with early commitment governed by a collective bias.

2.7.2.2 Predicate

Summed residual log latencies in the predicate region are presented in Table 2.36. The model fitted to those latencies is reported in Table 2.37.

Table 2.36: Conditional means and measures of spread for the predicate region in Experiment 4. Standard errors are reported over summed raw response latencies, and bootstrapped 95% confidence intervals are reported over the critical measure, summed residualized log response latencies.

Meaning	Position	Sum RT	SE	Sum Resid. Log RT	95% CI
together	Early	4017	76	0.15	(0.05, 0.25)
together	Late	2287	63	0.14	(0.07, 0.22)
each	Early	3836	83	-0.02	(-0.11, 0.08)
each	Late	2221	66	0.09	(0.02, 0.17)

I observed a main effect of disambiguator meaning, such that predicates associated with *each* received faster latencies than predicates associated with *together*, $\hat{\beta}$ = -0.06, 95% CRI = (-0.10, -0.02). Marginal comparisons reveal that this is driven by a credible difference in conditions with pre-verbal disambiguation, $\hat{\delta}$ = -0.18, $P(\delta < 0)$ = 0.99, while predicates which have not yet been disambiguated at this point show no credible difference, $\hat{\delta}$ = -0.06, $P(\delta < 0)$ = 0.87, as expected. This difference drove a positive interaction term approaching credibility, $\hat{\beta}$ = 0.03, 95% CRI = (-0.01, 0.07). The cost for processing preceding disambiguation with *together* is broadly consistent with the costs for *together* observed on the pre-verbal adverbs themselves. Of the generalizations reviewed in Table 2.24, this is most compatible with models with early specification and a distributive bias, and unexpected given the apparent collective bias driving garden-path effects on late adverbs.

Bayes factor analysis concludes that these results provide very strong evidence against the presence of the expected interaction, but this is presumably because the main effect of meaning is in the opposite direction from what was observed in Dotlačil and Brasoveanu (2021) (BF_{10} = 0.01).

2.7.2.3 First Spillover

Residual log latencies at the first word of the spillover region are presented in Table 2.38. The model fitted to those latencies is reported in Table 2.39. I observe no credible main effects of disambiguator position, $\hat{\beta}$ = 0.00, 95% CRI = (-0.02, 0.03), or meaning, $\hat{\beta}$ = 0.01, 95% CRI = (-0.01, 0.03), but a credibly-positive interaction term, $\hat{\beta}$ = 0.04, 95% CRI = (0.02, 0.06) Post-hoc marginal comparisons revealed that this was a cross-over interaction,

Table 2.37: Bayesian linear mixed-effects model fit to summed residual log response latencies in the predicate region in Experiment 4.

Effect	Posterior $\hat{\beta}$	Posterior σ_β	95% CRI Lower	95% CRI Upper
Intercept	0.02	0.04	(-0.06, 0.11)	
Disambiguator (Late)	0.03	0.03	(-0.02, 0.09)	
Meaning (*each*)	-0.06	0.02	(-0.10, -0.02)	
Disambig. × Meaning	0.03	0.02	(-0.01, 0.07)	

Table 2.38: Conditional means and measures of spread for the first word of the spillover region in Experiment 4. Standard errors are reported over raw response latencies, and bootstrapped 95% confidence intervals are reported over the critical measure, residualized log response latencies.

Meaning	Position	RT	SE	Resid. Log RT	95% CI
together	Early	902	24	-0.05	(-0.09, -0.02)
together	Late	827	17	-0.12	(-0.15, -0.09)
each	Early	854	25	-0.09	(-0.13, -0.06)
each	Late	957	30	-0.01	(-0.05, 0.03)

reflecting that in conditions with pre-verbal disambiguation this region was read faster after *each* than after *together*, $\hat{\delta}$ = -0.06, $P(\delta < 0)$ = 0.97, while in conditions with post-verbal disambiguation, this region was read slower after *each* than after *together*, $\hat{\delta}$ = 0.10, $P(\delta > 0)$ = 0.99. This is consistent with both major patterns observed in other regions, in which the computation of collective readings seemed to be more difficult than distributive readings, in the pre-verbal baseline, but post-verbal distributive disambiguation comes with a particular cost.

Bayes factor analysis suggests extreme evidence for the predicted interaction (BF_{10} = 135.61). These patterns remain consistent with the presence of the predicted reanalysis costs associated with late distributive disambiguation.

2.7.2.4 Full Spillover

The model fitted to summed residual log latencies across the full spillover region is reported in Table 2.41. Unlike the first spillover, I observe a near-credible main effect of

Table 2.39: Bayesian linear mixed-effects model fit to residual log response latencies on the first word of the spillover region in Experiment 4.

Effect	Posterior $\hat{\beta}$	Posterior σ_β	95% CRI Lower	95% CRI Upper
Intercept	-0.06	0.02	(-0.10, -0.02)	
Disambiguator (Late)	0.00	0.01	(-0.02, 0.03)	
Meaning (*each*)	0.01	0.01	(-0.01, 0.03)	
Disambig. × Meaning	0.04	0.01	(0.02, 0.06)	

Table 2.40: Conditional means and measures of spread for the full spillover region in Experiment 4. Standard errors are reported over summed raw response latencies, and bootstrapped 95% confidence intervals are reported over the critical measure, summed residualized log response latencies.

Meaning	Position	Sum RT	SE	Sum Resid. Log RT	95% CI
together	Early	2768	52	-0.15	(-0.23, -0.08)
together	Late	2750	52	-0.19	(-0.26, -0.12)
each	Early	2626	45	-0.26	(-0.33, -0.20)
each	Late	2832	57	-0.09	(-0.16, -0.01)

disambiguator position, such that this region received slower latencies when following a predicate with post-verbal disambiguation, $\hat{\beta}$ = 0.03, 95% CRI = (-0.00, 0.07). This was driven by a continued cost for late *each* (vs. early *each*), $\hat{\delta}$ = 0.19, $P(\delta > 0)$ = 0.99, while patterns for *together* reflected the opposite effect, $\hat{\delta}$ = -0.05, $P(\delta > 0)$ = 0.23. This was reflected in a positive interaction between disambiguator position and meaning, $\hat{\beta}$ = 0.06, 95% CRI = (0.02, 0.10), continuing the pattern from the first position of the spillover. Bayes factor analysis suggests moderate evidence for the predicted interaction (BF_{10}=8.05).

2.7.3 Discussion

The results of Experiment 4 demonstrate broad evidence for the specification of collective vs. distributive interpretations during the processing of verbs with plural subjects. Critical evidence comes from extra costs associated with late distributive disambiguation that emerge after *each* follows a predicate, which I interpret as a distributivity garden path. These patterns match what was found by Frazier et al. (1999) in eyetracking,

Table 2.41: Bayesian linear mixed-effects model fit to summed residual log response latencies in the full spillover region in Experiment 4.

Effect	Posterior $\hat{\beta}$	Posterior σ_β	95% CRI Lower	95% CRI Upper
Intercept	-0.10	0.03	(-0.17, -0.03)	
Disambiguator (Late)	0.03	0.02	(-0.00, 0.07)	
Meaning (*each*)	-0.02	0.02	(-0.07, 0.02)	
Disambig. × Meaning	0.06	0.02	(0.02, 0.10)	

and Dotlačil and Brasoveanu (2021) in self-paced reading, but notably mismatch the absence of such effects for these stimuli in my self-paced reading data from Experiment 3. Whatever the reason for the absence of reanalysis effects in Experiment 3, their presence here is another indication of motivated early commitment in the Maze.

Of secondary interest, I also replicated an unexpected difference between collective and distributive interpretations of these predicates: on the predicate and in the spillover region, I found evidence that constructing a collective reading incurs particular costs. This is at odds with preferences of interpretation when the distinction is ambiguous: in my norming data, collective readings were by far preferred to distributive readings. And response latencies in the Maze suggested that collective readings were often considered first, prompting a high likelihood of reanalysis when late distributive disambiguation was encountered. Rather than costs of selecting a collective meaning against a distributive bias, this slowdown might diagnose independent costs of interpreting a collective meaning. This is interestingly at odds with hypothetical structural complexity: in formal semantics, it is generally the distributive reading which is treated as derived, and the collective as basic. I do not challenge this structural proposal, but suggest that collective readings may be independently costly for other reasons more closely related to interpretation, such as the effort required to represent a pair of individuals acting as a group. More work with additional comparisons would be necessary to evaluate this kind of hypothesis. Regardless of the nature of the cost, I note that this is an interesting case of an apparent interpretive default towards a seemingly more difficult computation.

2.8 General discussion

Evidence from the processing of polysemy and distributivity in the Maze task supports the hypothesis that task motivations can motivate earlier commitments to lexical and relational meanings during incremental comprehension. I conclude by highlighting open questions about the nature of deferred selection, before discussing considerations for a theory of the interface between these task motivations and interpretive decision-making, and connecting the present results to work on how other factors of the environment may influence processing behavior.

2.8.1 Deferring decisions

I have concluded in this chapter that comprehenders somehow avoid making an immediate commitment about a particular aspect of meaning during incremental processing in a few cases, particularly polysemous nouns and distributivity ambiguities, in the absence of informative contexts, in standard (non-Maze) comprehension tasks. Evidence for this deferment comes from the absence of canonical garden-path and subordinate selection effects; that is, we have evidence for these cases that comprehenders do not select a single analysis.

We have less evidence about what exactly they do in the absence of selection, a question which has been of critical interest in the literature on the interpretation of polysemy since Frazier and Rayner (1990). The most plausible answers are either that (i) comprehenders do activate all maximally-detailed meanings as usual but do not select between them until a later date, or (ii) comprehenders avoid generating a plurality of meanings by adopting only a less-detailed underspecified representation (see especially Frisson and Pickering, 1999 and subsequent review in Frisson, 2009). The data reviewed here for polysemy cannot distinguish between these possibilities, although I add to a body of work that demonstrates that at least when a polyseme receives a fully specified interpretation, its senses do compete and interact much like the meanings of a homonym (Klein & Murphy, 2001; Klepousniotou et al., 2008; Foraker & Murphy, 2012; Lowder & Gordon, 2013; Fishbein & Harris, 2014; Brocher et al., 2016, 2018). On the other hand, the observation of deferred interpretation for distributivity may not lend itself so easily to an underspecification-type account, at least in that formal semantics has not generally proposed representations which are compatible with both collective and distributive interpretations, although see Schwarzschild (1994) for one exception.

Although they cannot be distinguished here, temporary parallel consideration and underspecification make different predictions about the availability of possible interpretations during the period of deferment before selection takes place. The latter chapters of this dissertation will present a better opportunity to probe for the rapid consideration of those interpretations in other cases of deferred selection, and largely concur that possible interpretations do become available to the comprehender far in advance of selection, ruling against an underspecification account for at least those phenomena. I will not rule on whether rapid generation of meanings in also occurring in the cases of this chapter, but hope that it might be investigated further in future work.

2.8.2 Risk-taking in strategic decision-making

For both of the linguistic phenomena studied here, I have provided evidence that, depending on the demands of a task, comprehenders may violate a typical heuristic for minimal effort. This heuristic has been proposed to explain the online underspecification of polysemes in neutral contexts, replicated in Experiment 1, and it plausibly underpins the lack of online commitment to the fine details of event structure in my Experiment 3. The intuition I have advanced, taken from Forster et al. (2009), is that the violation of that heuristic in my Maze experiments is crucially connected to the utility of a fixed interpretation for leftward context.

I posit the relationship between minimal commitment and the utility of full commitment can be better understood if we re-frame minimal commitment from a fixed heuristic to a derived optimization of cost and benefit within a given environment. Incremental parsing of any stimulus which is perceived through time poses a difficult problem (Hale, 2011). Any representational commitments which are made before the offset of the stimulus run the risk of being incorrect and requiring costly revision of some sort. On the other hand, delaying the representation of early portions of the stimulus while waiting for more information can be costly, because surface-level features of the input may be difficult to maintain and integrate at a long distance. A rational parser, then, should find some medium between eagerness and reticence, choosing to risk the formation of a representation from incomplete evidence only in order to avoid the potential failures or costs of integration at a long time delay.

I suggest that minimal commitment during incremental comprehension can be derived as a result of this typical state of affairs. While (some) grammatical representations, like the distinction between two distinct lexical items, offer enough utility that they are worth taking a risk on, more abstract interpretive distinctions, like the difference between two senses of a polyseme, may require reanalysis more often than they are worth. This is a more flexible hypothesis than a traditional linguistic/post-linguistic interpretational divide, and so it should be adopted cautiously, but by doing so here, we may come to a better understanding of the role of task effects.

Under this account, task demands will not encourage comprehenders to violate a stored heuristic, but instead can be understood to merely change the environment that determines our calculations of cost and benefit. The Maze task introduces exceptional costs for delayed decision-making, and exceptional benefits for early commitments. To the first point, it seems certain that the dynamics of the forced-choice decision task slow participants down and require participants to divert cognitive resources towards task-related decision-making, and so information about the stimulus may decay more rapidly than in everyday comprehension. To the second point, as discussed by Forster et al. (2009), the decision-making portion of the Maze is intuitively aided by a maximally informative leftward context. Decisions to commit to an interpretation for all incoming content are one way to maximize the information contained in that context. The combination of these factors can explain why the scale seems to tip towards commitment earlier in the Maze than typical reading.

Nevertheless, I note that neither of these claims has been independently validated. More work examining the precise motivations and demands instantiated by the Maze will be necessary to develop a complete understanding of the pattern of early commitment observed here.

Looking towards the future, I expect that more detailed models of participants' adaptation to this new environment, perhaps specified using domain general approaches to rational sequential decision-making like reinforcement learning (Sutton & Barto, 2018), will be of use in further theorizing. Even a simple computational model could help generate testable predictions, not only about the optimal timing for parsing decisions in various kinds of environments, but also about the dynamics of how participants progress towards that hypothetical local optimum.

2.8.3 Cross-linguistic and individual differences in comprehension

Cast in this way, we can see task-related differences in the timecourse of representational decision-making as part of a larger family of related effects on processing decisions from what may be described as the "parsing environment," broadly construed. For instance, distinctions across languages in the time course of representational decision-making are well known: e.g., different levels of phonological systematicity across English, French, and Japanese have been argued to induce different strategies for lexical segmentation and access across their native comprehenders (Cutler et al., 1986; Cutler & Norris, 1988; Otake et al., 1993). In the same vein, and closer to the focus of this study, recent work on the time course of aspect commitments across languages has uncovered language-by-language variation in the point at which comprehenders will generate a firm representation of a verb's aspect. Across several reading studies (O. Bott, 2013; O. Bott & Hamm, 2014; O. Bott & Gattnar, 2015), O. Bott and colleagues have demonstrated that in Russian, where aspect is morphologically specified to a high degree of precision, decisions about, e.g., telicity are made directly within the verb region, whereas in English, where there are fewer explicit cues, decisions are postponed until all arguments of the verb have been encountered, and in German, where telicity in the simple past is entirely unmarked, decisions can be postponed until the sentence boundary. We might reasonably consider all of these to be strategies in service of optimizing the trade-off between risk and representation.

Work on individual differences in low-level processing behaviors has also engaged with the question of adaptation to a "parsing environment," in particular, adaptation to the changing availability of cognitive resources across the human lifespan (see Stine-Morrow et al., 2006, for a comprehensive review). For instance, Rayner et al. (2006) observed that older adults are more likely to skip words during comprehension. Ruling out explanations related to oculomotor control, they concluded that older readers are more willing to take the risk of incorrectly representing the sentence and incur potential rereading costs later, perhaps to compensate for age-related differences in working memory span (Soederberg Miller & Stine-Morrow, 1998).

With further concerted efforts in theorizing and modeling, I see the potential to integrate the findings in these other literatures with the task effects evidenced in this study, in service of a general model of processing optimization in a given environment.

Chapter 3

Considering and cancelling scalar implicatures

Formal linguistic theory makes a split between strings which afford multiple interpretations because of representational ambiguity and strings which afford multiple interpretations only because of uncertainty about the intended message. In this chapter, I will discuss the processing of a paradigmatic case in the latter camp, assertions featuring the quantifier *some*. The multiple meanings of these assertions have a different nature from e.g. the homonyms and polysemes discussed in the last chapter. Nevertheless, I consider this to be another case of indeterminate meaning, and so I ask similar questions about the status of that meaning during incremental comprehension (when do we select a single interpretation?), and generally seek to explore this as a test case to better understand how comprehenders solve the general problem of decision timing.

Main contributions come from a detailed literature review and new evidence from two experiments using self-paced reading and the Maze. Although there is some debate in the literature, I will assert that known patterns here seem to reflect a process of Rapid Consideration Without Selection, such that comprehenders postpone a deterministic analysis of *some* as long as they can, even as they use the likelihood of different interpretations to reason about upcoming input. My own evidence is broadly in line with this picture, finding no evidence for the garden-path effects and subordinate selection costs which would be predicted if a single meaning were selected immediately. Given that comprehenders appear to defer selection, evidence for the rapid availability of distinct interpretations suggests that comprehenders may generate and maintain multiple interpretations in parallel at some stages of comprehension.

A canonical assertion featuring *some* is given in (5). *Some*, like all quantifiers, relates the set of individuals who satisfy the predicate in its restrictor (things which are *bottles of hand sanitizer*) to the set of individuals who satisfy the predicate in its nuclear scope (things which are *scented*). On one interpretation, a sentence like (5) asserts that there is at least one member of the restrictor who satisfies the nuclear scope; i.e. at least some of the bottles of hand sanitizer are scented. I'll call this, as is standard, the "lower-bound(ed)" reading. On another interpretation, the sentence can be taken to have a stronger meaning, asserting not only that there is at least one member of the restrictor who satisfies the nuclear scope, but also that not all of the members of the restrictor satisfy the nuclear scope. I.e. merely some of the bottles of hand sanitizer are scented. This interpretation has not only a lower bound (at least one), but also an upper bound (not all); I'll call it the "upper-bound(ed)" reading.

(5) Some of the bottles of hand sanitizer are scented.

Since Grice (1975), one standard analysis of such cases is that *some* entails the weaker, lower-bound reading only, and the stronger, upper-bound reading can be derived based on the principles of well-formed communication. Because speakers typically adhere to a pressure to be as informative as possible, when a speaker makes a notably weak assertion, they "implicate" from their avoidance of a stronger alternative (i) that they do not believe the stronger assertion would have been true, and from here, generally, (ii) that they believe the stronger assertion is not true. As a result, an upper-bound meaning can be derived without positing any ambiguity about the lexical meaning of *some*. Crucially, this makes the correct prediction that when certain features of the context disrupt this derivation, the upper-bound meaning is less likely—e.g. when the stronger alternative isn't relevant to the conversation (e.g. Zondervan et al., 2008), or when the speaker might lack the evidence to make the stronger assertion (e.g. Goodman and Stuhlmüller, 2013), among other cases (Degen, 2015). The upper bound for *some*, together with a variety of other cases where weak lexical items implicate the negation of a stronger scalar associate, is classed in particular as a "scalar implicature".

The major alternative to this analysis agrees that there is an optional strengthened upper-bound meaning for weak scalar items like *some*, derived by negation of the stronger alternative *all*, but contends that this takes place within the semantic interpretation proper of sentences like (5), where comprehenders may postulate a silent operator responsible for negating certain alternatives of its prejacent (e.g. Chierchia, 2004; Fox, 2007). While this is associated with a variety of different formal claims, it is worth noting, as do Chemla and Singh (2014a, 2014b), that a semantic theory of scalar implicature ultimately still imagines that a comprehender must engage in pragmatic reasoning to decide between a simple interpretation contingent on the lexical semantics of a weak scalar, and an enriched interpretation derived by some additional mechanism. For the cases I will focus on in this chapter, differences in the predictions for processing accounts depend more on linking assumptions between competence and performance than the choice of a Gricean vs. null-operator analysis.[1] In the chapter, I will sometimes discuss the upper-bound meaning as if it is uncontroversially the output of Gricean reasoning, but unless otherwise noted, my conclusions would be similar if working within a null-operator approach. I stress that I see nothing about the results I will discuss that argues directly for one competence theory over the other.

The chapter will proceed as follows. First, in the following section (§3.1), I will present an overview of the literature on the processing of *some*, arguing that a coherent picture emerges if we allow for the idea that awareness of a potential upper-bound meaning may drive expectations in online comprehension long in advance of a costly and resource-dependent selection process. This approach, dubbed "Rapid Consideration Without Selection," helps make sense of an asymmetry observed in Bergen and Grodner (2012), where comprehenders seem to derive facilitatory expectations from the consideration of an upper-bound meaning without demonstrating difficulty when an upper-bound meaning must be rejected. After an interlude where I explore what kinds of expectations could derive this asymmetry (§3.2), I present a design for a reading study to shore up our empirical understanding of the asymmetry (§3.3), which is then carried out in Experiment 5 in self-paced reading (§3.4) and Experiment 6 in the Maze (§3.5). The continued absence of reanalysis effects, even in the Maze, suggests that online selection of enriched meaning is indeed avoided for *some*, and in section 3.6 I discuss and relate this to the task-dependent reanalysis patterns observed in Chapter 2 before concluding.

[1] This is particularly the case because I do not follow the assumption that a Gricean account of upper-bound meanings should expect upper-bound meanings to require a multi-step derivation over complete propositional meanings before they can be considered during incremental comprehension, see discussion in Chemla and Singh (2014a, 2014b) and §3.1.3.

3.1 The processing of scalar implicature

The existing literature on the timecourse of *some* enrichment has focused largely on questions of the cost and immediacy associated with the enriched meaning: is the upper-bound meaning of *some* available for a comprehender to consider immediately, or does it emerge only as the result of a costly enrichment process? While early and influential evidence from judgment tasks was taken to diagnose a late, costly enrichment process, evidence from event-related potentials and visual world eyetracking experiments demonstrates that the enriched meaning can nevertheless rapidly influence lexical integration and referential prediction. In this section, I will review the status of this literature, attempting to make sense of the apparent contradictions by making a distinction which is often collapsed in this work. I will assume that the process of arriving at an upper-bound interpretation for *some* requires both the generation of an upper-bound meaning, and the selection of that upper-bound meaning. While comprehenders seem to generate enriched meanings quite early, the available evidence is consistent with a delay in selection, i.e. a period of online indecision, before a decision process where inhibiting the un-enriched meaning requires time and executive resources. Reading time experiments provide a crucial role in testing the predictions of this account, motivating the present study using reading methods to study cancellation.

3.1.1 Delayed influence of enriched meanings in judgment tasks

One of the simplest ways to examine the timecourse of *some* enrichment is to observe the dynamics of question responses that require the comprehender to select the upper-bound meaning. The method follows a standard subtractive logic: if responses which require the upper-bound meaning exhibit different patterns than responses which do not require the upper-bound meaning, we can take those differences as the signature of some process related to enrichment. Indeed, studies of this type, beginning with L. Bott and Noveck (2004), have uniformly observed that responses associated with an upper-bound interpretation of *some* are slower to arrive, and more dependent on the availability of executive resources, than responses associated with an unenriched, lower-bound interpretation.

Data from these tasks has generally been used to argue that upper-bound meanings emerge after complete comprehension of the stimulus via a costly additional step of Gricean reasoning. Here, I will argue for the existence of an alternative interpretation, which takes costs associated with enriched meaning to diagnose late, costly selection rather than late, costly generation.

The exact mechanisms of the designs in these studies have varied widely, but all involve administering binary judgments of "true" vs. "false" (or "felicitous" vs. "infelicitous") to a sentence, either using world knowledge or a visually-represented context. For instance, in L. Bott and Noveck (2004), participants were instructed and trained to use only an upper-bound or a lower-bound interpretation of *some* as they provided a truth value for critical sentences like (6a) or various control sentences which would be insensitive to the manipulation (6b-f)[2]. Rejections of critical sentences in the upper-bound training condition arrived almost 600ms later than acceptances in the lower-bound training condition, or indeed other correct rejections of control sentences (see also van Tiel et al., 2019 for similar results in a training paradigm with visual context verification). Participants in the upper-bound training condition were also quite inaccurate on the critical sentences, only achieving rejection rates of 60%, while their performance on controls matched the other group (80–90% accurate). L. Bott and Noveck (2004) also find the same delays when comparing free, untrained responses to the same prompts: rejections of critical sentences (consistent with an upper-bound meaning) are associated with slower responses than acceptances (consistent with a lower-bound meaning), a response time asymmetry which does not hold in control conditions (see also van Tiel and Schaeken, 2017 and van Tiel et al., 2019 for similar results in a free-response paradigm with visual context verification).

(6) **Stimuli from L. Bott and Noveck (2004)**

a. Some elephants are mammals.

b. Some mammals are elephants. (True)

c. Some elephants are insects. (False)

d. All elephants are mammals. (True)

e. All mammals are elephants. (False)

f. All elephants are insects. (False)

For theoretical reasons, it was appealing for L. Bott and Noveck (2004) and subsequent researchers to attribute these costs and resource-dependent outcomes to the stepwise so-

[2] The task itself was completed in French. The stimuli listed are the authors' English translations of their material.

cial calculations hypothesized to underlie a conversational implicature. Nevertheless, one should hesitate to take response time evidence as bearing on the interpretive process *per se*. The process of rejecting a *some* statement which is inconsistent with an upper-bound meaning indeed depends on the generation of an upper-bound meaning, but it also depends on the selection of that meaning to the exclusion of the lower-bound meaning (which would lead to a different, positive response). Without further evidence, we cannot be sure which of these stages is the source of the difficulty observed here.

And whether difficulty should be attributed to generation or selection, slower response times and reduced accuracy alone are not sufficient to demonstrate the presence of an additional costly processing step. Participants may also be strategically trading off on speed for improved accuracy due to differences in difficulty within a single process. These possibilities can be teased apart with a deadline procedure, wherein participants' responses are demanded at specific points along the temporal region of interest, allowing researchers to model the time-accuracy function in detail (Reed, 1973). McElree (1993) used this paradigm to argue that delays in response time associated with dispreferred argument frame were due to variable difficulty instead of an extra stage of serial processing. Although average response times in a grammaticality judgment task were slower when a frequently-transitive verb was presented in an intransitive frame, participants were able to respond correctly at short intervals when forced to, incompatible with a model where argument frames are checked serially, one-at-a-time. In comparison, in McElree and Griffith (1995), the paradigm provided evidence for a two-stage model where some information influences decision-making at a delay. For responses in a grammaticality judgment task associated with thematic violations, as compared to simple syntactic violations, accuracy at the earliest deadlines was sharply reduced, consistent with later-initiating and slower-accumulating processing dynamics.

To better determine whether later enrichment responses were indeed the result of an obligatory temporal delay, L. Bott et al. (2012) applied a deadline procedure to examine the time-accuracy function for trained upper-bound vs. lower-bound readings of *some* in a world-knowledge verification task. Their results support the presence of a delay, comparable to McElree and Griffith (1995): participants begin exhibiting accurate truth-value judgment responses for lower-bound *some* a few hundred milliseconds before upper-bound *some*, consistent with later-initiating and slower-accumulating processing dynamics. An early predominance of incorrect acceptance of critical sentences suggest that during this lag, the upper-bound-trained participants were responding based on an initial lower-bound interpretation. Indeed, mouse-tracking data from in the same design shows that participants initiate movements towards a lower-bound-consistent response in this window (Tomlinson et al., 2013).

Convergent evidence for a costly path towards the upper-bound reading comes from dual-task paradigms, where participants performed world-knowledge or depicted-context verification under cognitive load. For instance, De Neys and Schaeken (2007) had participants memorize a complex dot pattern (Bethell-Fox & Shepard, 1988) before each world-knowledge verification trial, known to demand executive functioning resources (Miyake et al., 2001). Participants were prompted to reproduce the dot pattern after their verification response. The authors observed a small but significant reduction in upper-bound-consistent responses from 79% of easy-pattern control trials to 73% of critical harder-pattern trials, while control conditions were unaffected (see also Dieussaert et al., 2011, and replication with depicted contexts in Marty et al., 2013; Marty and Chemla, 2013; van Tiel et al., 2019). Such evidence that comprehenders arrive at enriched meanings for *some* less frequently when executive function is taxed has been taken to as evidence that executive function is somehow required to derive the enriched meaning.

Pursuing the hypothesis that this is a generation-level effect, researchers attempted to locate exactly which component of generation was responsible for these costs. One attractive possibility was that enriched, upper-bound meanings engender costs due to the relative complexity of the upper-bound meaning itself. While a lower-bound meaning for *some* requires only the existence of some element of the restrictor which satisfies the nuclear scope, an upper-bound meaning requires the existence of a complement set whose members do not satisfy the nuclear scope. Another possibility was that such costs were associated with the hypothetical steps of an implicature derivation, in which an alternative to *some* must be identified (*all*) and negated (*not all*) before being added to the sentential meaning (*some but not all*). Natural language affords a useful comparison case in the string *only some*: here, the same meaning is instantiated (requiring the existence of a complement set with members for whom the nuclear scope is false). Moreover, the same derivational steps, by reasonable assumption, should underlie the comprehension of *only some*: *only* is generally taken to be a focus particle associated with a prejacent linguistic expression, and entail that all contextual alternatives to the prejacent would make the utterance false (e.g. Rooth, 1985). When the prejacent is *some*, then, comprehension presumably involves identification and negation of the *all* alternative.

Nevertheless, response behaviors reflecting difficulty for pragmatically-derived upper-bounded *some* can be observed above and beyond the costs associated with *only some*. In the second response deadline experiment reported in L. Bott et al. (2012), Bott and colleagues observed that accurate judgment responses which are dependent on trained upper-bound *some* are associated with a short delay compared to accurate responses dependent on *only some*. Likewise, Marty and Chemla (2013) found that upper-bound responses for *only some* were not reduced under cognitive load in the same way as those in the critical *some* condition. These costs suggest there is difficulty associated with the pragmatically-enriched meaning in judgment tasks above and beyond the calculation of an upper-bound meaning through the identification and negation of an alternative quantifier.

The major differences that remain between a pragmatically-enriched *some* and *only some* are twofold: (a) the theory-internal difference between strengthening via reasoning about the act of communication and strengthening via the entailments of *only*, and (b) the context-dependence of the strengthened meaning; e.g. pragmatic enrichment is context-dependent and prompts comprehender uncertainty.[3] That is, in the face of the evidence of some delay specific to pragmatically-enriched *some*, it would seem to be either an effect of costly generation through social reasoning, or an effect of costly selection due to comprehender uncertainty. Without particular evidence that the delay is associated with communicative reasoning, the latter seems like a more attractive explanation for the remaining variance (Marty and Chemla, 2013; see also Khorsheed et al., 2022 for useful discussion).

Nicely, this explanation affords an operationalization quite similar to the subordinate selection effect on the reading times of biased homonymous nouns in contexts which prompt an infrequent interpretation (see Pacht and Rayner, 1993 for a review, and also discussion in §2.2). A context-dependent meaning for an ambiguous string may engender cognitive effort simply as a result of a conflict between a pre-contextual bias for one meaning and the use of contextual cues.[4]

[3] Of course, (a) only applies on a Gricean theory of scalar implicature, and not silent-operator theories, which are largely indistinguishable from positing a silent *only*. Difference (b) holds regardless of your preferred account of implicature.

[4] This parallel is probably overly strong, as there may be unprobed differences in the dynamics of the

Table 3.1: Summarizing generalizations and conclusions from the judgment literature on *some*-enrichment.

Generation of UB meaning		**Selection of UB meaning**	
Timing	*Difficulty*	*Timing*	*Difficulty*
Verification responses are delayed when consistent with upper-bound meaning. (e.g. L. Bott and Noveck, 2004; L. Bott et al., 2012; Tomlinson et al., 2013)			
-	-	When prompted	Entails a slower decision process
Verification responses consistent with upper-bound meaning are reduced under cognitive load. (e.g. De Neys and Schaeken, 2007; Dieussaert et al., 2011)			
-	-	When prompted	Requires executive resources
Delays, resources required for *only some* verification are smaller. (e.g. L. Bott et al., 2012; Marty and Chemla, 2013)			
-	-	-	Time, resources are required due to LB inhibition

Adopting this interpretation of the truth-value judgment studies reviewed above, I summarize the relevant generalizations and associated conclusions in Table 3.1. Notice that, given this interpretation, we would take these tasks as evidence only for the difficulty of post-stimulus selection of the enriched meaning. Further evidence that a late-selection approach is preferable to a late-generation approach will come in the next section.

3.1.2 Earlier influence of enriched meanings in lexical integration and reference prediction

Indeed, in contrast to the original interpretation of the studies above, evidence from event-related potentials (ERPs) and visual world reference-resolution experiments suggests that comprehenders in fact rapidly activate and exploit enriched meanings during online comprehension. The contrast is most apparent in a pair of ERP studies which examine stimuli of the same form as L. Bott and Noveck (2004): Nieuwland et al. (2010) and Hunt et al. (2013). Both studies examined the envelope of the N400 response, taken to index lexical integration difficulty compatible with the absence of facilitatory expecta-

difficulty. E.g. metonymy in biased contexts admits a subordinate access effect much like homonymy (Lowder and Gordon, 2013, see also Ch. 2 of this dissertation), and yet L. Bott et al. (2016), in another deadline study, observe no evidence that dominant (literal) meanings of metonyms precede subordinate (figurative) meanings, in contrast to the delay for upper-bound *some* observed in L. Bott et al. (2012).

tions, on a critical noun of an assertion featuring *some*. Nieuwland et al. (2010) observed larger N400 responses for nouns in the nuclear scope which are incompatible with upper-bounded *some* (i.e. nouns which would only generate world-knowledge-consistent assertions with lower-bounded *some*) (7a), compared to those which are compatible with the upper-bounded meaning (7b).

(7) **Stimuli from Nieuwland et al. (2010)**

a. Some people have **lungs**... (inconsistent with upper-bounded *some*)

b. Some people have **pets**... (consistent with upper-bounded *some*)

Hunt et al. (2013) supplemented the earlier findings of Nieuwland et al. (2010) with visual contexts, allowing them to keep the linguistic signal identical, but manipulate truth in context across conditions. On stimuli where the critical noun is the restrictor of partitive *some*, they replicated the critical N400 when the noun creates an assertion which is incompatible with the upper-bounded reading of *some*. But crucially, they also compared the ERP on this noun when it creates an assertion which is incompatible with either reading of *some*. The authors observed a stairstep pattern: totally incompatible nouns yielded the largest N400 response, followed by the upper-bound incompatible nouns, followed by the upper-bound compatible nouns.

(8) **Stimulus from Hunt et al. (2013)**

The student cut some of the **steaks**...

(context: only some steaks cut, all steaks cut, or no steaks cut)

In both studies, the critical N400 on upper-bound inconsistent stimuli is found to be variable between participants. Nieuwland et al. (2010) observe the effect only for participants highly skilled in pragmatic reasoning (scoring highly on the Communication scale of the Autism-Spectrum Quotient questionnaire, Baron-Cohen et al., 2001), and Hunt et al. (2013) observe the effect only for participants whose offline verification responses are consistent in their preference for the upper-bounded meaning of *some*.

On the whole, ERP results suggest that at least some comprehenders consider the potential enriched meaning of *some* early enough to make upper-bound consistent predictions about critical nouns a few positions later. In particular tasks, comprehenders may even activate upper bound meanings as part of the out-of-context lexical entry of *some*. Barbet and Thierry (2018) report the results of what we might call a “homogeneity

Stroop task," where participants had to respond "Yes" when all letters of a presented word were uppercase. ERP envelopes revealed a Stroop-like N450 response, thought to index the presence of conflicting information, when participants were presented with the word *all* in mixed case, as would be expected if the lexical meaning of *all* were competing with the necessary response of *not all*. However, the authors observed the same N450 response when participants were presented with the word *some* in uniform uppercase, as would only be expected if the lexical meaning of *some* were competing with the necessary response of *all*. The authors note that this would be unexpected if early processing of *some* activated only the lower-bounded meaning, as this has no necessary conflict with an *all* response. This would seem to indicate that participants in the task associated the word *some* immediately with a potential upper-bounded meaning, although we might hesitate to compare this task directly with incremental sentence comprehension.

In comparison to the consistency of delayed costly judgment responses on one hand, and early anticipation in the ERP record on the other hand, studies which examine reference prediction effects in visual-world eyetracking have been the source of many conflicting conclusions, and remain a hotbed of debate for researchers interested in the timecourse of pragmatic reasoning. I will argue here that these results are most in line with the relatively rapid online use of upper-bound meaning to anticipate reference. Y. T. Huang and Snedeker (2009a, 2011), pioneering the use of this method in this literature, provided early evidence that comprehenders had only delayed access to upper-bounded meanings. They tracked comprehenders' gaze between two individuals that could be described by the head noun of a relative clause inside a definite nominal expression, a scenario where the relative clause is presumed to provide a further description which uniquely picks out a target individual. Relative clauses then contained a critical quantifier or numeral (9). In their critical condition, the stimulus described an individual that had *some* of a restrictor. Until comprehenders heard the restrictor, the instructions would be compatible with two individuals, *e.g.* a girl who had all of the soccer balls, and a girl who had only some of the socks. By hypothesis, as soon as participants can relate *some* to its alternative *all* and compute the potential enriched meaning given use of the weaker alternative,[5] they should begin anticipating reference to the girl who has only some of the socks (cf. similar effects of anticipatory use of prenominal adjectives in Sedivy et al., 1999 *et sequens*). In control

[5] Note that this enrichment is not a scalar implicature in the strictest sense, although it would be reasonable to derive it from Gricean reasoning over potential descriptions given the pressure to describe a unique referent. See Breheny et al. (2013) for further discussion.

conditions, the unambiguous meaning of *all*, the upper-bounded meaning of *two*, or either upper- or lower-bounded meaning of *three* was sufficient to pick out a unique individual. Gaze data revealed that participants anticipated the target referent equally in all conditions besides *some*, where no pre-restrictor anticipation effect was apparent (Y. T. Huang & Snedeker, 2009a) until the ambiguous region was extended in a follow-up study, revealing a delay of approximately 800ms (Y. T. Huang & Snedeker, 2011).

(9) **Stimuli from Y. T. Huang and Snedeker (2009a)**

Point to the girl that has {some, all, two, three} of the socks.

However, early studies from other research groups, using slightly different designs, were unable to replicate this evidence for delayed use of the upper-bounded meaning, reporting near-immediate anticipatory looks equivalent to *all* conditions (Grodner et al., 2010; Breheny et al., 2013). In subsequent attempts to probe this inconsistency in results (Degen & Tanenhaus, 2016; Y. T. Huang & Snedeker, 2018), researchers demonstrated that delays in anticipatory looking depend on the presence of trials where the distribution of critical objects is referred to using expressions other than *some* and *all* (i.e. the use of numerals). Degen and Tanenhaus (2016) suggested that the apparent delay is the result of flexible constraint-based processing of *some*, whereby the ease of interpreting *some* in the intended way is depressed by the presence of salient alternatives besides *all* (i.e. the numerals). Y. T. Huang and Snedeker (2018) weighed this explanation against another possibility, that in studies with little referential variation, participants may have pre-associated depicted individuals with partial sets and the expression *some*. In follow-up naturalness and production experiments, the latter authors challenged the offline evidence supporting the Degen and Tanenhaus (2016) explanation, demonstrating that *some* remained felicitous and frequent for their participants even when numeral alternatives are salient in context. Their evidence thus suggests that the Grodner et al. (2010) and Breheny et al. (2013) null findings are best thought of as showing task-specific facilitation of the enriched meaning of *some*, and the delays observed in other studies (Y. T. Huang & Snedeker, 2009a, 2011; Degen & Tanenhaus, 2016; Y. T. Huang & Snedeker, 2018) are characteristic of a slower, real-time process of enrichment.

Nevertheless, a final recent development in the literature has once again raised the possibility that the oft-observed delay may still be the result of task effects. Sun and Breheny (2020) reasoned, in line with some evidence in Degen and Tanenhaus (2016), that the discrepancy between *some* and *all* could be an illusion caused by facilitated processing of *all* in the Huang and Snedeker design. For instance, as they demonstrated in an initial offline judgment experiment, comprehenders have a basic preference for the use of *all* with large sets (like the target set in the *all* condition in e.g. Y. T. Huang and Snedeker, 2009a) and a basic dispreference for the use of *all* with sets with a cardinality of two (like the competitor set in the *all* condition in e.g. Y. T. Huang and Snedeker, 2009a). When they eliminated these factors which may have facilitated the processing of *all* in that condition, they observed that both *all* and *some* conditions featured the same delay relative to the numeral conditions. They attribute this delay to the dependence of both target meanings on comparison between the domain of a quantifier and the set of individuals which satisfy the nuclear scope. (The useful upper-bounded readings of numerals don't require this effort in the same conditions, as matching referents can be identified merely through rapid subitization of the objects in their posession.) Furthermore, they argued that the upper-bound reading of *some* was considered quite rapidly by their participants, pointing to evidence that comprehenders in the *some* condition began checking the unpossessed members of potential domains just as rapidly as comprehenders in the *all* condition, information which would only be relevant for the upper-bounded meaning of *some*.[6]

This new evidence for early consideration of an upper-bound meaning for *some* brings the visual world and ERP effects somewhat into alignment. Both suggest that the upper-bound meaning of *some* is considered quickly in online processing, and begins influencing predictive processes within the next few words. The stair-stepped N400 responses of Hunt et al. (2013) suggest that upper-bound consistent continuations are anticipated more strongly than continuations consistent with only a lower-bound interpretation, but that even the latter continuations are expected to some degree when compared to completely infelicitous options. Likewise, the quantification-driven gaze patterns of Sun and Breheny (2020) suggest that comprehenders begin considering the possibility of an upper-bound meaning early, and can exploit its potential contrastive value for reference prediction as early as in controls featuring *all*—however, in neither case does this prediction show up to the overwhelming degree demonstrated in cases like Grodner et al. (2010), which we may be able to comfortably consider a figment of strategic task performance

[6] The lower-bounded meaning of *some* would be satisfied by the presence of any members of the domain among a character's possessions. If participants were only considering this meaning during their early processing, looks to the unpossessed objects, what Sun and Breheny (2020) called the "residual set," should be somewhat delayed.

Table 3.2: Summarizing generalizations and conclusions from the ERP and visual world literature on *some*-enrichment.

Generation of UB meaning		**Selection of UB meaning**	
Timing	*Difficulty*	*Timing*	*Difficulty*
N400 responses signal some integration difficulty for nouns that mismatch upper-bound meanings. (Nieuwland et al., 2010; Hunt et al., 2013)			
Influences lexical prediction within a few words	-	-	-
Anticipatory gaze patterns reflect use of upper-bound meaning around 800ms after *some*. (e.g. Y. T. Huang and Snedeker, 2011; Sun and Breheny, 2020)			
Influences referent prediction within 800ms, influences other looking behavior even faster	Requires the same multiple-set comparison as *all*	-	-

(Y. T. Huang & Snedeker, 2018). While consideration is rapid, no evidence suggests that dominance or selection of this meaning is immediate.

I summarize the relevant generalizations and associated conclusions in Table 3.2. Notice that this is taken as evidence regarding the timecourse of generation, unlike the generalizations in Table 3.1.

3.1.3 Reconciling late verification and early prediction

How, then, do we reconcile these two sets of observations? On the one hand, responses consistent with the upper-bound meaning of *some* in a truth-value judgment task are dependent on slow and effortful processing. On the other hand, upper-bound meanings of *some* can begin influencing lexical prediction and reference anticipation rather quickly. The solution cannot be a hypothesis of strategic variation in behavior, because we can, in fact, observe both effects in the same experiment. Degen and Tanenhaus (2016) noted that in their visual-world eyetracking experiments, responses to a truth-value judgment prompt that were consistent with an upper-bound meaning for *some* consistently arrived later than those which were consistent with a lower-bound meaning for *some*, even in experiments where participants' gaze patterns revealed early awareness and exploitation of an upper-bound meaning. Again, following Y. T. Huang and Snedeker (2018), it may be the case that any exceptionally early use of the upper-bound meaning was the consequence of verbal pre-coding for the depicted referents, but it is still notable that early online awareness of an upper-bounded meaning could co-exist with a response delay purported to diagnose a post-stimulus process of costly enrichment.

As suggested by Degen and Tanenhaus (2016), and hinted at above, this picture can be comfortably resolved if we consider generation of the enriched meaning and selection of the enriched meaning for a forced-choice response separately. If we trust the online evidence, a participant in the critical condition of a truth-value judgment study has indeed had the chance to activate an upper-bound meaning for *some* well before their response is elicited, but the response requires that they take into consideration only this upper-bound meaning. That is, if a lower-bound meaning has been activated at all—and the partial facilitation of continuations only compatible with a lower-bound reading in Hunt et al. (2013) suggests it has—it must be inhibited or ignored. If the lower-bound meaning is somehow more dominant, due to its context-independence, or weaker entailments, inhibiting it will be harder than whatever inhibition would be required to ignore an upper-bound meaning and select a lower-bound meaning. The inhibition and selection required by a response prompt would not necessarily have any expected fingerprint in the critical lexical prediction and referential anticipation evidence. While it is indeed reasonable that in the course of free sentence interpretation comprehenders will select a single meaning for a *some* utterance at some point, we have no evidence in the studies reviewed above that participants are engaging in such a commitment in the early online window under investigation.

On this approach, the major difference between what we might call a "pragmatically-dependent response cost" and the subordinate selection cost observed online for homonyms (Rayner & Frazier, 1989) is simply that for homonymy, the selection process which engenders the difficulty is carried out among potential lexical representations during immediate online processing, and thus visible in reading behavior, while for pragmatic enrichments, the selection process is carried out among potential sentence representations at some later stage, possibly only when cued, and thus most obvious in response behavior. To be clear, proposing late selection does not pretend that an enriched meaning is not rapidly under consideration during online reading: indeed, evidence suggests that it must be. It simply imagines that cognitively effortful selection between the lower-bound and upper-bound meaning, and the associated inhibition of whatever is not selected, is not an obligatory part of the immediate processing of *some*.

Before moving on, I'd like to clarify two points. First: I have adopted without much worry the claim that the upper-bound meaning for *some* may be considered without particular cost compared to similarly-complex meanings that don't arrive via enrichment, like *only some* or even *all*. I have considered that for the comprehender, this might essentially resemble the process of deciding between pre-stored meanings for a homonym. This is in sharp contrast with the picture expected in most experimental pragmatic work on the topic. After all, the substantive contribution of Grice was to take the burden of meaning multiplicity away from the grammar for cases like *some*. I do not question a Gricean derivation of the upper-bounded meaning, but I would like to point out that positing a Gricean derivation for the meaning does not entail that comprehenders must engage in social reasoning to postulate that meaning every time they encounter *some*. In general, a claim for a pragmatic source of meaning at the level of competence need not assume that meaning is gated behind a costly pragmatic process in performance. A pragmatically competent language learner can rely on a single logical meaning for *some*, while still exploiting their experience with the word to begin weighing its two possible interpretations immediately whenever they encounter it. Much the same point has been demonstrated in the processing of regular polysemy, where certain senses of a word can be derived from a "literal" meaning together with a rule—e.g. the "producer-for-product metonymy" of using "Dickens" to refer to that author's works—yet comprehenders can apparently access these meanings without extra derivational processing during online reading (Frisson and Pickering, 1999; L. Bott et al., 2016).

Secondly: although we don't see such a case here, it is possible for response time delays to co-occur with delays in the use of information during incremental comprehension. Take for instance the case of verbal thematic structure, where we seem to observe a universal delay in availability to all components of sentence processing. Just as thematic information seems to influence participant responses in a judgment task at some delay (McElree & Griffith, 1995), it has a delayed effect on, for instance, online predictive structure-building: the argument structure of a verb is apparently not accessed quickly enough to contravene predictions for an object position in e.g. filler-gap processing, so-called "hyper-active gap filling" (Omaki et al., 2015, *pace* Staub, 2007). We can thus imagine at least three situations where information appears to be accessed at a delay in response behavior: (a) the information is not associated with any cost, but simply a higher stochastic rate of failed retrieval—e.g. considering infrequent vs. frequent argument frames in McElree (1993); (b) the information is indeed available only at a delay, as it requires the completion of an extra process—e.g. accessing thematic vs. syntactic information in McElree and Griffith (1995) and Omaki et al. (2015); and (c) the information is itself available early, but an exhaustive selection process engenders a costly delay. I argue here that pragmatic enrichment is an example of case (c).

One complication to this picture comes from the observation that different kinds of pragmatic enrichment do not exhibit consistent processing effects. For instance, comparable response delays aren't observed in judgment tasks with free choice or conditional perfection inferences (Chemla & Bott, 2014; van Tiel & Schaeken, 2017). In a recent line of work, van Tiel and colleagues have demonstrated that evidence for processing costs varies even within the domain of scalar implicatures (van Tiel et al., 2019; van Tiel and Pankratz, 2021; Sun et al., 2023; see Khorsheed et al., 2022 for review). On the one hand, delays for responses consistent with upper-bound meanings are substantial for many other weak scalar items besides *some*, e.g. *or*, *might*, *most* (van Tiel et al., 2019), as well as weak adjectives like *content*, *passable*, *warm* (van Tiel & Pankratz, 2021). On the other hand, for many weak scalar items, costs are reliably absent, e.g. *scarce*, *low* (van Tiel et al., 2019), *cool*, and *mediocre* (van Tiel & Pankratz, 2021). In fact, exceptional items also show a lack of resource-dependence in a cognitive load task (van Tiel et al., 2019), and no lag in enrichment-based behavior in visual-world gaze patterns or incremental forced referential prediction (Sun et al., 2023), although in this latter case we might not have expected a lag to begin with (Sun & Breheny, 2020).

Noticing that the only cases where difficulties are observed are with weak positive scale items, van Tiel and colleagues have hypothesized that such difficulties are related to an interaction between scalar implicature and polarity. The proposal is as follows: enriched meanings of statements with weak positive scale items feature the negation of a stronger positive scale item. Administering a judgment to sentences with an explicit (*not above*) or "implicit" (*below*) negative meaning has long been observed to come at a delay (Clark & Chase, 1972), at least without supporting context (cf. Nieuwland and Kuperberg, 2008). Administering a judgment to sentences with enriched weak positive scale items, then, would also undergo such a cost. In contrast, enrichment of weak negative scale items requires the negation of a negative meaning, equivalent to a positive meaning and thus perhaps thus easier to process than an un-negated negative, if you believe that negative valence, and not negation as an operator, drives the classic negative delay effects (see Sherman, 1976). Nevertheless, this negativity-based proposal would incorrectly expect the same costs for verification of *only some*, given the negation introduced by *only*, whereas we see evidence that enriched-*some* verification remains more costly than *only some* verification (L. Bott et al., 2012; Marty & Chemla, 2013).

Accounting for the scale-specific variation in processing costs remains troubling under the ambiguity resolution account discussed above, but not impossible. If costs in verification paradigms arise as a symptom of the difficulty of inhibiting a favored un-enriched meaning, and if the un-enriched meaning is less favored in cases like *low* and *mediocre*, then we could derive the absence of costs. Offline judgment studies (e.g. van Tiel et al., 2016) do find wide variation in endorsement rates for implicatures, and *low* to *low but not depleted* and similar cases are indeed among the most endorsed, but so is *some* to *some but not all*: more work is needed to determine whether degrees of bias could indeed explain the spectrum of processing costs.

3.1.4 The outlook from reading time experiments

Against the background presented in the previous sections, reading time studies offer a unique opportunity to observe the timing of *some*-enrichment during a less directed task. They are also the one paradigm where researchers have examined the interpretation of *some* at a wider time window, looking for consequences of an enriched upper-bound meaning beyond the *some* assertion itself. On the whole, results from these studies fail to provide consistent evidence that readers require a costly generation process to access the upper-bounded meaning of *some*. Nevertheless, they add to the ERP and visual-world studies reviewed above in demonstrating that upper-bounded meanings are activated quickly enough to influence processing of later material.

As in the visual-world paradigm, early research on *some* in reading found evidence consistent with an immediate but costly enrichment process. Breheny et al. (2006) instantiated what has come to be the typical design for such a study, investigating reading times at *some* and immediately following positions while manipulating whether the context supports an upper-bounded reading. Researchers also typically examine the processing of context-sensitive *only some* as a control: a context-dependent slowdown at *some*—but not *only some*—in upper-bound-consistent contexts would be taken to index an immediate costly enrichment process. Breheny et al. (2006) manipulated support for the upper-bound reading by introducing explicit narrative-internal questions, as in (10). (Stimuli are presented here in translation, as the study itself was conducted in Greek.) For validation that scalar enrichment is sensitive to relevance manipulations via such questions, see e.g. Zondervan et al. (2008).

(10) **Stimuli from Breheny et al. (2006), Experiment 3 (critical segment in bold)**

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

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

Breheny and colleagues observed that, in a segmented self-paced reading task, the segment containing *some* was read 100ms slower in the upper-bound context, taking this as evidence of costly enrichment. Although a control condition with *only some* was shown to participants, the authors declined to analyze reading times at *some* for this condition, presumably because the extra particle was present within the critical segment.

Bergen and Grodner (2012) improved somewhat on the quality of the evidence for costly enrichment in a later word-by-word self-paced reading study in English. Rather than manipulating the relevance of the upper bound in context, they controlled the availability of the upper-bound reading by manipulating the characterized knowledgeability of the speaker (11). As scalar enrichment here depends on the assumption that the speaker has the competence to have asserted *all* (see e.g. Goodman and Stuhlmüller, 2013 and Tsvilodub et al., 2023 for offline validation), slowdowns particular to the knowledgeable-speaker condition may index enrichment costs. Indeed, at the quantifier itself, and the three words immediately following it, Bergen and Grodner observed that self-paced reading times are delayed by roughly 20ms in this condition. Controls featuring *only some* in the same window showed no comparable effect. They attributed this delay to context-specific generation of the upper-bound meaning, blocked when the prerequisites for scalar enrichment were not met.

(11) **Stimuli from Bergen and Grodner (2012) (critical segments in bold)**:

a. **Knowledgeable**: At my clients request, I meticulously compiled the investment report. **Some of the real estate** investments lost money. The rest were successful despite the recent economic downturn.

b. **Neutral**: At my clients request, I skimmed the investment report. **Some of the**

real estate investments lost money. The rest were successful despite the recent economic downturn.

While both these studies would suggest immediate and costly enrichment, a trio of studies aiming to replicate the original Breheny et al. (2006) finding failed to observe the same penalties at *some* (Politzer-Ahles & Fiorentino, 2013; S. Lewis, 2013; Hartshorne & Snedeker, 2014). Politzer-Ahles and Fiorentino (2013) ran a quite close self-paced reading replication in English, standardizing the contextual manipulation to a comparison between a contextual *all* question vs. a contextual *any* question, where the former should support an upper-bound reading. Nevertheless, they found no differences at *some* across contexts. In a quite similar eyetracking-while-reading experiment, S. Lewis (2013) introduced implicit *all* vs. *any* questions via the intentions of a protagonist, and again observed no differences at *some*, in any eye-movement measure. Finally, in another self-paced reading study Hartshorne and Snedeker (2014) attempted to control the availability of the upper-bound reading by manipulating the monotonicity of the immediate semantic environment, following evidence that scalar implicatures are largely not computed in downward entailing environments like the antecedent of a conditional (e.g. Chierchia et al., 2001 re: *or*). While this might be expected to be a particularly strong manipulation, as it relies on an interaction between enrichment and grammatical context, Hartshorne and Snedeker again observe no differences in the reading profile of *some* in either of the two experiments they report, or the many replications mentioned in their discussion.

Evidence is thus split as regards a contextually-gated, immediate, and costly enrichment process during reading. To date, there has not been an attempt in the literature to untangle the exact factors which control the difference in findings between Breheny et al. (2006) and Bergen and Grodner (2012) on one hand, and the null effects of Politzer-Ahles and Fiorentino (2013), S. Lewis (2013), and Hartshorne and Snedeker (2014) on the other. Nevertheless, given the preponderance of null observations, and the relatively more precise manipulations and measurements used in the studies which produced null observations, it seems most likely that the purported costly enrichment process is not characteristic of typical reading.

While the nature of the initial comprehension processes at *some* remains unclear, reading time studies have consistently found evidence that context modulates consideration of the upper-bound reading by the following clause. All five of the studies discussed above also probe reading times at a following elliptical referring expression like *the rest*, which would most naturally refer to a complementary subset of the restrictor quantified by the critical *some*—e.g. in (10), those relatives who John will not be hosting, or in (11), those investments which did not lose money. Breheny et al. (2006) and subsequent authors assume that the interpretation of *the rest* depends on a partition of the restrictor which will already have been made salient to the extent that participants have considered an upper-bound reading of *some*. Across the board, all observe that reading times at *the rest* are controlled by their contextual manipulations: faster following *some* given an upper-bound-consistent contextual question (Breheny et al., 2006; Politzer-Ahles & Fiorentino, 2013; S. Lewis, 2013), a knowledgeable speaker (Bergen & Grodner, 2012), or an upward-entailing environment (Hartshorne & Snedeker, 2014). In no case is the relevant effect observed following *only some*, after which *the rest* always evokes relatively fast reading times, consistent with the expectation that it always has an upper-bound reading which makes the complement set salient.

This finding offers two interesting conclusions. First, it validates that these contextual manipulations did modulate the degree to which an upper-bound reading was considered. Second, it offers some evidence that this consideration has accumulated with some substance by the reading of material which follows *some* a few positions later, consistent with what was observed in comprehenders' ERPs and anticipatory gaze behavior. Hartshorne and Snedeker (2014) suggest that this effect is somewhat late-emerging, finding that it does not emerge when probed directly after the quantified noun phrase (a temporal lag of about 900ms), although Politzer-Ahles and Fiorentino (2013) found no evidence that lag time affected the size of their effect in this way.[7]

Notice that it would be incorrect to take effects at *the rest* to diagnose costly reanalysis, and therefore the timing of selection and commitment. The presence of a complement set is just as possible regardless of whether an enriched meaning of *some* is interpreted, as the lower-bounded meaning for *some* does not entail the absence of a complement. As a result, we still have only evidence for the timecourse of consideration of the enriched meaning, not its selection.

The critical evidence to support an incremental commitment would be the costly

[7] Interestingly, Hartshorne and Snedeker (2014) investigated cases where *the rest* was within the same sentence as *some*, in a coordinate or the consequent of a conditional, while Politzer-Ahles and Fiorentino (2013) uniformly investigated cases where a sentence boundary intervened. This may be a case where optionally-delayed enrichment processing is flexible within a sentence, but likely to be taken care of at sentence boundaries, cf. the case of sense selection for polysemes (Frisson, 2009; Foraker & Murphy, 2012).

processing of some material which is incompatible with the enriched meaning, and thereby requires cancellation. A brief note in Bergen and Grodner (2012) offers the only known evidence as to the presence of such an effect. In addition to passages with codas containing *the rest*, Bergen and Grodner also had participants read passages with codas which affirm the absence of a complement, i.e. which affirm that all members of the restrictor satisfy the nuclear scope (12).

(12) **Cancellation stimulus from Bergen and Grodner (2012):**

a. At my clients request, I meticulously compiled the investment report. Some of the real estate investments lost money. In fact, they all did because of the recent economic downturn.

Much in contrast with their effect at *the rest*, the authors observed no effect of speaker knowledgeability at any point in the reading of these cancellation-prompting codas, declining to analyze the result. If this lack of an effect is systematic, it would suggest that selection between lower-bound and upper-bound readings of *some* is indeed absent during typical reading, despite evidence that upper-bound readings have been activated.

I summarize the relevant generalizations and associated conclusions in Table 3.3. Failures to consistently observe reading costs in implicature-supporting environments, together with evidence for downstream effects within a few words, add to the studies reviewed in Table 3.2, suggesting that upper-bound meanings can enter quickly into consideration without particular cost. The tentative cancellation finding, in turn, is consonant with the interpretation of the response time literature presented in Table 3.1, suggesting that costly specification of enriched meanings is not volunteered during online reading, but instead adopted only when necessary for further reasoning.

3.1.5 Decoupling consideration from selection

Going forward, as a shorthand, I will refer to the view developed here as the hypothesis of Rapid Consideration Without Selection, as summarized in (13).

(13) Rapid Consideration Without Selection

Optional enrichments derived from pragmatic principles are considered within the process of incremental comprehension, and drive expectations for future input gradiently, modulated by the amount of evidence which supports the enrichment. Selection and commitment to an enriched meaning does not happen in the course of

Table 3.3: Summarizing generalizations and conclusions from the reading literature on *some*-enrichment.

Generation of UB meaning		**Selection of UB meaning**	
Timing	*Difficulty*	*Timing*	*Difficulty*
Reading times don't reflect difficulty at *some* in upper-bound contexts. (e.g. Politzer-Ahles and Fiorentino, 2013; S. Lewis, 2013)			
-	Not costly (or else difficult to observe)	-	-
Reading times refect facilitation at *the rest* in upper-bound contexts (e.g. Breheny et al., 2006; Bergen and Grodner, 2012)			
Influences reference processing before 900ms	-	-	-
Reading times don't reflect difficulty upon cancellation in upper-bound contexts. (Bergen & Grodner, 2012)			
-	-	Not until much later	-

normal comprehension, unless prompted, at which point it may be difficult for a given construction when a context-independent bias favors the un-enriched meaning of that construction.

We can contrast this with two simpler alternatives which don't decouple consideration from selection, Strictly Offline Enrichment (14) and Rapid Enrichment (15). These are admittedly overly naive, but they will help us see why this decoupling is necessary.

(14) Strictly Offline Enrichment

Comprehenders do not consider pragmatic enrichment during online comprehension. Enriched meanings are considered and selected only once the basic logical meaning of the input has been fully established, in an offline process, when necessary.

(15) Rapid Enrichment

As comprehenders encounter material which might be pragmatically enriched, they select between an un-enriched and an enriched meaning based on contextual support, much like the process of homonymy resolution.

Strictly Offline Enrichment is *a priori* quite implausible, and indeed does not hold up well when we consider the robust evidence from ERP and visual world studies for expectations based on upper-bound meaning for *some* before clause offset, reviewed in §3.1.2.

These cases of rapid expectations based on upper-bound meanings receive a more natural account if selection of an upper-bound meaning has already occurred, per Rapid Enrichment, as does the consistent evidence for complement reference facilitation in reading studies, reviewed in §3.1.4.

Indeed, on the whole, Rapid Enrichment faces just a few challenges given the current empirical landscape. First, there is inconsistent evidence for any homonym-like process of costly selection in incremental processing, as also reviewed in §3.1.4. Breheny (2019) suggests that this is not so damning; if we follow the corpus evidence presented in Degen (2015), lower-bound and upper-bound interpretations of *some* are about equally as likely in real contexts, and per e.g. Duffy et al. (1988), selecting an interpretation for an equibiased ambiguity is not inherently costly when context supports one interpretation.

A more serious obstacle is how to account for the delays and resource-contingencies observed for upper-bound-based responses in judgment tasks, as reviewed in §3.1.1. On the one hand, if an upper-bound interpretation has already been selected at stimulus offset, what is responsible for the delay? On the other hand, how is selection of an upper-bound meaning especially contingent on executive resources if the ambiguity in these cases is equibiased? One recourse here is to explain these factors as consequences of properties of the generated meaning which are unrelated to the presence of the implicature itself, like the account based on negativity costs advanced by van Tiel et al. (2019), but as noted, these face difficulty in accounting for the contrasts between *some* and *only some* observed by L. Bott et al. (2012) and Marty and Chemla (2013).

The most critical place where Rapid Enrichment and Rapid Consideration Without Selection diverge, however, is their predictions vis-à-vis reanalysis effects. Because Rapid Enrichment explains within-clause facilitation effects as the consequence of early selected upper-bound meaning, it expects that any downstream content incompatible with upper-bound meaning should be associated with costly reanalysis as comprehenders revise their decision, much as observed for homonymy. Rapid Consideration Without Selection, however, holds that early facilitation effects are strictly the result of pre-selection consideration, that *some* does not receive a selected interpretation online, and thus that comprehenders will be under no obligation of reanalysis when faced with later disambiguation. We have some data from Bergen and Grodner (2012) that suggests an absence of costly reanalysis, validating Rapid Consideration Without Selection. My own data provides a further test case which I will take as stronger evidence for Rapid Consideration

Without Selection.[8] Starting in §3.3, I will lay out how the experiments in this chapter were designed to build on that result, so as to evaluate these critical predictions and arbitrate between these two hypotheses.

First, in the next section, I will pause for a moment to consider how exactly we might model gradient facilitation of expected content without comparable difficulty with unexpected content, as Rapid Consideration Without Selection seems to require.

3.2 Interlude: No Worries If Not

The above discussion suggests a remarkable generalization about comprehenders' behavior in the period after the introduction of *some*, which I'll call affectionately No Worries If Not (16).

(16) No Worries If Not

In contexts favoring interpretation X of an element which may mean X or Y, comprehenders exhibit:

a. facilitation with X-compatible continuations

b. but no difficulty with Y-compatible continuations

In particular, this is the state of affairs suggested by the null effect of context on reanalysis costs in Bergen and Grodner (2012): the same change in context causes facilitation of continuations consistent with an upper-bound meaning, but no change in the reading of continuations which necessitate a lower-bound meaning. More robust evidence for the slightly more general phenomenon that expectations conditioned on a likely parse do not preclude expectations conditioned on an unlikely parse comes from the partial facilitation of lower-bound consistent continuations in the Hunt et al. (2013) ERP study.

[8] I note at this point that all critical argumentation will be made about the timeline of selection of the upper-bound meaning in contexts which support it. In these contexts, we have good evidence that upper-bound meaning is considered rapidly due to evidence from ERP, visual world, and reading time facilitations on complement reference, but we also have good evidence that selection is delayed, from late-occuring selection costs in judgment latencies and the absence of garden-path effects in reading. We lack this quality of evidence for selection of lower-bound meaning in contexts which support that meaning: it is unclear the extent to which upper-bound meaning is considered here during online comprehension, only that is considered less than in upper-bound-biased contexts. Likewise, there is no evidence for garden-path effects in such contexts, but a shift from lower-bound to upper-bound meaning may simply proceed additively without requiring reanalysis, so it is unclear how to interpret this fact. It thus remains possible that lower-bound readings could be selected early in contexts which support them, although we will remain more focused on the timecourse of interpretation in the other conditions. I thank Jess Law for the encouragement to clarify this point.

I have so far simply said that this is incompatible with the idea of early selection, as selection would entail costly reanalysis in the latter case. The facilitation therefore must come from some kind of pre-selection consideration of the potential meanings, sensitive to context. But more must be said here. A natural formalization of this pre-selection consideration would be a model based on distributional probability. In such a model, facilitation effects would be achieved as comprehenders reason over multiple interpretations of their current input and pre-allocate resources based on the conditional probability of various continuations. As interpretation X becomes more likely, more resources are pre-allocated towards expectation of X-consistent content, which is in turn integrated more quickly than it would have been in contexts where X was less likely. Such a model closely resembles the landmark proposals of e.g. Hale (2001) and Levy (2008), where consideration of a possible syntactic parse X drives expectations for possible continuations proportional to their conditional probability given X. However, note that these models in their basic form predict reanalysis-like costs for Y-consistent continuations as X becomes more likely: this is indeed exactly how Levy (2008) models classic garden-path effects without assuming commitment to a single parse.

But the facts, at least those known to us now, suggest that facilitation for X-consistent input need not result in difficulty for Y-consistent input. So, in order to capture No Worries If Not, a classic probabilistic model won't do.

The spitefulness of these models is essentially a result of a resource limitation: because there is only a single unit of probability to distribute, if the probability of analysis X and its likely continuations rises, the probability of analysis Y and its likely continuations must fall. Any system with substantial resource limitations would recapitulate the same dependency, even if it were unmoored from probability, so long as increasing expectation of X-based continuations required decreasing the expectation of Y-based continuations.

We need not assume such stringent resource limitations in our modeling of context-based expectations. Even if there is a finite pool of "activation" which the comprehension mechanism draws upon to anticipate certain likely continuations, if the mechanism does not exhaust that pool at all times, then it would be able to increase resources committed to a certain likely continuation by drawing on the stockpiled pool, without necessarily decreasing the resources committed to an unlikely continuation. Could such a system, which doesn't use all resources at its disposal, be rational? Consider the behavior of an investor with a million dollars of capital and an imaginary set of 100 possible investments. They will not, generally, have the full million invested at any given moment, or seek to invest some in all 100 possible ventures. There are two good reasons for this: first, many investments are unlikely to yield profits, and second, they may have uses for their money beyond investment. If the resources used to anticipate certain continuations in incremental comprehension likewise come from some pool of general cognitive resources (say, attention) which can also be consumed by processes outside the comprehension mechanism, then a rational comprehender would use those resources for expectation when they will be useful, and otherwise leave them in reserve.

To demonstrate that resource-rich models are able to capture No Worries If Not patterns better than resource-limited models, I will walk through a toy example based on the Bergen and Grodner (2012) design. Comprehenders will encounter the word *some*, which has a possible lower-bound or a possible upper-bound meaning, in either a neutral context or an upper-bound biased context, and then they will encounter the word *rest* or the word *all*. I will assume that comprehenders are already aware of (simplified) distributional differences in the targets based on various meanings of *some*: an upper-bound meaning entails that *rest* will appear next, while a lower-bound meaning might be followed by either *rest* or *all* with equal likelihood. The two toy models will both attempt to distribute points of activation based on likely interpretations of *some* and likely upcoming tokens, and when a token receives more pre-activation, it will be read faster. The major difference between the resource-limited model and the resource-rich model is the total number of points of activation they may distribute: the resource limited model has 200 and the resource-rich model has at least 400.

I begin by working through the predictions of the resource-limited model. After encountering *some* in the neutral context, the model will commit 200 points of activation to predicting continuations, the total amount of resources available. Those 200 points will be allocated based on reasoning about the meaning of *some*: since there's no bias here, 100 will be allocated to anticipate continuations based on the upper-bound meaning (all thus expecting *rest*), and 100 will be allocated to anticipate continuations based on the lower-bound meaning (half expecting *rest*, half *all*). *Rest* is ultimately facilitated by 150 points of pre-activation, while *all* has 50 points.

In the biased context, the resource-constrained model will allocate its 200 points differently: given stronger anticipation for the upper-bound meaning, 75% will be allocated based on the upper-bound, all expecting *rest*. The remaining 25% will be allocated based on the lower-bound, split between *rest* and *all*. In this case, *rest* is more pre-activated than it was in the baseline, at 175 points, but *all* is less so, at 25 points. This incorrectly predicts greater facilitation with *rest* and greater difficulty on *all* in this context.

The resource-rich model here differs across conditions not only in its proportional distribution of resources, but also in its absolute amount of resources devoted to pre-activation. Imagine that the model invests 200 points in expectations for every useful cue it receives. In the neutral context, it receives only one useful cue, that *some* is present in the input. Allocating 200 points, it will distribute these to *rest* and *all* just like the resource-constrained model, yielding 150 points for *rest* and 50 points for *all*. The biased context for *some* is another piece of useful information that changes expectations, meriting another 200 points of activation. Now a total of 400 points will be allocated in this case, 75% based on the upper-bound, expecting *rest*, and the remaining 25% based on the lower-bound, split between *rest* and *all*. As a result, *rest* concludes with much more pre-activation than in the baseline, 350 points, while *all* has not changed, ending again with 50 points. These predictions match the No Worries If Not effect: greater facilitation with *rest* without any added difficulty for *all*.[9]

To the extent that we observe credible No Worries If Not patterns, then, we have evidence for a stage of anticipatory processing which is best modeled by a resource-rich system of interactive expectations. In the remainder of this chapter, I will present two experiments which aim to determine whether these patterns are indeed present with regard to the interpretation of *some*.

3.3 The present study: Extending Bergen & Grodner (2012)

The present study aims to bring more evidence to bear on the timecourse and costs associated with the generation and selection of the upper-bound meaning of *some*,

[9] Note that this outcome depends on the total allocation of resources. If resources were not at least doubled, *all* would have been subject to some loss in activation, and if resources were more than doubled, *all* would in fact gain in activation compared to the neutral context, expecting even more facilitation. To make this model generate testable predictions, we would have to specify exactly how resources are allocated. While the vague "cue"-based approach adopted in the text above depends on the implausible assumption of discrete units of information, one more rigorous approach would be to use an information-theoretic operationalization, such that the amount of resources invested is inversely proportional to the entropy of the probability distribution over potential meanings. This is intuitively reasonable: the more you think you know the future, the more resources you should commit to exploiting that expectation.

in particular by expanding on the design set out by Bergen and Grodner (2012). This particular design is chosen for a few reasons. First, while there have been two failed attempts to replicate the context-based relevance manipulations of Breheny et al. (2006) (Politzer-Ahles & Fiorentino, 2013; S. Lewis, 2013), I am not aware of any attempts to replicate Bergen and Grodner's speaker knowledge manipulations. Perhaps that effect is more robust. Second, their study is the only case in the literature where cancellation was probed. Testing for reanalysis effects offers a critical test for the presence of selection in online reading, and it seems wise to begin where an attempt has already been made. However, there is at least one feature of their design which makes interpreting their cancellation findings challenging. In this section, I will go into detail about the parameters of their design, this potential confound, and how the materials for the present study were created to improve on the design.

3.3.1 The original design

Bergen and Grodner constructed twenty-four first-person narratives featuring an initial context-setting sentence (S1), a target sentence featuring *some* (S2), and a coda consistent with one of the potential meanings of *some* (S3). Critical data came from a subset of four conditions derived by crossing CONTEXT, the knowledge of the speaker as communicated in S1 (Knowledgeable vs. Neutral), and INFORMATION STATUS, the nature of the potential upper-bound meaning as communicated in S2 (Implicature vs. Entailment, i.e. *some* vs. *only some*).[10] All four of these critical conditions were displayed before an S3 which remained consistent with an upper-bounded meaning by reference to a complement set via *the rest* or *the others*. In two supplementary conditions, using only the Implicature (*some*) targets, they also investigated the processing of an S3 which mandated a lower-bound interpretation of the target quantifier by asserting that indeed the nuclear scope of S2 was true for all members of the quantified domain, again across the Knowledgeable and Neutral contexts. All the building blocks which make up one sample item are laid out in (17).

(17) **The full design of Bergen and Grodner (2012)**

[10] By their terminology, these were "full-knowledge" vs. "partial-knowledge" contexts and "scalar" vs. "focused" triggers.

Context	
Knowledgeable	I carefully inspected the new shipment of jewelry.
Neutral	I helped unload the new shipment of jewelry.
Information Status	
Implicature	Some of the gold watches were fakes.
Entailment	Only some of the gold watches were fakes.
Coda	
Affirmation	The rest were real, but the company is still planning to sue.
Cancellation	In fact, they all were, so the company is planning to sue.

The authors expected implicature calculation to occur most regularly in the Knowledgeable contexts, and to be largely blocked in the Neutral contexts, because derivation of the upper-bound meaning requires an assumption that the speaker would know if it would be appropriate to use "all". Contexts were normed in a likelihood rating study with twenty participants to ensure they reliably modulated this particular assumption.

The resulting twenty-four item sets across six conditions were presented to forty-two participants in a Latin square, using word-by-word self-paced reading task followed by true/false comprehension questions. This makes for a total of 168 observations per condition, or four observations per condition per participant and seven observations per condition per item set, before exclusions. One goal of the present study is to improve on this power, as it is not up to current standards for convincingly demonstrating a reading time effect in the ballpark of 10ms. E.g. Vasishth et al. (2022) present an instructive power simulation suggesting that somewhere around 1600 observations per condition (in their case, 200 participants and 16 items over 2 conditions) would be ideal to observe strong evidence for a reading time effect of 16–81ms in the Bayesian mixed-effects regression analytical pipeline I will use here. While the experiments reported here will not quite meet that threshold, they will approach it far more closely than the original study.

Bergen and Grodner focused their analysis on two critical regions: the critical target *some* and the positions which followed, and the complement set expression (e.g. *the rest*) and the positions which followed. In the first case, they observed a predicted effect similar to Breheny et al. (2006), a small penalty in response latencies at *some* after the Knowledgable context vs. the Neutral context, not observed in the somewhat-faster Entailment controls, attributed to the cost of generating (and/or selecting) an enriched upper-bound meaning of *some*. In the second case, they observed a second predicted effect similar to Breheny et al. (2006), a small benefit in response latencies at the predicate in S3 following *the rest*, in the Knowledgeable context vs. the Neutral context, not observed in the somewhat-faster Entailment controls, attributed to implicature-driven facilitation of the complement reference.

The connection the authors hypothesized between the effects of generation and facilitation is supported by a correlational analysis performed on the response latencies in the critical regions by trial. To prepare for this correlation analysis, the authors first constructed a model of typical within-trial latency patterns by predicting latencies at the predicate following *the rest* from latencies at *some* in the control Entailment sentences. (In the general case, response latencies in an early region are positively correlated with later response latencies, as latency is often quite consistent within a given self-paced reading trial.) They then applied this model to the critical Implicature sentences, and extracted the residuals, to retain only the variance which could not be explained by the typical effect of within-trial consistency. They observe a predicted negative correlation with the residuals, such that longer latencies in the *some* region predicted faster-than-usual latencies at the predicate following *the rest.* This correlation is convincing support for a connection between costly processing at *some* to the ease of processing later complement reference, consistent with the proposal that the penalty at *some* involves generation of an upper-bound meaning.

Analysis of the cancellation codas, by comparison, was largely parenthetical in the paper as published. Very little effect of the Context manipulation was observed. Again, if we believe that the costs at *some* indexed a process that included selection of the upper-bound meaning, we would expect a mirror-image effect for upper-bound-consistent vs. upper-bound-inconsistent continuations. As the context better supports an implicature, latencies at the implicature trigger should increase, latencies at a consistent continuation should decrease due to facilitation, and latencies at an inconsistent continuation should increase due to costly reanalysis. The presence of facilitation in lieu of reanalysis is more compatible with a model where an upper-bound meaning may be differentially activated in the absence of firm commitment. However, the small size of the expected reading time effect, the inferential fragility of a null result in a null-hypothesis-significance-testing framework, and the solitude of this one attempt to observe the critical effect leave conclusions uncertain. The present study aims to offer a better opportunity to quantify the evidence for or against the predicted cancellation effect. In addition to increasing power and applying modern Bayesian methods, we will also have to address a potential confound obscuring detection of a cancellation effect, to which I will now turn.

3.3.2 A potential confound

In a modern theory of scalar implicature derivation in the tradition of Grice, in cases where speaker competence is in doubt, the implicature featuring an upper-bounded meaning will be replaced by an "ignorance inference," essentially that the speaker does not know whether it would be appropriate to use a stronger alternative (e.g. see the derivation in Chemla and Singh, 2014a). That is, speakers that utter *some* are essentially either presumed competent, and thus taken to implicate *some but not all*, or else taken to implicate their own ignorance.

This state of affairs is a problem for the design of the cancellation condition in Bergen and Grodner (2012). The logic of the comparison they examine relies on the high likelihood of an implicated meaning being present in the Knowledgeable context, and thus having to be cancelled in S3, while the same S3 in the Neutral context is associated with no comparable cancellation process. However, the theory of implicature described above predicts that whenever participants do not compute the upper-bound implicature, they will calculate speaker ignorance as to *all*. As a result, in both the Knowledgeable and Neutral contexts, the Cancellation coda requires participants to cancel an inference that the context makes particularly likely. Either they infer *some but not all* and must retract it in the face of the *all* assertion, or they infer ignorance and must retract it in the face of the *all* assertion. There is no opportunity to selectively estimate the size of any reading time effect specific to cancellation by comparing these conditions. It indeed is quite possible that the absence of any difference between the conditions is the result of slowdowns in both cases rather than an absence of a slowdown in either.

In order to appropriately measure a potential cancellation effect related to an upper-bound meaning for *some*, a design must feature, minimally, a condition where the critical region will be expected to trigger cancellation of that meaning, and a condition where it will be compatible with all the commitments expected to have been generated earlier in the sentence. The solution I will adopt here is to set up a more complex meaning, where readers can use contextual information to avoid constructing an upper-bound meaning for *some* without necessarily concluding speaker ignorance. I describe how this is achieved below.

3.3.3 Materials

I created 40 three-sentence narratives building from the template used in Bergen and Grodner (2012) with three major differences of design. Like Bergen and Grodner's item sets, the narratives always began with an sentence which provided information about someone's knowledgeability (S1), manipulating whether they were presented as Knowledgeable or Neutral. However, in this case, the individual introduced is a named third-party protagonist, distinct from the speaker. In the second sentence (S2), a proposition containing *some* (Implicature) or *only some* (Entailment) is then introduced as a shared belief of the protagonist and the narrator, using a factive embedding verb like *notice* or *realize*. Finally, in the third sentence (S3), the narrator makes an unembedded assertion which clarifies whether an upper-bound meaning would be an appropriate inference at the matrix level of the narrative, either making upper-bound-consistent reference to a complement set, or directly contradicting an upper-bound meaning by making an *all* assertion. I'll refer to these two latter conditions as the Affirmation or the Cancellation of an upper-bounded meaning. A complete item set is presented in (18).

(18) **A sample item set from Experiments 5 and 6**

Context	
Knowledgeable	Petra wrote an article about the company's response to the scandal.
Neutral	Petra heard a bit about the company's response to the scandal.
Information Status	
Implicature	She realized that some of the marketing executives were fired.
Entailment	She realized that only some of the marketing executives were fired.
Coda	
Affirmation	The rest suffered a huge pay cut, which seemed fair.
Cancellation	In fact, they all were, which seemed fair.

Note that unlike the partially-crossed design of Bergen and Grodner, this study fully crosses all three factors of interest, resulting in eight critical conditions, including two where the upper-bound meaning is entailed via *only some*, but followed by an incompatible Cancellation coda. In such cases, the coda is in fact a global contradiction, and there should be no way to integrate it with an accurate representation of the narrative to that point. Entailment plus Cancellation conditions, then, offer the opportunity to observe how the comprehender handles irreconcilable contradiction, and how this may or may not be different from difficult but grammatically-permissible reanalysis.

To demonstrate how manipulating the likelihood of an upper-bound meaning through the knowledgeability of a third-person attitude holder makes it possible to avoid the confound described above, I will step through the assumed reasoning an eager comprehender might exploit in the conditions of interest to select a meaning for S2.

The first step is to understand the possible pragmatic interpretations of S2, independently of a particular context. When the narrator attributes to Petra the proposition that *some of the marketing executives were fired*, they entail that she holds a belief consistent with at least the lower-bounded meaning, and they also presuppose the truth of at least the lower-bounded meaning (indicating that they too believe at least that *some* were fired). I will assume there are two central dimensions along which relevant alternative utterances could have varied. First, of course, the narrator has used *some*, rather than the stronger *all*, which would attribute a stronger belief to Petra, and which would thus be a stronger assertion for the narrator. Second, the narrator has used a factive predicate, indicating their own belief in the truth of the complement, rather than the weaker option of merely attributing a belief to Petra using a non-factive epistemic predicate like *believe*. The other possible combinations of these elements vary in their strength. Two notable combinations are stronger assertions than the one in S2. If the speaker wanted to communicate that Petra believed *all* executives were fired, and that they shared this belief, they could have unambiguously communicated this using a factive with *all*, e.g. *Petra realized that all of the executives were fired.* Given they did not do this, the Gricean reasoner can imagine we are not in that scenario. Alternatively, if the speaker wanted to communicate any world where Petra believed *all* executives were fired, without committing to the same proposition for themselves, they could have used a slightly weaker assertion featuring a non-factive with *all*, e.g. *Petra believed that all of the executives were fired.* (They would then implicate either their own ignorance or belief that *not all* were fired, given their avoidance of the stronger factive alternative.) That they did not use this option, then, implicates that Petra does not believe *all* executives were fired. Taking these implicatures into account, *realize that some* is consistent with six possible scenarios, which involve Petra either being ignorant or believing *some but not all*, and which could involve any of the narrator's possible positions (*all*, *some but not all*, or ignorance).

Now, consider when Petra is portrayed with implied exhaustive knowledge of the situation. The comprehender should trust that she is not ignorant, and if so could conclude that she believes *some but not all* were fired. The narrator's epistemic status remains unclear. I will add another assumption, that when the protagonist is painted as credible, and protagonist beliefs are presented with a factive predicate, comprehenders are most likely to imagine that the narrator shares any implicit enrichments of belief licensed for the protagonist. As a result, an eager comprehender has sufficient evidence to select an upper-bound interpretation at the matrix level of the narrative, consistent with the presence of a complement set. If the upper-bound meaning is negated in S3, the eager comprehender must at this point revise in order to maintain a coherent understanding of the narrative. They might, for instance, adopt an alternative reading where the narrator maintains a distinct epistemic stance from the speaker.

Alternatively, consider when Petra is portrayed with implied partial knowledge of the situation. It is quite possible now that she would be ignorant as to whether *all* executives were fired, presumably more likely than her having the exhaustive knowledge that would license the implicature that she believed *some but not all* were fired. And in the absence of a particularly credible protagonist, a comprehender has no particular information about the narrator's beliefs: they might share the protagonist's ignorance, or they might have knowledge consistent with *some but not all* or even *all*. It would be appropriate not to draw any implicature under this uncertainty, including the weak implicature of speaker uncertainty. That is, I anticipate that eager comprehenders will be more likely to select the weaker, lower-bound meaning for S2, while nevertheless allowing for the possibility that the narrator may have further information.[11] In this case, for that eager comprehender, reference to a complement set will not be particularly facilitated, but no revision would be necessary if the upper-bound meaning was negated.

We can compare both of these cases, which engender various possibilities for enrichment, to the expected profile for *only some*. Because *only some* entails the upper-bound meaning of *some*, that upper-bound meaning becomes a necessary part of the content of both the protagonist's beliefs and, due to the factive, the narrator's beliefs. This is unambiguous and context-independent. Regardless of context, an Affirmative coda featuring complement set reference should be facilitated, and a Cancellation coda should prompt severe comprehender difficulty, given that input does not permit any coherent interpretation.

The above line of reasoning derives the desired scenario, dissociating selection of

[11] When stated in this way, this almost sounds as if I am suggesting that comprehenders underspecify a comprehension decision. This isn't the case: I am merely suggesting that they might adopt a weak meaning. Its compatibility with more possible continuations is a function of its weakness rather than underspecification, and in this particular case, it is the complexity of the relevant proposition which disassociates the choice of that weak meaning from the supplementary implicature of ignorance.

a lower-bound *some* meaning from an ignorance implicature. However, to do so it relies on two assumptions that could be called into question: (i) that comprehenders consider scales of informativity for embedding verbs and embedded scalar items together to derive potential implicatures in the way sketched, and (ii) that comprehenders reliably attribute an implicated upper-bound meaning under factive embedding to the level of narrative truth in particular when an attitude-holder is portrayed as knowledgeable. Given the burden of these assumptions, before proceeding, I ran a small offline judgment study to verify that the knowledgeability of the protagonist affected rates of endorsement for the matrix-level implicature in the way predicted.

3.3.4 Norming

Sixty participants, all native English speakers between the ages of 18 and 40 raised in the US and participating on Prolific, read and administered judgments for the forty critical items in an questionnaire administered on PCIbex (Zehr & Schwarz, 2018). In particular, they were asked to judge the likelihood that there was at least one member of the quantified domain in S2 who did not satisfy the nuclear scope, taken to measure the likelihood of an upper-bounded meaning. E.g. for the item set in (18), comprehenders were asked "How likely is it that at least one marketing executive kept their job?" Efforts were made to use lexical material to pick out the opposite of the nuclear scope predicate rather than employ negation, to keep the likelihood prompts as simple as possible. The likelihood judgments were solicited on a four-point scale, where 1 indicated "unlikely" and 4 "likely".

The narratives were displayed to participants for judgment in one of eight conditions, derived from a 2×2×2 design. In all cases participants saw only the first two sentences of an Implicature version of the narrative, i.e. all items featured *some* and not *only some*, and S3 was never included. As in the intended reading study, participants saw either Knowledgeable or Neutral versions of S1, allowing us to observe the expected effect of context on offline implicature endorsement. But in addition to the critical manipulation, I varied two aspects of the design, to help probe whether the intended use of factive predicates with third-person protagonists would affect the potency of the knowledgeability manipulation. Items were varied in the Individual whose knowledge was manipulated in S1 (1st-person vs. 3rd-person), and the Embedding of the critical implicature trigger in S2 (Matrix vs. Embedded). This allowed comparison of the knowledgeability effect across four potential designs: in addition to the original first-person matrix-level implicature design of Bergen and Grodner (2012) as a control, we can examine the effect in the third-person factive-embedding design proposed above, as well as first-person embedding and third-person matrix designs which adopted partial qualities of the proposed design. For demonstration, the four versions of the Knowledgeable condition for the item in (18) as used in the norming experiment are presented in (19). Sentences were revealed one at a time in a cumulative self-paced reading format, intended to encourage careful reading of both sentences, with the final button press revealing the likelihood prompt.

(19) **The conditions used in the norming task for E5–6**
1st, Matrix I wrote an article about the company's response to the scandal.
Some of the marketing executives were fired.
1st, Embed I wrote an article about the company's response to the scandal.
I realized that some of the marketing executives were fired.
3rd, Matrix Petra wrote an article about the company's response to the scandal.
Some of the marketing executives were fired.
3rd, Embed Petra wrote an article about the company's response to the scandal.
She realized that some of the marketing executives were fired.

Prompt How likely is it that at least one marketing executive kept their job?

Responses are summarized in Figure 3.1. The data was fit to a cumulative-link ordinal mixed-effects model using `brms` (Bürkner, 2017, 2018), with sum-coded predictors, non-default normalizing priors of $\mathcal{N}(0, 1)$ for slopes and $\mathcal{N}(0, 5)$ for the three thresholds, and with parameters initialized at 0, estimated across 6 chains of 10,000 iterations (including 2,000 iterations of warmup). The fixed effects of the resulting posterior model are presented in Table 3.4. We observe that narratives with Neutral characters were on the whole associated with reduced endorsement of the upper-bound meaning, $\hat{\beta}$ = -0.08, 95% CRI = (-0.14, -0.03), as expected, a replication of the competence effect on implicature endorsements previously observed in e.g. Goodman and Stuhlmüller (2013) and Tsvilodub et al. (2023). 95% CrIs for the value of all other main effects and interactions did not exclude 0. Planned marginal comparisons reveal that the Knowledgeability manipulation yielded a credible difference in the expected direction for, in fact, all hypothetical designs besides the original template of Bergen and Grodner. That is, we observe robust context effects for narratives featuring 1st-person subjects with factive embedding, $\hat{\hat{\delta}}$ = -0.19, $P(\delta < 0)$ = 0.96, 3rd-person subjects without embedding, $\hat{\hat{\delta}}$ = -0.20, $P(\delta < 0)$ = 0.97,[12] and crucially 3rd-person

[12] This would be surprising if you believed participants were indeed tracking a 3rd-person character's knowledge and a speaker's matrix-level assertions separately. However, we could understand this effect if comprehenders are assuming the speaker has at least the same amount of knowledge as the relevant charac-

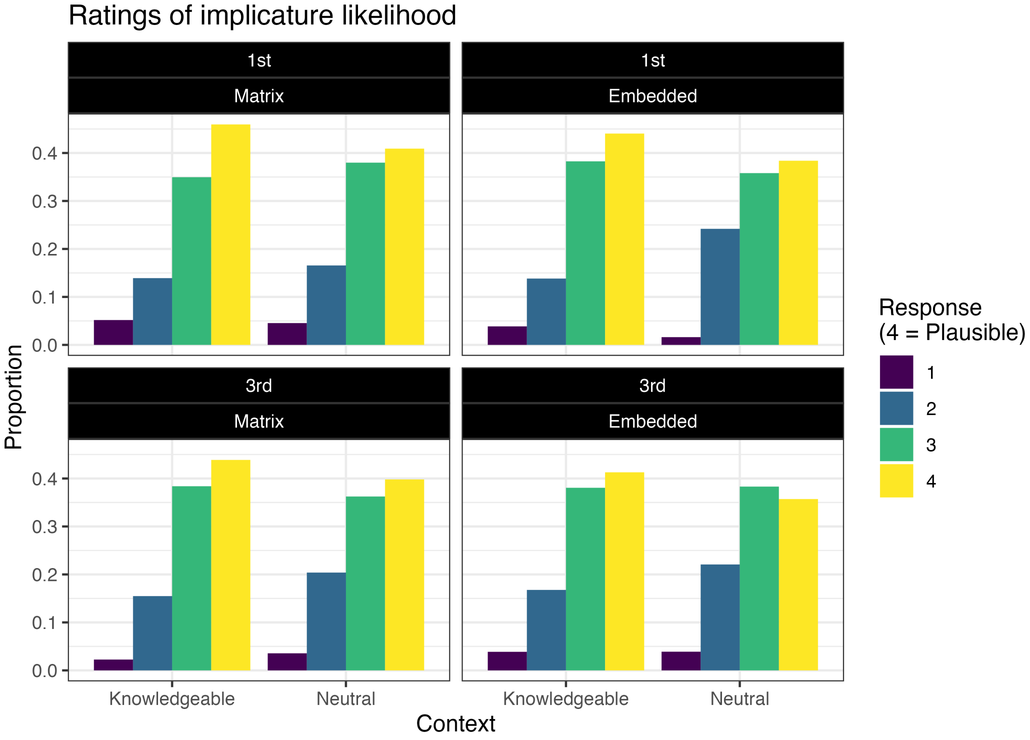

Figure 3.1: Response distributions from the norming task for Experiments 5 and 6

subjects with factive embedding, $\hat{\delta}$ = -0.18, $P(\delta < 0)$ = 0.96. For 1st-person characters without embedding, the effect is in the same direction, but weaker, to the extent that it is not credibly non-zero, $\hat{\delta}$ = -0.11, $P(\delta < 0)$ = 0.85. As for the overall likelihood of an implicature in each design, inspection of the plots reveals that endorsements were somewhat weaker for the intended narratives featuring 3rd-person subjects with factive embeddings compared to 1st-person subjects without embedding, but post-hoc marginal comparisons in the Knowledgeable condition reveal no credible difference, $\hat{\delta}$ = -0.11, $P(\delta < 0)$ = 0.82.

I conclude that the intended design behaves as expected, with a robust sensitivity to protagonist knowledge, and without any substantial reduction in overall derivation of an upper-bound meaning. Although the design introduces some undesirable complexity, it offers a way of handling the confound introduced above, and so it is adopted for the reading studies that follow.

ter, as would be natural.

Table 3.4: Bayesian ordinal mixed-effects model fit to 4-point likelihood responses in the norming task for Experiments 5 and 6. Factor levels in parentheses were coded as positive.

Effect	Posterior $\hat{\beta}$	Posterior σ_β	95% CRI Lower	95% CRI Upper
Threshold 1|2	-2.27	0.13	-2.52	-2.03
Threshold 2|3	-1.02	0.12	-1.25	-0.79
Threshold 3|4	0.27	0.11	0.05	0.50
Context (Neut)	-0.08	0.03	-0.14	-0.03
Embedding (Emb)	-0.04	0.03	-0.09	0.02
Individual (3rd)	-0.03	0.03	-0.10	0.03
Context × Embed	-0.01	0.03	-0.06	0.04
Context × Indiv	-0.01	0.03	-0.06	0.04
Embed × Indiv	-0.02	0.03	-0.07	0.03
C × Emb × Ind	0.01	0.03	-0.04	0.06

3.4 Experiment 5: Scalar implicatures in self-paced reading

For maximal comparability with previous reading time studies (not only Bergen and Grodner, 2012 but also Breheny et al., 2006; Politzer-Ahles and Fiorentino, 2013; Hartshorne and Snedeker, 2014), these narratives were shown to participants in a word-by-word self-paced reading task. Before detailing the method, I will briefly review how reading times on these narratives will speak to the open questions which have concerned this chapter so far, and the particular predictions of Rapid Consideration Without Selection.

One critical question is whether upper-bound meanings of sentences with *some* are considered online with strength modulated by contextual support. Consistent with the evidence in all known reading studies to date, Rapid Consideration Without Selection expects that upper-bound meanings will indeed be active, and will drive the ease of comprehension at the nominal expression *the rest* in Affirmation S3s. In particular, we predict that just when the upper-bound meaning of *some* is a potential Implicature, response latencies at this expression and following positions will be faster in Knowledgeable contexts, which support the implicature.

Next, and most critically, we are interested in determining whether the upper-bound meaning is selected online with a probability modulated by contextual support. Rapid Consideration Without Selection expects that this is not the case, and no commitment will arise during the reading of the narrative, such that online reanalysis costs will not be observed. In particular, we predict that when the upper-bound meaning of *some* is a potential Implicature, response latencies at a region which requires cancellation of that implicature will not be modulated by contextual support for the implicature. The design also permits us the opportunity to compare these latencies to cases where the same region contradicts an upper-bound Entailment. We predict that such cases should provoke uniformly slower response latencies than cases which require cancellation of an implicature. The major alternative hypothesis, that selection of upper-bound meanings is indeed happening online, would predict that cancellation in Implicature conditions will be associated with slower response latencies in Knowledgeable contexts, although presumably still faster latencies than when the same region contradicts an Entailment.

Finally, we are also concerned with whether the upper-bound meaning comes under consideration via an immediate and costly generation process, as originally concluded by Breheny et al. (2006). Rapid Consideration Without Selection holds that generation is more or less immediate, but it does not have a particular view regarding costs. What evidence we have from reading experiments (Politzer-Ahles & Fiorentino, 2013; S. Lewis, 2013; Hartshorne & Snedeker, 2014) has suggested that such experiments are not well-equipped to detect such a cost. Nevertheless, if there were a cost, we might observe it through a few analytical tools. In the first place, costly generation would predict that just when the upper-bound meaning of *some* is a potential Implicature, response latencies at *some* and following positions will be slower in Knowledgeable contexts, which support the implicature. By the same token, because we expect response latencies at *the rest* in S3 to reduce as a function of the activation of the upper-bound meaning, costly generation would predict that across all trials, the size of this reduction should be correlated with slower latencies at *some*, as observed by Bergen and Grodner (2012). As a footnote, if context-dependent cancellation costs were observed as entertained above, costly generation expects that they should also be correlated with slower latencies at *some*.

3.4.1 Method

The design and analysis plan for Experiments 5 and 6 were pre-registered (https://doi.org/10.17605/OSF.IO/4ZJD6), including the plan for norming as carried out above.[13] All materials, data, and analysis scripts are available for review in an OSF repository (https://osf.io/6t7jd/?view_only=14948ba7eea14373b3df62cd790207df).

[13] The study was planned first as a Maze study. After observation of the data, this self-paced reading experiment was carried out for the purposes of task comparison. For ease of comparison to prior work, I present the self-paced reading study first before moving on to the originally-planned Maze study.

3.4.1.1 Participants

80 native English speakers participated in the experiment on Prolific in early 2023, compensated according to a $12 hourly wage. All participants had US nationality, at least the equivalent of a high school degree, and a minimum of 20 prior submissions with an acceptance rate of 90% on the platform. Ages were within the range of 18 to 40 (with a mean of 31).

3.4.1.2 Procedure

The experiment was prepared in Ibex (Drummond, 2010), and deployed on PCIbexFarm. For each item, participants read a context sentence presented all at once (S1), followed by two critical sentences (S2 and S3) presented non-cumulatively in fixed-window, word-by-word self-paced reading. This particular variety of self-paced reading was adopted so as to allow the most minimal comparison with the Maze task of Experiment 6.

Items were followed by binary forced-choice comprehension questions. For one third of the test items, these probed information presented only in S1 (20a), to encourage non-trivial attention to the context sentence. For another third of the test items, these probed information presented only in S2 (20b). For another sixth of the test items, these probed miscellaneous information presented only in S3 (20c). In the final sixth of the test items, these were yes/no questions probing the critical interpretation of *some* as disambiguated in S3 (20d). Accuracy on each type of comprehension question for the 80 participants in the final sample is summarized in Table 3.5. A post-hoc analysis comparing accuracy on questions probing the interpretation of *some* finds that success in arriving at the S3-disambiguated meaning did not meaningfully vary across conditions. Although participants in Entailment + Cancellation conditions were presented with conflicting information, they reliably adopted responses to these questions which reflected the universal quantification introduced in S3 rather than the upper-bounded quantification introduced in S2 (>92%).[14]

(20) **Example narratives with comprehension questions from Experiment 5**

a. Earlier today, Homer was leading a small group of tourists around the sights downtown. He figured out that some of the tourists got soaked by the rain

[14] These answers were counted as correct for the purposes of the accuracy analysis reported in Table 20.

Table 3.5: Accuracy on comprehension questions in Experiment 5.

Question Target	% Correct	SE
S1	91%	1%
S2	96%	1%
S3 (Misc.)	96%	1%
S3 (*some* Interp.)	90%	1%

storm. The rest were dry because they had remembered their umbrellas.

Where were the tourists? {Downtown, The beach}

b. To prepare for his Spanish test, Sherman spent hours studying the new vocabulary items. He figured out that some of the words sounded like they do in English. The others were totally unfamiliar, which made the test somewhat challenging.

   What language did some of the words resemble? {Russian, English}

c. This morning, Garth took attendance at an important meeting with the manager. He figured out that only some of the company's accountants were there. The rest were missing because they had to audit the company's finances before the end of the quarter.

   Why were the accountants missing? {They had work, They were lazy}

d. As the new librarian, it was John's resposibility to catalog every book in the reference section. He noticed that some of the dictionaries were labeled incorrectly. The others were labeled appropriately, though a few of them had been shelved in the wrong place.

   Were all the dictionaries labeled correctly? {Yes, No}

Test items were presented in pseudo-randomized order across eight Latin-squared forms, balanced across our final sample of 80 participants. Information Status was manipulated as a between-subjects factor, so 40 participants saw test items with *some*, and 40 participants saw test items with *only some*, with each sub-sample distributed across four forms manipulating the presentation of item sets across Context and Coda manipulations. Test items were mixed with 70 filler items from various sources, including the 40 six-sentence narratives from Experiment 8, and 20 additional six-sentence narratives constructed to resemble those item sets. A final 10 three-sentence filler items were constructed to camouflage the test items from this experiment, and exemplify the Information Status condition that was not used in the test items for each participant. That is, participants who encountered the 40 test items with *only some* also saw 10 filler items with *some*, and vice versa. These fillers all finished with Cancellation-type S3s, so that each participant always saw at least some sentences with contradicted entailments, and some sentences with cancelled implicatures. All fillers were also followed by comprehension questions. Three of the six-sentence fillers were presented to participants as practice items, and four six-sentence fillers and two three-sentence fillers were reserved as "burn-in" items, presented at the beginning of the main body of the experiment before participants were shown the first critical item.

After completing all 110 trials, participants completed short exit questionnaires on their experience in the study, and demographic information regarding their language history, before receiving compensation. The procedure in entirety was estimated to take about 35 minutes.

3.4.1.3 Analysis

208 test trials in which a participant responded incorrectly to a comprehension question or a software error prevented accurate latency collection were excluded from analysis. The remaining sample includes data from 2992 critical trials.

Seven measures of word-by-word response latencies were computed for the analysis of test items. All measures relied on log-transformed response latencies, either summed or averaged per word within critical multi-word regions adopted from Bergen and Grodner (2012). The seven critical measures were: in S2, (i) summed log RTs at *some of*, (ii) summed log RTs within a first spillover region containing the following two words, e.g. *the marketing*, and (iii) summed log RTs within a second spillover region containing the following two words, e.g. *executives were*; in Affirmation S3s, (iv) summed log RTs within the two-word sentence-initial nominal instantiating reference to the complement set, e.g. *the rest*, and (v) average log RTs within the variable-length predicate which followed, e.g. *suffered a huge pay cut*; and in Cancellation S3s, (vi) summed log RTs within the sentence-initial adverbial *in fact*, and (vii) summed log RTs within the three-word clause which followed, instantiating *all* affirmation, e.g. *they all were*. While the large number of regions examined amplifies the possibility of over-interpreting illusory systematicities in reading behavior, and collapsing latencies within multi-word regions leads to imprecise measures, expectations for the timing of effects of interest were not precise enough to allow a more targeted analysis plan.

For analysis, Bayesian linear mixed-effects models were fitted to these seven measures with Stan (Stan Development Team, 2019) using the `brms` package in `R` (Bürkner, 2017, 2018) with principled weakly-informative priors, maximal random effects structures, and sum-coded predictors. Neutral protagonists, Entailment status for the upper-bounded meaning, and Cancellation codas were coded as positive. Weakly-informative non-default priors were adopted for fixed effects, $\mathcal{N}(0, 1)$, and the intercept, $\mathcal{N}(6.5n, 1)$, where n is the number of words in the region (if summed) and 6.5 is the average log-ms response latency typically observed in self-paced reading, per Nicenboim et al. (2022). Models were fit on 6 chains of 10,000 iterations (including 2,000 warmup interations), with all other `brms` parameters left to their defaults. I take model parameters whose 95% credible intervals (CRIs) do not contain 0 to indicate noteworthy effects. All models reported feature $\hat{R} = 1.00$ for the parameters of interest.

3.4.2 Results

All 80 participants retained for analysis answered comprehension questions with accuracy of greater than 75%, and recorded average response latencies of 1000ms or more on the context sentence. Other participants who had been recruited were excluded from analysis to ensure a base level of attentive comprehension. Mean accuracy in the final sample was 93%. In this section, we report the response latencies in the various regions of interest.

3.4.2.1 S2 Regions

Average log response latencies in the three critical regions of S2 at and following *some* are presented in Figure 3.2 and Table 3.6. Posterior values for $\hat{\beta}$ and σ_{β} from our model along with 95% CRIs are provided for fixed parameters of interest in Table 3.7. We observe no main effects or interactions which are credibly non-zero in any of the three regions. This includes the predicted interaction of Context and Information Status, expected to diagnose a penalty in Knowledgeable contexts specific to Implicature conditions, but estimated close to zero at *some of*, $\hat{\beta}$ = -0.00, 95% CRI = (-0.02, 0.01), the first spillover region, $\hat{\beta}$ = 0.00, 95% CRI = (-0.01, 0.02), and the second spillover region, $\hat{\beta}$ = -0.01, 95% CRI = (-0.03, 0.01). Examination of marginal contrasts in the Implicature conditions reveals that Neutral contexts did not credibly yield expected faster latencies at *some of*, $\hat{\delta}$ =

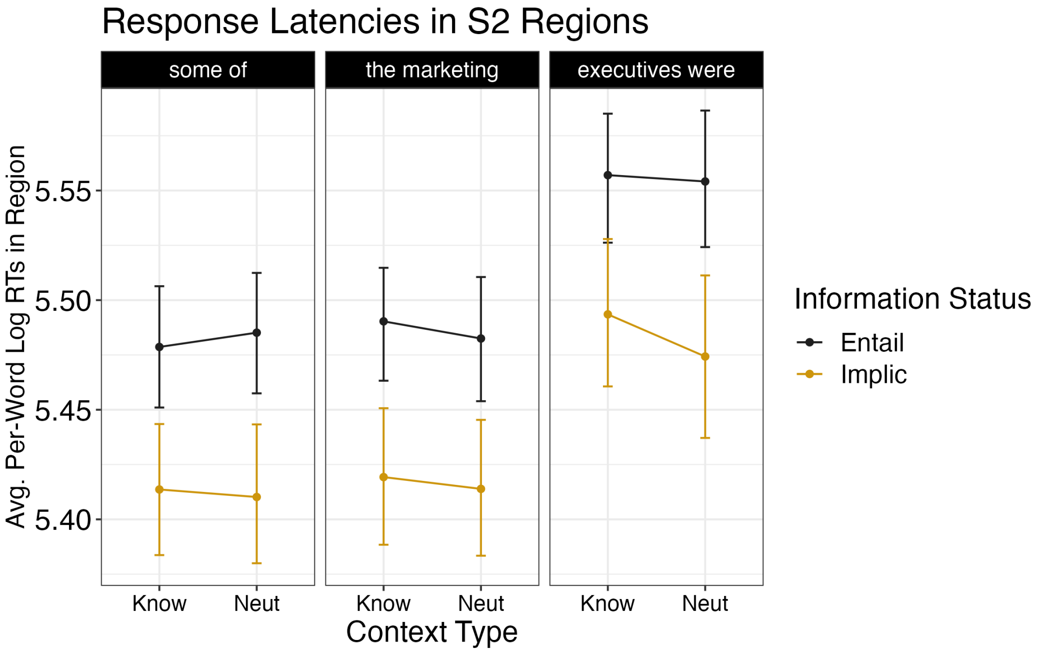

Figure 3.2: Average per-word log response latencies in the critical regions of S2 in Experiment 5, by condition. Error bars represent bootstrapped 95% confidence intervals around the mean.

0.00, $P(\delta < 0)$ = 0.57, the first spillover region, $\hat{\delta}$ = -0.01, $P(\delta < 0)$ = 0.63, or the second spillover region, $\hat{\delta}$ = -0.04, $P(\delta < 0)$ = 0.88, although the final region approaches credibility. An unexpected difference associated with Information Status is present across all three regions, such that these regions were read slower in Entailment conditions, but this does not emerge as credibly non-zero in the model in any region.

To assess the informativity of these findings with regard to our particular hypotheses, I computed Bayes factors between models which contained a term which corresponded to a particular hypothesized effect, and reduced models without that term, BF_{10} (Schad et al., 2022; Nicenboim et al., 2022). Informative priors for the models used in Bayes factor analysis were adopted from the model parameters reported in Bergen and Grodner (2012) (Table 3.8), corresponding to the weak expectation for a penalty of about 15ms particular to Implicature-triggering *some* in contexts with a Knowledgeable speaker. In the taxonomy of Lee and Wagenmakers (2013), results of the subsequent Bayes factor analysis indicate "moderate evidence" for the absence of the predicted interaction at *some of* (BF_{10} = 0.32) and in the first spillover region (BF_{10} = 0.16), and "anecdotal evidence" for its absence in the second spillover region (BF_{10} = 0.43). That is, although differences in

Table 3.6: Conditional means and measures of spread for the S2 regions in Experiment 5. Standard errors are reported over summed raw response latencies, and bootstrapped 95% confidence intervals are reported over average per-word log response latencies.

	Info. Status	Context	Sum RT	SE	Avg. Log RT	95% CI
some of	Entail	Know	520	9	5.48	(5.45, 5.50)
	Entail	Neut	543	20	5.49	(5.46, 5.51)
	Implic	Know	490	8	5.41	(5.38, 5.44)
	Implic	Neut	500	11	5.41	(5.38, 5.44)
Spill #1	Entail	Know	527	9	5.49	(5.47, 5.52)
	Entail	Neut	539	16	5.48	(5.45, 5.51)
	Implic	Know	530	31	5.42	(5.39, 5.45)
	Implic	Neut	503	9	5.41	(5.38, 5.45)
Spill #2	Entail	Know	583	15	5.56	(5.53, 5.58)
	Entail	Neut	587	15	5.55	(5.52, 5.58)
	Implic	Know	626	70	5.49	(5.46, 5.53)
	Implic	Neut	562	14	5.47	(5.44, 5.51)

Table 3.7: Bayesian linear mixed-effects models fit to summed log response latencies in the critical regions of S2 in Experiment 5. Factor levels in parentheses were coded as positive.

	Effect	Posterior $\hat{\beta}$	Posterior σ_β	95% CRI
some of	Intercept	10.89	0.07	(10.75, 11.03)
	Context (Neut.)	0.00	0.01	(-0.02, 0.02)
	Info. Status (Implic.)	-0.07	0.07	(-0.22, 0.08)
	C × IS	-0.00	0.01	(-0.02, 0.01)
Spill #1	Intercept	10.89	0.08	(10.75, 11.04)
	Context (Neut.)	-0.01	0.01	(-0.03, 0.01)
	Info. Status (Implic.)	-0.07	0.08	(-0.22, 0.08)
	C × IS	0.00	0.01	(-0.01, 0.02)
Spill #2	Intercept	11.03	0.09	(10.86, 11.20)
	Context (Neut.)	-0.01	0.01	(-0.03, 0.01)
	Info. Status (Implic.)	-0.07	0.09	(-0.24, 0.10)
	C × IS	-0.01	0.01	(-0.03, 0.01)

Table 3.8: Informative priors used for Bayes factor analysis of the S2 regions in Experiment 5, derived from Bergen and Grodner's self-paced reading results.

Effect	Distribution
Intercept	$\mathcal{N}(13.00, 0.10)$
Context (Neut.)	$\mathcal{N}(0.00, 0.01)$
Info. Status (Implic.)	$\mathcal{N}(0.04, 0.01)$
C × IS	$\mathcal{N}(-0.02, 0.01)$

response latencies were sometimes slightly in the direction expected, they were on the whole more compatible with a model which expected no particular increases in latency in the Knowledgeable-protagonist Implicature condition.

3.4.2.2 Affirmation Regions

Average log response latencies in the two critical regions of Affirmation S3s are presented in Figure 3.3 and Table 3.9. Posterior values for $\hat{\beta}$ and σ_β from our model along with 95% CRIs are provided for fixed parameters of interest in Table 3.10. We again observe no main effects or interactions which are credibly non-zero in any region, although a few near-credible trends are of interest. We observe a near-credible negative main effect of Context at *the rest*, $\hat{\beta}$ = -0.03, 95% CRI = (-0.05, 0.00), and the following predicate, $\hat{\beta}$ = -

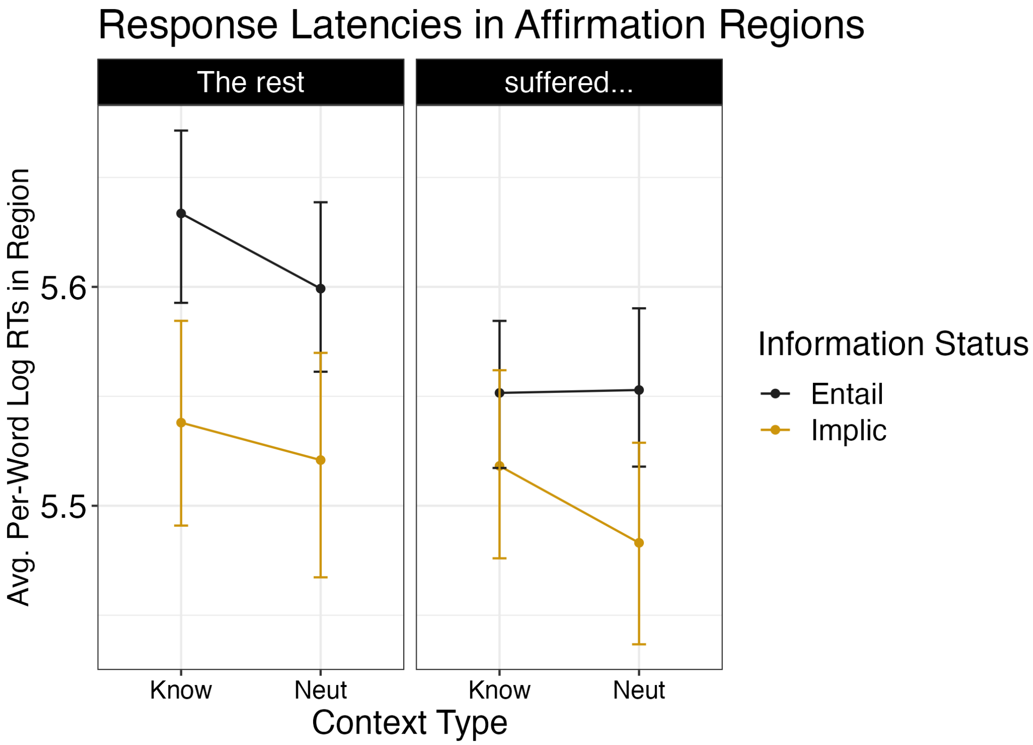

Figure 3.3: Average per-word log response latencies in the critical regions of Affirmation S3s in Experiment 5, by condition. Error bars represent bootstrapped 95% confidence intervals around the mean.

0.01, 95% CRI = (-0.02, 0.00), in the opposite direction of the facilitatory effect expected for Implicature conditions, suggesting that in Knowledgeable contexts, references to the complement set were read somewhat more slowly. At the predicate, this was qualified by a near-credible interaction of Context and Information Status, $\hat{\beta}$ = -0.01, 95% CRI = (-0.02, 0.00). Examination of marginal contrasts revealed that the slowdown in Knowledgeable conditions was driven by a large slowdown in Implicature conditions, $\hat{\delta}$ = -0.03, $P(\delta < 0)$ = 0.97, with no corresponding slowdown in the Entailment conditions, $\hat{\delta}$ = 0.00, $P(\delta < 0)$ = 0.50. While an Implicature-specific Context effect was expected, the direction here is not as predicted by any theory, or matching any previous study—I do not take it to be an effect of interest.

A trend for slower reading in Entailment conditions continues here, without emerging as credibly non-zero.

I again computed BF_{10} Bayes factors to determine the strength of evidence for or against the expected interaction, here specifying priors based on the facilitation effects

Table 3.9: Conditional means and measures of spread for the critical regions of Affirmation S3s in Experiment 5. Standard errors are reported over summed raw response latencies, and bootstrapped 95% confidence intervals are reported over average per-word log response latencies.

	Info. Status	Context	Sum RT	SE	Avg. Log RT	95% CI
the rest	Entail	Know	626	16	5.63	(5.59, 5.68)
	Entail	Neut	600	15	5.60	(5.56, 5.64)
	Implic	Know	577	17	5.54	(5.49, 5.58)
	Implic	Neut	577	16	5.52	(5.47, 5.57)
Predicate	Entail	Know	887	54	5.55	(5.52, 5.59)
	Entail	Neut	854	29	5.55	(5.52, 5.59)
	Implic	Know	861	40	5.52	(5.47, 5.56)
	Implic	Neut	824	32	5.48	(5.44, 5.53)

Table 3.10: Bayesian linear mixed-effects models fit to summed log response latencies in the critical regions of Affirmative S3s in Experiment 5. Factor levels in parentheses were coded as positive.

	Effect	Posterior $\hat{\beta}$	Posterior σ_{β}	95% CRI
the rest	Intercept	11.13	0.08	(10.98, 11.29)
	Context (Neut.)	-0.03	0.01	(-0.05, 0.00)
	Info. Status (Implic.)	-0.08	0.08	(-0.24, 0.07)
	C $\times$ IS	0.01	0.01	(-0.02, 0.04)
Predicate	Intercept	5.52	0.04	(5.44, 5.60)
	Context (Neut.)	-0.01	0.01	(-0.02, 0.00)
	Info. Status (Implic.)	-0.02	0.04	(-0.10, 0.05)
	C $\times$ IS	-0.01	0.01	(-0.02, 0.00)

Table 3.11: Informative priors used for Bayes factor analysis of Affirmative S3s in Experiment 5, derived from Bergen and Grodner's self-paced reading results.

Effect	Distribution
Intercept	$\mathcal{N}(13.00, 0.10)$
Context (Neut.)	$\mathcal{N}(0.01, 0.01)$
Info. Status (Implic.)	$\mathcal{N}(0.01, 0.01)$
C × IS	$\mathcal{N}(0.02, 0.01)$

observed in Bergen and Grodner (2012) (Table 3.11). The resulting analysis indicates anecdotal evidence against the expected interaction at *the rest* (BF_{10} = 0.97) and moderate evidence against it at the following predicate (BF_{10} = 0.14).

3.4.2.3 Cancellation Regions

Average log response latencies in the two critical regions of Cancellation S3s are presented in Figure 3.4 and Table 3.12. Posterior values for $\hat{\beta}$ and σ_β from our model along with 95% CRIs are provided for fixed parameters of interest in Table 3.13. We again observe no main effects or interactions which are credibly non-zero in any region. This includes the predicted interaction of Context and Information Status, expected to diagnose difficulty in Knowledgeable contexts specific to Implicature conditions, at *in fact*, $\hat{\beta}$ = -0.01, 95% CRI = (-0.03, 0.02) and on the following clause, $\hat{\beta}$ = 0.01, 95% CRI = (-0.03, 0.06). A near-credible negative main effect of Context was observed in the direction of the effect expected for Implicature conditions at *in fact*, $\hat{\beta}$ = -0.02, 95% CRI = (-0.04, 0.01), and the following clause, $\hat{\beta}$ = -0.03, 95% CRI = (-0.07, 0.01), suggesting that in all Knowledgeable contexts, cancellation-cuing content was read somewhat more slowly. Examination of marginal contrasts in the Implicature conditions reveals that Neutral contexts did not credibly yield expected faster latencies at *in fact*, $\hat{\delta}$ = -0.05, $P(\delta < 0)$ = 0.88 or the following clause, $\hat{\delta}$ = -0.03, $P(\delta < 0)$ = 0.67. Finally, a trend for slower reading in Entailment conditions continues here, still without emerging as credibly non-zero.

I again computed BF_{10} Bayes factors to determine the strength of evidence for or against the expected interaction, here, in the absence of known pragmatic reanalysis effects, specifying priors based on the homonymy reanalysis effects observed in Experiment 1 (Table 3.14). The resulting analysis indicates moderate evidence against expected

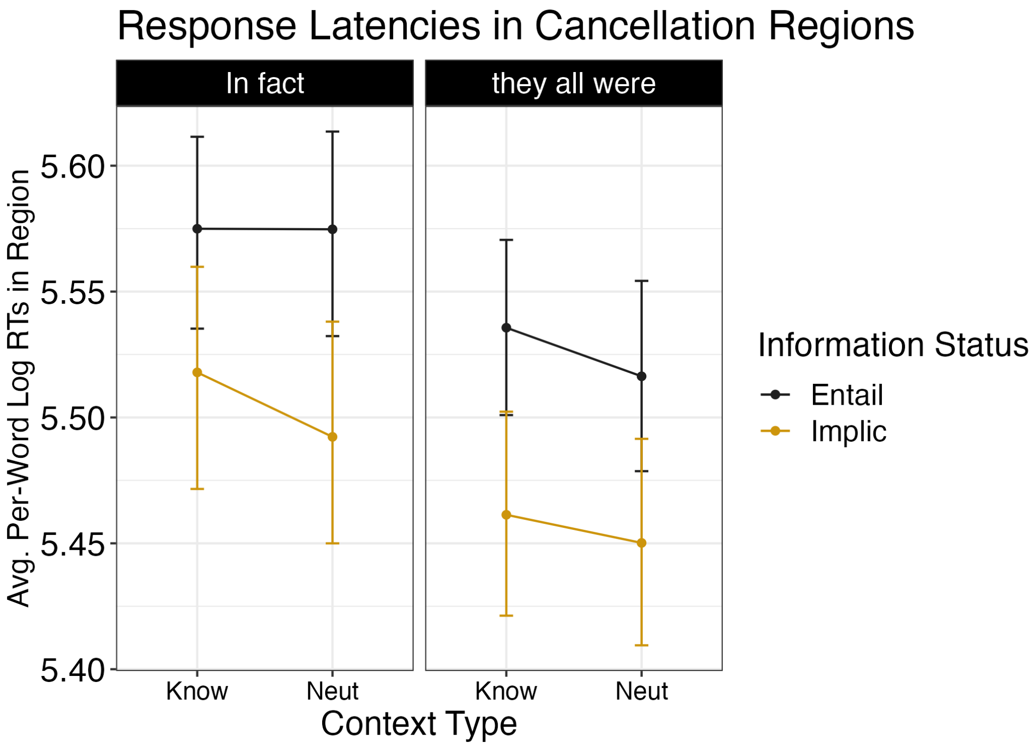

Figure 3.4: Average per-word log response latencies in the critical regions of Cancellation S3s in Experiment 5, by condition. Error bars represent bootstrapped 95% confidence intervals around the mean.

Table 3.12: Conditional means and measures of spread for the critical regions of Cancellation S3s in Experiment 5. Standard errors are reported over summed raw response latencies, and bootstrapped 95% confidence intervals are reported over average per-word log response latencies.

	Info. Status	Context	Sum RT	SE	Avg. Log RT	95% CI
in fact	Entail	Know	580	12	5.57	(5.54, 5.61)
	Entail	Neut	599	19	5.57	(5.53, 5.62)
	Implic	Know	558	13	5.52	(5.47, 5.56)
	Implic	Neut	546	14	5.49	(5.44, 5.54)
all-Clause	Entail	Know	830	16	5.54	(5.50, 5.57)
	Entail	Neut	822	20	5.52	(5.48, 5.55)
	Implic	Know	788	20	5.46	(5.42, 5.51)
	Implic	Neut	780	20	5.45	(5.41, 5.49)

Table 3.13: Bayesian linear mixed-effects models fit to summed log response latencies in the critical regions of Cancellation S3s in Experiment 5. Factor levels in parentheses were coded as positive.

	Effect	Posterior $\hat{\beta}$	Posterior σ_β	95% CRI
in fact	Intercept	11.08	0.08	(10.92, 11.23)
	Context (Neut.)	-0.02	0.01	(-0.04, 0.01)
	Info. Status (Implic.)	-0.07	0.08	(-0.22, 0.08)
	C × IS	-0.01	0.01	(-0.03, 0.02)
all-Clause	Intercept	16.47	0.11	(16.25, 16.69)
	Context (Neut.)	-0.03	0.02	(-0.07, 0.01)
	Info. Status (Implic.)	-0.10	0.11	(-0.32, 0.12)
	C × IS	0.01	0.02	(-0.03, 0.06)

Table 3.14: Informative priors used for Bayes factor analysis of Cancellation S3s in Experiment 5, derived from reanalysis costs observed for homonyms in Experiment 1.

Effect	Distribution
Intercept	$\mathcal{N}(11.00, 0.10)$
Context (Neut.)	$\mathcal{N}(-0.04, 0.01)$
Info. Status (Implic.)	$\mathcal{N}(-0.04, 0.01)$
C × IS	$\mathcal{N}(-0.04, 0.01)$

interaction at *in fact* (BF_{10} = 0.16) and strong evidence against it at the following predicate (BF_{10} = 0.06).

3.4.2.4 Within-Trial Correlation

As an additional, post-hoc examination of the data, I replicated Bergen and Grodner's correlational analysis relating response latencies for the *some of* region to response latencies for the various critical regions of S3. Following their procedure, I first fit simple linear mixed-effects models predicting average latencies in S3 from average latencies in S2 within the Entailment + Affirmation conditions, where that dependency is presumed to resemble the standard dependency between any two regions within the same trial (i.e. by assumption there is no implicature calculation going on). Indeed, these models recorded the expected positive relationships in the case of the initial regions of S3, $\hat{\beta}$ = 0.43, 95% CRI = (0.35, 0.50) and the second regions of S3, $\hat{\beta}$ = 0.37, 95% CRI = (0.31, 0.43). These models were then used to generate predicted S3 latencies from S2 latencies in the Implica-

Table 3.15: Bayesian linear mixed-effects models fit to residual response latencies in the critical regions of S3, using S2 latencies as a predictor, in Experiment 5.

	Effect	Posterior $\hat{\beta}$	Posterior σ_β	95% CRI
Region 1	Intercept	-1.00	0.15	(-1.29, -0.71)
	S2 Latency	0.17	0.03	(0.12, 0.23)
	Coda (Cancel)	-0.04	0.09	(-0.21, 0.13)
	S2 × Coda	0.01	0.02	(-0.02, 0.04)
Region 2	Intercept	-1.46	0.13	(-1.79, -1.21)
	S2 Latency	0.26	0.02	(0.22, 0.31)
	Coda (Cancel)	0.15	0.08	(-0.00, 0.29)
	S2 × Coda	-0.03	0.01	(-0.06, -0.00)

ture conditions, and residuals were extracted by comparing these predictions to the actual values. Final, more complex models were fit to predict these residuals based on the S2 latencies, the Coda condition, and a potential interaction between the two (Table 3.15). I observe that even after residualization, strong positive relationships between S2 and the S3 regions remain, e.g. at the initial regions, $\hat{\beta}$ = 0.17, 95% CRI = (0.12, 0.23). In the second region, I observe a near-credible main effect of Coda, $\hat{\beta}$ = 0.15, 95% CRI = (-0.00, 0.29), such that Cancellation continuations were generally associated with slower latencies, and a small just-credible interaction, $\hat{\beta}$ = -0.03, 95% CRI = (-0.06, -0.00), suggesting a slightly weaker positive relationship between S2 and Coda latencies in Cancellation items, $\hat{\delta}$ = 0.46, $P(\delta > 0)$ = 0.99, compared to Affirmation items, $\hat{\delta}$ = 0.59, $P(\delta > 0)$ = 0.99. On the whole, in this experiment, latencies at *some* predict latencies later on in the narrative with a slope which exceeds the pattern observable in the *only some* conditions.

3.4.3 Discussion

In a self-paced reading study attempting to extend the design of Bergen and Grodner (2012), I observe none of the effects found in that study. Bayes factors allow us to quantify the nature and quantity of this evidence, suggesting moderate evidence against both the generation and the facilitation effects observed there, and somewhere between strong evidence against to weak evidence in favor of a cancellation effect. These results are largely in line with the Rapid Consideration Without Selection hypothesis developed here. For one, the data concurs with the absence of generation effects in the small majority of reading studies that have looked for them. This extends to the failure to replicate the patterns in their correlational analysis: if reading times on *some* can't be attributed to a costly generation process, then it follows that there should be no particular special relationship between latencies on that region and our S3 regions.[15] We also continue to observe no convincing evidence for costly cancellation, even after attempting to surmount the confound I argue is present in the Bergen and Grodner design. To be specific, models do not estimate credibly non-zero parameters for the expected interactions, and Bayes factor analysis reveals evidence for the expected interaction is at best anecdotal, on just one short region which itself merely serves as a possible early cue to cancellation. Crucially, the absence of any cancellation cost is as predicted in the absence of selection.

However, a few details resist easy interpretation. In the first place, I do not replicate the facilitation effect observed throughout the literature on complement reference in implicature-supporting contexts. I can see two possibilities for interpreting the absence of this effect. On the one hand, it may be possible that comprehenders sometimes are less motivated to use unselected meanings to anticipate and facilitate later comprehension. Perhaps rapid consideration is a typical phenomenon, but in this particular sample, either that consideration or its influence on later processes is withheld for an unknown reason. On the other hand, in a more particular proposal, it may be possible that the design of this study has produced a substantially different processing environment than other studies. The use of factive embedding makes the relevant implicature more complex—perhaps this level of inference no longer has the same online status as the enrichment of matrix *some*. Or perhaps enrichment proceeded as expected, but despite the outcome of my norming study, this particular enrichment is not as readily modulated by protagonist knowledgeability. If any of these latter possibilities are true, we should hesitate to infer anything from these results about the processing of more standard cases of *some*, not to mention implicature in general.

Another troubling observation is the consistent relative difficulty of *only some* narratives relative to *some* narratives. Hypotheses that involve costly generation or selection of an enriched meaning of *some* online would expect the opposite effect, with *only some* narratives being read relatively faster. Other evidence here suggests those hypotheses are incorrect, and no such costly process happens during incremental comprehension, but even under the alternative proposal of Rapid Consideration Without Selection, we would expect parity, not slower reading for *only some*. And to be sure, none of the pre-

[15] I have no confident explanation for the residual positive correlation, where we would perhaps expect the absence of any systematic dependence remaining between S3 residuals and S2 latencies. Perhaps this reveals that the *only some* conditions were a poor control in this case, a concern echoed below.

vious studies reviewed above which used this control have found it to provoke elevated reading times. It would be reasonable to avoid over-interpreting this trend, as it is only a trend, and, because this manipulation was between-subjects, it could be attributable to an accident of sampling. If we sought a more detailed explanation, I would suggest that it might arise from a problem particular to this design, where *only* is at the left edge of an attitude complement. In this position, the relative linear order of *that* and *only* has important consequences for the interpretation of *only*: *Petra realized that only some executives were fired* has a single meaning, where *only* negates alternatives to *some*, but *Petra realized* ***only that*** *some executives were fired* now relates *only* to the alternatives for the entire embedded clause. Participants are rapidly fallible to this sort of transposition in various comprehension and judgment tasks, especially when words are short and matched in length (Poppels & Levy, 2016; Mirault et al., 2018; K.-J. Huang & Staub, 2021). Crucially, considering an exchange in this particular case could allow participants in the *only some* condition to rescue an interpretation in the case of the contradictions which they encountered with some regularity. While the string as presented has no coherent interpretation, strings like *Petra realized only that some executives were fired. In fact all of them were.* have a quite coherent reading where *only* highlights the things Petra didn't realize, which might include in particular that all the executives were fired. If the presence of these contradictions drove participants to consider exchanges like this, and generally reduce their confidence in the input as they have encoded it, this could explain their slowdown on sentences with *only*.[16]

On the whole, these unexplained patterns raise some doubt about the possibility to interpret evidence against costly generation and costly reanalysis. The Maze task reported in the next section will offer some insight, after which I will return to discuss what to make of these self-paced reading results in the general discussion.

[16] Note that there is a simpler explanation that seems less feasible. One might suggest that participants in the *only some* condition encountered more contradictions than the participants in the *some* condition. They might, then, have slowed down because they received more challenging input. Indeed, the 10 fillers which attempted to counter-balance the number of contradictions for each participant did not wholly offset this differential: participants in the *only some* condition saw a total of 20 sentences featuring *only* and a contradiction, while *some* participants saw this pattern only in the 10 fillers. Could encountering 10 further contradictions drive a global slowdown? As we will see, the task-dependence of this global slowdown make the exchange-related explanation more appealing.

3.5 Experiment 6: Scalar implicatures in the Maze

As reviewed in Chapter 2, the Maze task has been observed to allow for tighter localization and larger effect sizes than self-paced reading (Forster et al., 2009; Witzel et al., 2012), attributed to the utility of a more careful incremental reading strategy. I demonstrated there that this careful strategy can include earlier decision-making for uncertain meaning, e.g. motivating atypical, immediate sense selection for polysemes in neutral contexts.

With this in mind, a comparison with the Maze task can offer a few key insights towards the research questions of this chapter. First, the hypothesis of Rapid Consideration Without Selection suggests that the decision between enriched and unenriched meanings for *some*, like the decision between senses of a polyseme, is regularly postponed in online comprehension. While there are important differences—e.g. that sense selection does happen online in certain specifying contexts, and that it is not postponed beyond a sentence boundary—this parallel leaves open the question of how strategy interacts with the timing of the decision to select an enriched meaning. Is the apparent delay strategic in nature? Can earlier selection be motivated when it might be useful? If so, we would expect that garden-path-like cancellation costs could emerge in the Maze where they have not in Bergen and Grodner (2012) or Experiment 5, so that Cancellation regions receive slower latencies in a Knowledgeable context, particularly when the cancelled upper-bound meaning is an Implicature.

But the surprising absence of evidence for implicature-induced facilitation for complement reference in Experiment 5 raised a theory-independent worry about performance in the self-paced reading task. Perhaps participants were engaging only shallowly with the text, failing to generate expectations specific to the possible enriched interpretation of *some*. If this is the case, we might expect that the general pressure for incrementality in the Maze would lead to more robust expectations, and thus a chance to at least observe the expected facilitation effect, so that Affirmation regions receive faster latencies in a Knowledgeable context, particularly for conditions where the complement-favoring upper-bound meaning is a context-dependent Implicature.

Different combinations of these two effects offer different insight into the results of Experiment 5 and the evidence for Rapid Consideration Without Selection. If both facilitation and reanalysis effects emerge in the Maze, it is evidence that, at least when strategic, comprehenders will engage in rapid selection. If facilitation emerges in the absence of reanalysis, it is strong evidence that consideration and selection can be dissociated, and further suggests that postponed selection of enriched meaning is not as flexible as postponed sense selection for polysemy. If we continue here to observe neither effect, we fail to provide evidence for any consideration of enriched meaning online, and raise doubts about whether this design is well-suited to investigating the timecourse of typical pragmatic enrichment.

3.5.1 Method

All materials, data, and analysis scripts are available for review in an OSF repository (https://osf.io/6t7jd/?view_only=14948ba7eea14373b3df62cd790207df).

3.5.1.1 Participants

80 native English speakers participated in the experiment on Prolific in 2023, compensated according to a $12 hourly wage. All participants had US nationality, at least the equivalent of a high school degree, and a minimum of 20 prior submissions with an acceptance rate of 90% on the platform. Ages were within the range of 18 to 40 (with a mean of 31).

3.5.1.2 Procedure

The experiment was prepared in Ibex (Drummond, 2010), and deployed on PCIbexFarm. For each item, participants read a context sentence presented all at once (S1), followed by two critical sentences (S2 and S3) presented in a Maze task.

Following Boyce et al. (2020), Maze foils were sampled from length-matched high-surprisal words with reference to the language model of Gulordava et al. (2018). Separate foil strings were prepared for S2, Affirmation S3s, and Cancellation S3s, each beginning with a foil of “x-x-x” due to the high entropy of sentence-initial positions. S2 foils were initially prepared for Entailment conditions, with *only some*, and foils for the Implicature conditions, which omitted *only*, simply omitted the foil for *only* as well. Foils that were too repetitive or were judged as plausible continuations were replaced by hand. In previous studies, high error rates on capitalized words have been observed, e.g. on named characters and place names, possibly because many phonotactically licit words in English are somewhat plausible names. To prevent excess data loss for this reason, foils for capitalized words were always same-length strings of *x*, matched for capitalization (e.g. *Xxxxxx* served as the foil for *Polish*). An example set of foil strings is given in (21).

(21) **Example foils for Maze portions of Experiment 6**

a. She realized that (only) some of the marketing executives were fired.
x-x-x unlikely lurches (goes) spot oh seat celebrate comprehend hill clock.

b. The rest suffered a huge pay cut, which seemed fair.
x-x-x gone stunning idea whom kids holy, wavers bishop lady.

c. In fact, they all were, which seemed fair.
x-x-x lose, anti sir yeah, china miller cent.

As in Experiments 2 and 4, in order to encourage accurate performance of the Maze task, participants saw a counter at the top of the screen during each Maze decision measuring how many targets they had chosen correctly without an error. This number reset to 0 when participants chose a foil, and the ongoing sentence was immediately terminated, moving prematurely to the comprehension question.

Other than the mechanics of the Maze task, presentation, form assignment, and randomization was carried out as in Experiment 1. The same questions, fillers, practice, and burn-in items were used. This procedure was estimated to take about 60 minutes.

3.5.1.3 Analysis

898 test trials in which a participant responded incorrectly to a Maze decision or a comprehension question were excluded from analysis of response latencies. The remaining sample includes data from 2302 test trials. Maze decision errors were analyzed separately as a secondary measure of incremental difficulty, but revealed no patterns of interest. Critical response latency measures and their analysis were computed using the same procedures as in Experiment 5.

3.5.2 Results

All 80 participants retained for analysis answered comprehension questions with accuracy of greater than 75% and an average Maze depth of greater than 50%, which is to say that more often than not, they successfully made it past the halfway point of the Maze stimuli. Other participants who had been recruited were excluded from analysis to ensure a base level of attentive comprehension. Mean comprehension accuracy in the final sample was 89%, Maze completion rate was 75%, and average Maze depth was 86%. In this section, I report the response latencies in the various regions of interest.

3.5.2.1 S2 Regions

Average log response latencies in the three critical regions of S2 at and following *some* are presented in Figure 3.5 and Table 3.16. Posterior values for $\hat{\beta}$ and σ_β from our model along with 95% CRIs are provided for fixed parameters of interest in Table 3.17. We observe no main effects or interactions which are credibly non-zero in any of the three regions. This includes the predicted interaction of Context and Information Status, expected to diagnose a penalty in Knowledgeable contexts specific to Implicature conditions, but estimated close to zero at *some of*, $\hat{\beta}$ = 0.00, 95% CRI = (-0.04, 0.04), the first spillover region, $\hat{\beta}$ = 0.01, 95% CRI = (-0.01, 0.03), and the second spillover region, $\hat{\beta}$ = 0.00, 95% CRI = (-0.01, 0.02). Examination of marginal contrasts in the Implicature conditions reveals that Neutral contexts did not credibly yield expected faster latencies at *some of*, $\hat{\delta}$ = 0.02, $P(\delta < 0)$ = 0.22, or the second spillover region, $\hat{\delta}$ = 0.02, $P(\delta < 0)$ = 0.27, and indeed yielded credibly slower latencies in the first spillover region, $\hat{\delta}$ = 0.06, $P(\delta > 0)$ = 0.99, driving a main effect of context there which approached a credible positive value, $\hat{\beta}$ = 0.02, 95% CRI = (-0.00, 0.04). If anything, it would appear that comprehenders struggled somewhat in their reading of this region when context did not support an implicature, an effect in the opposite direction of predictions under a theory of Rapid Enrichment.

Subsequent Bayes factor analysis using priors informed by the Bergen and Grodner effects (see §3.4.2.1) indicated consistent moderate evidence for the absence of the predicted interaction at *some of* (BF_{10} = 0.11), the first spillover region (BF_{10} = 0.10), and the second spillover region (BF_{10} = 0.16). Note that at *some of*, there was, nevertheless, anecdotal evidence for the presence of overall slower processing in implicature conditions (BF_{10} = 1.27).

3.5.2.2 Affirmation Regions

Average log response latencies in the two critical regions of Affirmation S3s are presented in Figure 3.6 and Table 3.18. Posterior values for $\hat{\beta}$ and σ_β from our model along with 95% CRIs are provided for fixed parameters of interest in Table 3.19. We again observe

Table 3.16: Conditional means and measures of spread for the S2 regions in Experiment 6. Standard errors are reported over summed raw response latencies, and bootstrapped 95% confidence intervals are reported over average per-word log response latencies.

	Info. Status	Context	Sum RT	SE	Avg. Log RT	95% CI
some of	Entail	Know	1262	13	6.41	(6.40, 6.43)
	Entail	Neut	1265	14	6.41	(6.40, 6.43)
	Implic	Know	1291	13	6.43	(6.41, 6.45)
	Implic	Neut	1308	16	6.44	(6.42, 6.46)
Spill #1	Entail	Know	1501	19	6.56	(6.54, 6.58)
	Entail	Neut	1527	19	6.57	(6.55, 6.59)
	Implic	Know	1508	31	6.54	(6.52, 6.56)
	Implic	Neut	1550	22	6.58	(6.55, 6.60)
Spill #2	Entail	Know	1633	22	6.64	(6.62, 6.67)
	Entail	Neut	1636	21	6.65	(6.63, 6.67)
	Implic	Know	1607	27	6.62	(6.60, 6.64)
	Implic	Neut	1823	208	6.64	(6.62, 6.66)

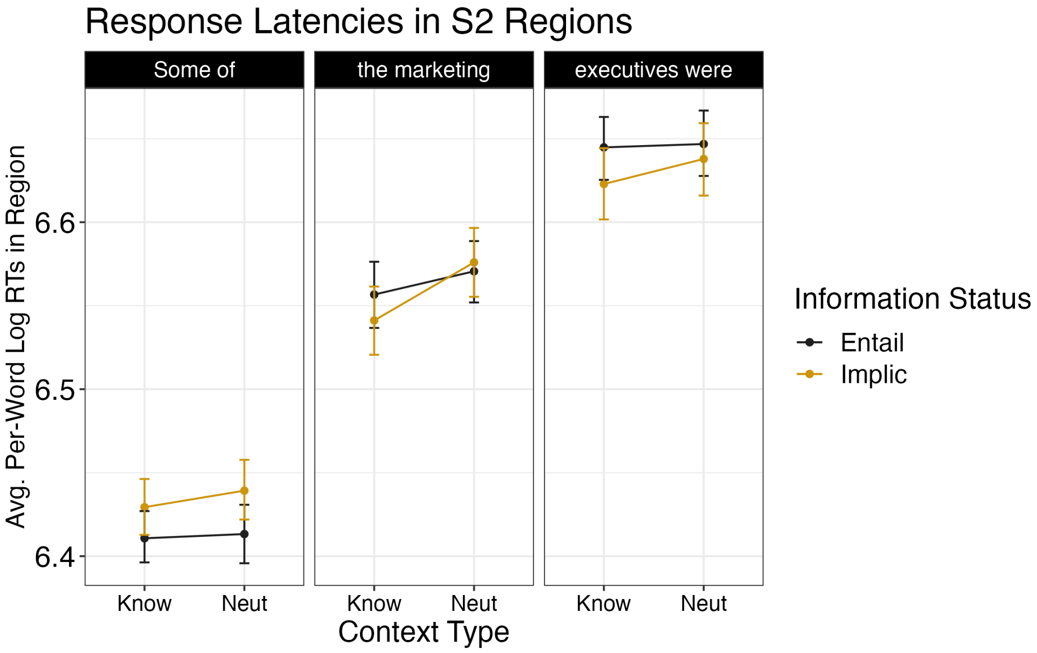

Figure 3.5: Average per-word log response latencies in the critical regions of S2 in Experiment 6, by condition. Error bars represent bootstrapped 95% confidence intervals around the mean.

no main effects or interactions which are credibly non-zero in any region. This includes the predicted interaction of Context and Information Status, expected to diagnose facilitation in Knowledgeable contexts specific to Implicature conditions, but estimated close to zero at *the rest*, $\hat{\beta}$ = -0.01, 95% CRI = (-0.03, 0.02) and the following predicate, $\hat{\beta}$ = -0.00, 95% CRI = (-0.01, 0.01). Examination of marginal contrasts in the Implicature conditions reveals that Neutral contexts did not credibly yield expected slower latencies at *the rest*, $\hat{\delta}$ = 0.00, $P(\delta > 0)$ = 0.44 or the predicate, $\hat{\delta}$ = 0.00, $P(\delta > 0)$ = 0.52.

Subsequent Bayes factor analysis indicated moderate evidence for the absence of the predicted interaction at *the rest* (BF_{10} = 0.13), and strong evidence for the absence of the predicted interaction at the following predicate (BF_{10} = 0.09).

3.5.2.3 Cancellation Regions

Average log response latencies in the two critical regions of Cancellation S3s are presented in Figure 3.7 and Table 3.20. Posterior values for $\hat{\beta}$ and σ_β from our model along with 95% CRIs are provided for fixed parameters of interest in Table 3.21. We again observe

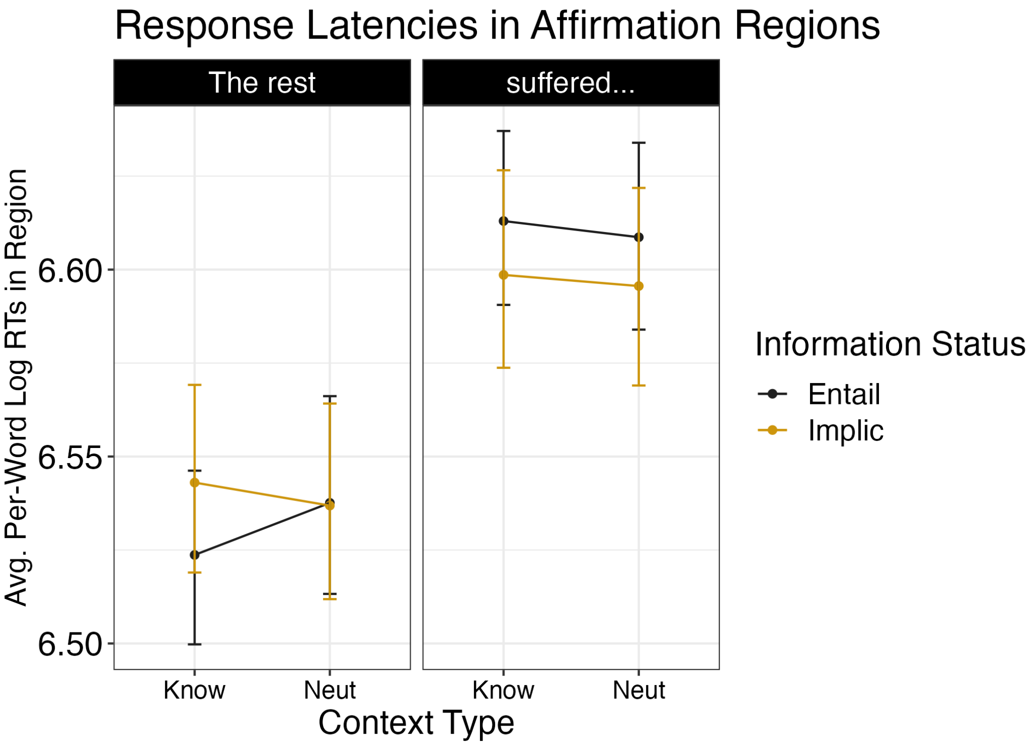

Figure 3.6: Average per-word log response latencies in the critical regions of Affirmation S3s in Experiment 6, by condition. Error bars represent bootstrapped 95% confidence intervals around the mean.

Table 3.17: Bayesian linear mixed-effects models fit to summed log response latencies in the critical regions of S2 in Experiment 6. Factor levels in parentheses were coded as positive.

	Effect	Posterior $\hat{\beta}$	Posterior σ_β	95% CRI
some of	Intercept	12.84	0.03	(12.79, 12.89)
	Context (Neut.)	0.01	0.01	(-0.01, 0.02)
	Info. Status (Implic.)	0.02	0.02	(-0.02, 0.07)
	C × IS	0.00	0.01	(-0.01, 0.02)
Spill #1	Intercept	13.12	0.04	(13.04, 13.21)
	Context (Neut.)	0.02	0.01	(-0.00, 0.04)
	Info. Status (Implic.)	0.00	0.03	(-0.06, 0.06)
	C × IS	0.01	0.01	(-0.01, 0.03)
Spill #2	Intercept	13.27	0.05	(13.18, 13.37)
	Context (Neut.)	0.01	0.01	(-0.02, 0.03)
	Info. Status (Implic.)	-0.01	0.03	(-0.07, 0.04)
	C × IS	0.00	0.01	(-0.01, 0.02)

no main effects or interactions which are credibly non-zero in any region. This includes the predicted interaction of Context and Information Status, expected to diagnose difficulty in Knowledgeable contexts specific to Implicature conditions, but estimated close to zero at *in fact*, $\hat{\beta}$ = 0.01, 95% CRI = (-0.01, 0.03) and the following clause, $\hat{\beta}$ = 0.01, 95% CRI = (-0.02, 0.05). Examination of marginal contrasts in the Implicature conditions reveals that Neutral contexts did not credibly yield expected faster latencies at *in fact*, $\hat{\delta}$ = 0.04, $P(\delta < 0)$ = 0.08 or the following clause, $\hat{\delta}$ = 0.06, $P(\delta < 0)$ = 0.11. A near-credibly-negative main effect of Information Status within the *all* clause may indicate some general difficulty in Entailment conditions, $\hat{\beta}$ = -0.08, 95% CRI = (-0.17, 0.01), where this region introduced a proposition which contradicted the meaning of S2.

Subsequent Bayes factor analysis indicated extreme evidence for the absence of the predicted interaction at *in fact* ($BF_{10} <$ 0.01), and strong evidence for the absence of the predicted interaction at the following predicate (BF_{10} = 0.02). Note that in this latter region, there was, nevertheless, moderate evidence for the presence of overall slower processing in Entailment conditions (BF_{10} = 3.02).

3.5.2.4 Within-Trial Correlation

I again repeated Bergen and Grodner's correlational analysis relating response latencies for the *some of* region to response latencies for the various critical regions of

Table 3.18: Conditional means and measures of spread for the critical regions of Affirmation S3s in Experiment 6. Standard errors are reported over summed raw response latencies, and bootstrapped 95% confidence intervals are reported over average per-word log response latencies.

	Info. Status	Context	Sum RT	SE	Avg. Log RT	95% CI
the rest	Entail	Know	1433	25	6.52	(6.50, 6.55)
	Entail	Neut	1450	26	6.54	(6.51, 6.56)
	Implic	Know	1459	22	6.54	(6.52, 6.57)
	Implic	Neut	1467	28	6.54	(6.51, 6.56)
Predicate	Entail	Know	2305	73	6.61	(6.59, 6.64)
	Entail	Neut	2289	72	6.61	(6.58, 6.63)
	Implic	Know	2259	69	6.60	(6.57, 6.62)
	Implic	Neut	2329	78	6.60	(6.57, 6.62)

Table 3.19: Bayesian linear mixed-effects models fit to summed log response latencies in the critical regions of Affirmative S3s in Experiment 6. Factor levels in parentheses were coded as positive.

	Effect	Posterior $\hat{\beta}$	Posterior σ_β	95% CRI
the rest	Intercept	13.06	0.03	(13.00, 13.12)
	Context (Neut.)	0.00	0.01	(-0.02, 0.03)
	Info. Status (Implic.)	0.01	0.03	(-0.04, 0.07)
	C × IS	-0.01	0.01	(-0.03, 0.02)
Predicate	Intercept	6.60	0.02	(6.56, 6.64)
	Context (Neut.)	0.00	0.01	(-0.01, 0.01)
	Info. Status (Implic.)	-0.01	0.01	(-0.03, 0.02)
	C × IS	-0.00	0.01	(-0.01, 0.01)

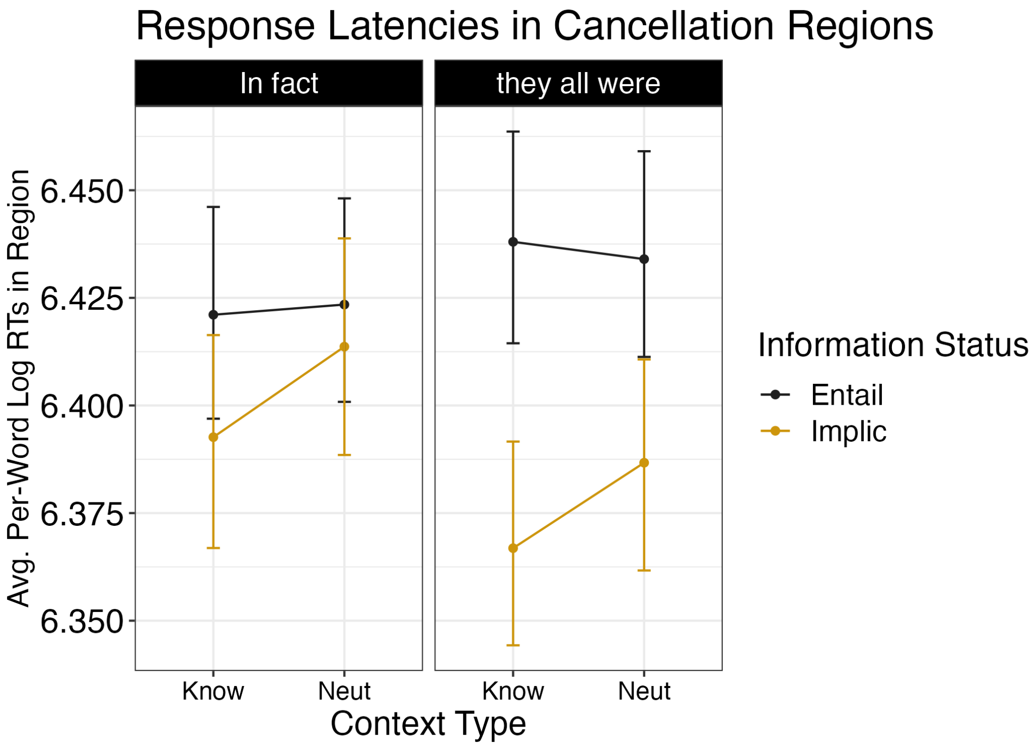

Figure 3.7: Average per-word log response latencies in the critical regions of Cancellation S3s in Experiment 6, by condition. Error bars represent bootstrapped 95% confidence intervals around the mean.

Table 3.20: Conditional means and measures of spread for the critical regions of Cancellation S3s in Experiment 6. Standard errors are reported over summed raw response latencies, and bootstrapped 95% confidence intervals are reported over average per-word log response latencies.

	Info. Status	Context	Sum RT	SE	Avg. Log RT	95% CI
in fact	Entail	Know	1281	20	6.42	(6.40, 6.44)
	Entail	Neut	1285	26	6.42	(6.40, 6.45)
	Implic	Know	1258	30	6.39	(6.37, 6.42)
	Implic	Neut	1305	47	6.41	(6.39, 6.44)
all-Clause	Entail	Know	1953	28	6.44	(6.41, 6.46)
	Entail	Neut	1954	37	6.43	(6.41, 6.46)
	Implic	Know	1817	30	6.37	(6.34, 6.39)
	Implic	Neut	1875	37	6.39	(6.36, 6.41)

Table 3.21: Bayesian linear mixed-effects models fit to summed log response latencies in the critical regions of Cancellation S3s in Experiment 6. Factor levels in parentheses were coded as positive.

	Effect	Posterior $\hat{\beta}$	Posterior σ_β	95% CRI
in fact	Intercept	12.82	0.03	(12.77, 12.88)
	Context (Neut.)	0.01	0.01	(-0.01, 0.03)
	Info. Status (Implic.)	-0.02	0.03	(-0.07, 0.04)
	C $\times$ IS	0.01	0.01	(-0.01, 0.03)
all-Clause	Intercept	19.22	0.05	(19.13, 19.31)
	Context (Neut.)	0.02	0.02	(-0.02, 0.05)
	Info. Status (Implic.)	-0.08	0.05	(-0.17, 0.01)
	C $\times$ IS	0.01	0.02	(-0.02, 0.05)

S3. Simple linear models fit to the Entailment + Affirmation conditions again recorded the expected positive relationships between latencies at *some of* and the initial regions of S3, $\hat{\beta}$ = 0.34, 95% CRI = (0.26, 0.43), and the second regions of S3, $\hat{\beta}$ = 0.27, 95% CRI = (0.19, 0.35). A second pair of models were fit to predict residual variance in the Implicature conditions, based on the S2 latencies, the Coda condition, and a potential interaction between the two (Table 3.22). If latencies at *some of* reflected the generation of an upper-bound meaning which facilitates processing of reference to the complement set, and engenders difficult reanalysis, we would expect a negative relationship here for Affirmation S3s and a positive relationship with Cancellation S3s. In contrast, I observe a general negative relationship between S2 and the first S3 regions, $\hat{\beta}$ = -0.10, 95% CRI = (-0.16, -0.04), which falls below credibility in the second regions, $\hat{\beta}$ = -0.04, 95% CRI = (-0.09, 0.02). Marginal comparisons in the first regions reveal this is especially strong for the Cancellation regions, $\hat{\delta}$ = -0.14, $P(\delta < 0)$ = 0.99, compared to the Affirmation regions, $\hat{\delta}$ = -0.06, $P(\delta < 0)$ = 0.93, driving a trending interaction with Coda opposite the predicted direction, $\hat{\beta}$ = -0.04, 95% CRI = (-0.09, 0.01). It would appear that the amount of processing time devoted to the critical implicature trigger in this experiment was predictive of facilitation in downstream regions, perhaps especially on regions which necessitate a cancellation of the potential implicature.

If this is some type of facilitation effect arising from effort earlier in the sentence, it's useful to know whether it arises from standard comprehension principles, or if it might be related to participants' growing experience with the shape of this experiment's stimuli. A secondary analysis was conducted for the first regions in S3, adding by-condition exposure counts as a predictor for residual latency. This model captures a near-credible

Table 3.22: Bayesian linear mixed-effects models fit to residual response latencies in the critical regions of S3, using S2 latencies as a predictor, in Experiment 6.

	Effect	Posterior $\hat{\beta}$	Posterior σ_{β}	95% CRI
Region 1	Intercept	0.59	0.20	(0.20, 0.99)
	S2 Latency	-0.10	0.03	(-0.16, -0.04)
	Coda (Cancel)	0.19	0.17	(-0.14, 0.52)
	S2 × Coda	-0.04	0.03	(-0.09, 0.01)
Region 2	Intercept	0.11	0.19	(-0.27, 0.48)
	S2 Latency	-0.04	0.03	(-0.09, 0.02)
	Coda (Cancel)	-0.09	0.16	(-0.42, 0.23)
	S2 × Coda	-0.00	0.03	(-0.05, 0.05)

Table 3.23: Supplementary Bayesian linear mixed-effects model fit to residual response latencies in the first critical regions of S3, using S2 latencies and by-condition exposure counts as a predictor, in Experiment 6.

Effect	Posterior $\hat{\beta}$	Posterior σ_{β}	95% CRI
Intercept	0.49	0.36	(-0.21, 1.19)
S2 Latency	-0.07	0.06	(-0.18, 0.04)
Coda (Cancel)	-0.20	0.33	(-0.84, 0.45)
Exposure	0.09	0.06	(-0.03, 0.20)
S2 × Coda	0.02	0.05	(-0.08, 0.12)
S2 × Exp	-0.02	0.01	(-0.03, 0.00)
Coda × Exp	0.08	0.06	(-0.03, 0.19)
S2 × C × E	-0.01	0.01	(-0.03, 0.00)

first-order interaction between the effect of S2 latencies and the number of exposures, $\hat{\beta}$ = -0.02, 95% CRI = (-0.03, 0.00), such that as exposures increase, so does the strength of this facilitatory effect. This is qualified by a near-credible second-order interaction with Coda, $\hat{\beta}$ = -0.01, 95% CRI = (-0.03, 0.00), capturing that this is driven almost exclusively by Cancellation conditions. Marginal comparisons reveal a strong facilitatory relationship in Affirmation conditions at first exposure, $\hat{\delta}$ = -0.10, $P(\delta < 0)$ = 0.94 increasing only slightly by the tenth and final exposure, $\hat{\delta}$ = -0.12, $P(\delta < 0)$ = 0.95, in contrast with Cancellation conditions, where there is a small, non-credible facilitatory relationship at first exposure, $\hat{\delta}$ = -0.08, $P(\delta < 0)$ = 0.89, which increases many times over by the tenth and final exposure, $\hat{\delta}$ = -0.34, $P(\delta < 0)$ = 0.99. It would seem that the facilitatory effect in Affirmation sentences is a property of typical behavior, while the larger facilitatory effect in Cancellation sentences is the result of experience with these stimuli.

Table 3.24: Accuracy on comprehension questions among trials with accurate Maze responses in Experiment 6.

Question Target	% Correct	SE
S1	91%	1%
S2	96%	1%
S3 (Misc.)	99%	0%
S3 (*some* Interp.)	92%	1%

3.5.2.5 Comprehension Questions

Patterns of accuracy on comprehension questions in this experiment also provide some potential useful insights. Overall accuracy (Table 3.24) was more or less equivalent to Experiment 5, with particular difficulty with questions probing information from the context sentence S1 and the proper interpretation of *some* given S3. Conditional accuracy data for these critical questions is given in Table 3.25. A post-hoc analysis over this accuracy data was performed using a Bayesian logistic mixed-effects model in `brms` (Table 3.26). Accuracy was predicted by an interaction between Context and Information Status, $\hat{\beta}$ = 0.55, 95% CRI = (0.08, 1.05), which was in turn qualified by a near-credible second-order interaction with Coda, $\hat{\beta}$ = -0.33, 95% CRI = (-0.82, 0.14). Marginal comparisons over Context reveal that Entailment + Affirmation stimuli were harder in Neutral contexts, $\hat{\delta}$ = -1.79, $P(\delta < 0)$ = 0.98, while Implicature + Cancellation stimuli were easier in Neutral contexts, $\hat{\delta}$ = 1.51, $P(\delta > 0)$ = 0.97. Other types of stimuli were less dependent on context. Implicature + Cancellation was a condition where Context was indeed expected to modulate the likelihood of drawing the implicature, and thus the difficulty of cancellation: this result suggests that participants were indeed sensitive to the context manipulation in their ability to arrive at a final interpretation of the sentence. Context-mediated difficulty is less expected in the case of the former stimuli, where an upper-bound meaning should have been straightforwardly entailed, independent of context. This seems to demonstrate that participants reasoned about entailed upper-bound meaning using other sources of information.

3.5.3 Discussion

In a Maze study investigating the same stimuli as Experiment 5, I again observe the general lack of the effects reported in Bergen and Grodner (2012). The critical contextual manipulation again had no credible effect on response latencies at the implicature

Table 3.25: Accuracy on critical implicature-probing comprehension questions in Experiment 6.

Info. Status	Context	Coda	% Correct	SE
Entail	Know	Affirm	98%	2%
Entail	Neut	Affirm	85%	5%
Implic	Know	Affirm	90%	4%
Implic	Neut	Affirm	95%	3%
Entail	Know	Cancel	91%	4%
Entail	Neut	Cancel	91%	4%
Implic	Know	Cancel	91%	4%
Implic	Neut	Cancel	95%	3%

Table 3.26: Bayesian logistic mixed-effects model fit to accuracy on comprehension questions targeting the final interpretation of *some* in Experiment 6.

Effect	Posterior $\hat{\beta}$	Posterior σ_β	95% CRI
Intercept	3.08	0.53	(2.03, 4.14)
Context (Neut.)	-0.07	0.24	(-0.55, 0.41)
Info. Status (Implic.)	0.05	0.28	(-0.51, 0.60)
Coda (Cancel.)	-0.09	0.28	(-0.64, 0.46)
Ctxt × IS	0.55	0.25	(0.08, 1.05)
Ctxt × Coda	0.28	0.24	(-0.19, 0.76)
IS × Coda	0.09	0.28	(-0.46, 0.65)
Ctxt × IS × Coda	-0.33	0.24	(-0.82, 0.14)

trigger *some* or in regions of the following sentence meant to probe the online consideration or selection of an upper-bound meaning, with Bayes factors uniformly finding moderate to strong evidence against such effects.

In two cases, I do observe some evidence for a general contrast between implicature and entailment. First, at *some* itself, there is anecdotal evidence for generally slower reading in Implicature conditions. While this is consistent with the idea that *some* may invoke particular processes which *only some* does not (related to social reasoning or to meaning uncertainty), the low quality of the evidence makes me hesitate to interpret it. More convincingly, in the S3 region which directly obligates cancellation of upper-bound meaning, we see strong evidence for slower reading when the upper-bound meaning was entailed through the contribution of *only*. This is evidence for a slowdown particular to the contradiction of entailments, something that any theory of incremental comprehension would expect to observe. If there is any cost incurred in the case of cancellation of an implicated upper-bound meaning (evidence suggests no), it must be smaller than this contradiction cost.

The lack of any effect of context is somewhat discouraging, and opens up the possibility that the manipulation of context was simply not an effective way to control participants' online interpretations in this experiment. We have one tentative piece of evidence against this, however, from the comprehension question accuracy data, where we observe context-dependence in particular for the successful offline interpretation of cancelled implicatures: participants were more likely to arrive at the post-cancellation, lower-bound reading of *some* successfully when the context was Neutral, and thus less supportive of the implicature. So, context had some effect on at least participants' offline interpretations of the stimuli, as it did in the norming experiment. What we lack is evidence of context on online interpretations, e.g. the missing facilitation effect on complement reference in S3. Its continued absence here in the Maze suggests that its absence in self-paced reading was not merely due to shallow engagement, to the extent that the Maze forces participants to engage more deeply.

In the absence of any effects of the context manipulation on reading times, evidence from within-trial correlations helps to make a partial argument for some rapid anticipation effects using *some*. Two patterns of time distribution emerged in this study. In regions which affirmed an upper-bound meaning by referencing a contrast set, we observed that faster-than-normal reading times were achieved on trials where participants engaged in slower reading at *some*, a pattern which held from the earliest exposures. This resembles the pattern observed in Bergen and Grodner (2012), which was not replicated in Experiment 5. Elsewhere, in regions which anticipated cancellation of an upper-bound meaning by featuring the contrastive connective *in fact*, we observed that faster-than-normal reading times were also achieved on trials where participants engaged in slower reading at *some*, but this was a pattern which emerged over a series of exposures to these conditions.

The exact interpretation of these effects depends on a theory of how trials with longer response latencies at *some* were special. Bergen and Grodner (2012) claim that these are trials where an enriched upper-bound meaning was considered, at cost, but the widespread failure to observe context-dependent costs in this region, including in the current experiment, makes this less convincing. An alternative theory might take these to be trials where *some* was processed more carefully, and participants had a more detailed representation as a result. This may be intuitively attractive, but work in sentence processing has largely found that reading times do not predict higher accuracy in offline measures (Weiss et al., 2018), suggesting more that when they occur, delays are necessary to achieve typical comprehension performance, rather than an option to achieve atypically precise comprehension performance. Indeed a post-hoc investigation of the relationship between latencies at *some of* and comprehension question performance in this experiment revealed no effect on accuracy for questions which targeted the interpretation of *some*, or accuracy in general.

I suggest instead that the most important quality of trials with slower latencies in the *some* region was simply that they gave comprehenders more time to develop expectations conditioned on that material. Research on expectation in comprehension generally finds that expectation strength grows over time—as a very concrete example, Wlotko and Federmeier (2015) observe that contextual facilitation effects in ERPs are much smaller when presentation rates are increased. Although intentional delays were presumably associated with processing difficulty, and do not improve the quality of comprehenders' final interpretations, I will assume that in at least some cases, the simple fact of slower movement through the sentence can yield stronger online expectations. If this is the case, then the connections between slower latencies in S2 and faster latencies downstream reveal the nature of the expectations which participants are fostering in S2. The facilitatory relationship at *the rest* would diagnose that observing *some* fosters a general expectation for upcoming reference to a complement set, presumably mediated by the upper-bound meaning, although we cannot observe that dependency directly here. On the other hand, the acquired facilitatory relationship at *in fact* would diagnose that comprehending *some* can foster a general expectation for upcoming corrective elaboration, whether through merely the distribution of word forms, or through sophisticated discourse-structural prediction. The pattern of a narrator attributing a less specific belief to a protagonist, and then pointing out what they do not know has plenty of support in narrative, and could have been drawn on here as a predictable schema. It is crucial that the first relationship seems to have been present before participants had extensive experience with these stimuli, and it is critical that we do not see this effect reversed for cancellation regions. I take this particular state of affairs as evidence that comprehenders are developing upper-bound-related expectations rapidly during online reading, without inhibiting lower-bound interpretations.

3.6 General discussion

Results from self-paced reading and Maze studies examining the processing of the upper-bound meaning of *some* have demonstrated that contexts supportive of an upper-bound meaning are associated with neither (a) difficulty at *some* itself, nor (b) difficulty at later assertions incompatible with upper-bound meaning. These findings are in line with the predictions of a hypothesis of Rapid Consideration Without Selection as introduced in §3.1. Where previous studies have observed evidence that comprehenders are considering upper-bound meaning, I find no evidence for costly generation or an online selection process entailing reanalysis.

But a wrinkle remains: both studies also failed to observe the facilitation effects expected if a upper-bound meaning had been generated and was under consideration, driving expectations. This opens up competing explanations for the absence of context effects here which are unrelated to the critical research question.

The first potential deflationary explanation hinges on the nature of the stimuli in these experiments. In an effort to surmount a potential confound present in simpler designs, I substantially complicated the nature of the critical upper-bound meaning, as reviewed in §3.3.2. As a result, inferences about the true relationship between the quantifier's restrictor and nuclear scope set depend on reasoning about scalar alternatives to *some* and to factive embedding predicates, along with assumptions about epistemic parallelism between protagonists and narrators. We easily might imagine that this complication changed the way comprehenders engaged with the meaning of *some*. Evidence from a judgment study suggests that such changes could not have been extreme: I observe that in offline judgments comprehenders are just as sensitive to differences in protagonist knowledgeability as they are to differences in speaker knowledgeability in the simpler cases (§3.3.4). Indeed, this sensitivity seems to fuel differential difficulty dealing with cancellation in offline responses in Experiment 6 as well (§3.5.2.5). These observations can reassure us that context was still taken into account in the comprehension of *some*. Nevertheless, it is *a priori* possible that, in light of the increased complexity of the contextual calculus, context was less frequently exploited to direct resources in online processing here than it was in simpler designs. If this is the case, we cannot interpret the present null findings as directly informative for the critical questions of this chapter.

However, even if context did not have the expected effect here, within-trial correlations from Experiment 6 suggests that comprehenders were engaged in online consideration of upper-bound meaning without online selection of upper-bound meaning. Under the assumption that downstream differences correlated with reading times at *some* can be used to diagnose the expectations launched at *some*, the correlated facilitation of complement reference provides alternative evidence that comprehenders were engaging in some amount of expectation for the upper-bound meaning of *some*. If this expectation was due to online selection of a upper-bound reading, we would expect similar patterns of correlated inhibition for content inconsistent with the upper-bound meaning, but we see no such effect. In fact, I observe that some aspects of this meaning (the connective *in fact*) begin to be anticipated as well.

I thus take the overall picture to provide additional evidence in favor of a hypothesis like Rapid Consideration Without Selection. To conclude this chapter, I will add two further comments on the data discussed above, before returning to the larger picture.

3.6.1 The comprehension of "in fact"

One peculiarity of this experimental design is the relationship between cancellation continuations and the discourse marker *in fact* in S3. The consistency of this relationship might have had unpredicted consequences for our ability to measure costs associated with the cancellation material itself in S3.

In particular, if comprehenders took *in fact* to signal some contrast between context and the assertion to come, a model with online selection could predict that cancellation costs in the Knowledgeable condition were masked by a different sort of difficulty in the Neutral condition.[17] The logic is that in Neutral conditions, comprehenders would quickly select a lower-bound interpretation, and when they encountered *in fact* and its subsequent *all*-assertion, they would struggle to make *in fact* coherent, as its prejacent contributed information which was consistent with their current interpretation. Such an effect would be reasonable in at least one way, as comprehenders do show rapid slowdowns in reading tasks when the content of a clause is less obviously coherent with a discourse connective (see Cozijn, 2000 and much discussion in Chapter 4).

On deliberation, I think such a cost in this case is unlikely, however. There is little evidence that *in fact* is prototypically used to mark contrast. In the third version of the Penn Discourse Treebank, for instance, annotators associated *in fact* with the introduction of detailed elaborations on the current topic—e.g. (22) and 73 other cases associated with an Expansion relation—-much more regularly than contrast with a previous statement—e.g. 23 and only 10 other cases associated with a Comparison relation. A pattern common across many examples is the use of *in fact* after a weak assertion to introduce an assertion with a compatible and indeed stronger meaning, as in (24). Indeed, this last case is rather directly comparable to the case where comprehenders settle on a lower-bound meaning for *some* in S2 before encountering an *all* assertion in S3.

(22) [Even though the market started to slide on Friday,] no special bulletins or emergency meetings of the investors' clubs are planned. In fact, some of the association's members… welcomed the drop in prices.

(23) When UAL Corp. stock finally opened on the New York Stock Exchange at 11:08 a.m., the price was listed at $324.75 a share, up about $45 from Friday; in fact, its true price was $224.75, down $55.

(24) And while the job is half done, Brooks is still bitter. In fact, there's only one person involved who's happy, and that's Floyd String.

Indeed, I have already pointed out that in Experiment 6, comprehenders seemed to develop an expectation for *in fact* given *some* over the course of the experiment, potentially conditioned on their recognition of a weak meaning that might be made more precise in later discourse. On the whole, it does not seem likely that comprehenders would struggle

[17] I thank Richard Breheny for this observation.

to access an elaboration meaning here, and so the potential confound is not so worrying. Future work should, nevertheless, verify that the absence of apparent reanalysis extends to discourses which do not feature *in fact*.

3.6.2 Task effects

Unlike the experiments reported in Chapter 2, in this pair of experiments I do not observe a pressure in the Maze for earlier selection of a meaning for uncertain content. This yields some insight into the nature of the task effects discussed there. I argued that the Maze is motivating earlier decisions about polyemy and distributivity because these factors yield predictive value that supports accurate Maze task performance. If that's the case, I might assume here that decisions about the interpretation of *some* yield less predictive value. This seems reasonable, as the exact nature of the relationship between two sets may be less informative about context than the particular lexical meaning of a noun—which has implications for the setting and the entities within it—and the way in which a plurality of agents were related to discrete events picked out by the verb—which has implications for the total number of objects which must be represented in the discourse model. Nevertheless, this should be tested empirically, if it is to be a critical conclusion here.

Alternatively, Maze pressures may have been present here, but not sufficient to overcome biases for later selection of this kind of meaning. A stronger bias to postpone selection here would also make sense. Where reanalysis of a polyseme or a distributivity ambiguity involves a switch within the possible meaning of a single lexical representation, or what I have treated as a difference in the postulation of an implicit operator, revising from a selected upper-bounded meaning to a lower-bounded meaning here requires revisiting the entire pragmatic lens through which the sentence was interpreted. To the extent that this would be a more undesirable cost, the functional approach to decision timing could expect that the gain in utility would not be worth the risk.

I do observe at least three differences between the two tasks here, related to other components of processing behavior. First, recall that in Experiment 5, general costs were observed in conditions with the string *that only some*. In the subsequent discussion (§3.4.3) I proposed that this might be the result of comprehenders' uncertainty over the order of *that* and *only*. Experiment 6 revealed that no such costs emerge in the Maze. This seems compatible with an explanation tied to uncertainty in order; in the Maze task, comprehenders are presumably encouraged to engage in much more exact encoding of the order of adjacent words. The elimination of this secondary cost allows us to observe a particular cost for the contradiction of entailments in the Maze. In self-paced reading, we don't see evidence for a contradiction cost, either because it is obscured by the global difficulty in this condition, or indeed because adopting the non-veridical order would circumvent the contradiction altogether.

Next, participants in the Maze exhibited a novel context-dependence in their successful integration of S3. That comprehenders particularly failed when S3 required the cancellation of what should have been a relatively highly-expected upper-bound meaning is an indication that the Maze specifically taxed comprehenders ability to resolve conflicts between multiple sources of information. I'll note that we might have actually expected to see a reduction in comprehenders' willingness to provide upper-bound readings on the whole here, in the comprehension questions, if the apparent task demands of the Maze resembled the kind of cognitive load examined in De Neys and Schaeken (2007) and subsequent studies. There's no such reduction on display here, but that isn't particularly surprising: in this study, interpretation of *some* is fully determined by S3. For instance, Affirmation conditions entail the presence of a non-empty complement regardless of whether comprehenders settle on a lower-bound or upper-bound meaning for *some* in S2. As a result, comprehenders need not select an upper-bound meaning of *some* in order to answer comprehension questions accurately in those conditions. An investigation of Maze reading with the simpler stimuli and informative questions of De Neys and Schaeken (2007) would be an interesting follow-up.

Finally, in post-hoc within-trial latency correlations between critical *some* and S3-initial regions inspired by Bergen and Grodner (2012), there is a clear difference between the two studies. In Experiment 5, latencies at *some* predicted more difficulty in S3 than latencies at *only some* did (in the Entailment conditions used to set the baseline). One potential interpretation of this finding is that latencies on *some* when its interpretation is not fixed in context were more likely to reflect generic difficulty with the critical quantificational meaning, as compared to the control conditions where slower processing was just a matter of typical random variation in latencies. In Experiment 6, we observe much the opposite: latencies at *some* predicted ease in S3, according to two different patterns (generic ease with Affirmation continuations, and ease with Cancellation continuations acquired over the course of the experiment). In §3.5.3 I advanced an interpretation of these effects as expectations unfolding in time, benefiting from random variation in latencies at *some*. Why the difference between the two studies, then? Tentatively, the correlations observed in Experiment 6 may not show up in Experiment 5 because self-paced reading participants were less likely to engage in predictions based on hypotheses about the meaning of partial input. That is, although we do not observe an increase in online selection behavior in the Maze comparable to Experiments 2 and 4, we might be observing an increase in other optional processes that are particularly useful due to the utility of interpreted contexts in the Maze.

3.6.3 Conclusions

In this chapter, I presented a case of indeterminate meaning in online comprehension which is entirely unlike the paradigmatic cases introduced in Chapter 2. A review of the literature on the interpretation of *some* finds evidence that comprehenders generate and consider both ("logical") lower-bound and ("pragmatic") upper-bound meanings rapidly online. Nevertheless, we see no evidence for subordinate selection or reanalysis costs during early processing in the literature to date: it seems that the consideration being probed here does not result in the selection of a single interpretation. I advance a hypothesis informed by these results, Rapid Consideration Without Selection, and demonstrate one possible model of interactive expectation-based processing that could capture the particular ("No Worries If Not") facilitation effects attributed to pre-selection consideration. Across self-paced reading and Maze studies, I revisit and extend partial evidence from Bergen and Grodner (2012) that online consideration of upper-bound meanings does not engender reanalysis costs in reading. Although facilitation effects attributed to consideration were much weaker here than in previous work, emerging only in a post-hoc correlational analysis of the Maze data, I take the overall picture as consistent with Rapid Consideration Without Selection. I also observe that the preferences for postponed selection here overwhelm whatever task-specific pressures there are for early selection in the Maze.

If Rapid Consideration Without Selection is the appropriate model for the interpretation of *some*, it exists as a rather unique phenomenon in our understanding of comprehension decisions. Debate about the nature of comprehension decisions has often hinged on whether purported reanalysis effects should be explained as the costs associated with revising a single analysis of the input, or the costs associated with encountering evidence inconsistent with the most probable of multiple possible analyses of the input. Here, this picture is disturbed by evidence for a second phenomenon that needs modeling, “No Worries If Not” facilitation associated with certain analyses of the input, in the absence of corresponding reanalysis costs. Such effects lend themselves to a certain parallel interactive view of expectation-driving comprehension, but those models contrast with the expectation-driven models that have been used to account for reanalysis. If expectations need not lead to costs when they are not met, some of the intuitive argument for a general expectation-based model of comprehension is undermined. Either expectations must somehow differ in consequence between those which do and do not trigger reanalysis effects, or reanalysis effects depend on the presence of a single deterministic analysis. I will revisit this topic in Chapter 5.

Chapter 4

Considering and cancelling causal inferences

Frequently in natural language, a pair of two sentences $\langle S_1, S_2 \rangle$, unmarked by any particular discourse marker, prompt the inference that the eventuality described in S_1 caused the eventuality described in S_2. See for instance, a few examples from spontaneous American English narratives given in (25). I will call this a Result inference.

(25) a. The guy comes over and smiles at me. I blew up.

b. The girl said, "I'm sorry, I don't cook the food, it's precooked." He picked up the meal and threw it on the floor.

c. [Right Guard] was positioned at that time for men. It was not going anywhere. (Terkel, 1974)

To give a more quantitative sense of the frequency of Result inferences: in a subset of the Penn Discourse TreeBank 2.0 (Prasad et al., 2008), a corpus of newswire from the Wall Street Journal, Asr and Demberg (2012) count 2,240 cases where S_2 describes an event that closely follows S_1, without any explicit discourse marker.[1] Of these, 1,704, 76%, are cases where annotators agreed that the text invites an implicit Result inference.

It is also very common for an inference to be generated with the opposite temporal orientation, that the eventuality described in S_2 caused the eventuality described in

[1] That is, 2,240 cases that were implicit, and received the label TEMPORAL.Asynchronous.precedence or CONTINGENCY.Cause.result. No other PDTB2 tag marks an inference about forward temporal movement. This is about 14% of the 16,327 annotated implicit pairs used by Asr and Demberg (2012). Note also, that *closely* here is relative to the temporal grain of the narrative.

S_1 (26).

(26) a. The doctor advised me to quit work. My heart got bad to where I couldn't get enough oxygen.

b. I admired the men ... that were stonemasons. They knew their trade.

c. Until recently I'd cry in the morning. I didn't want to get up.

(Terkel, 1974)

These, which I'll call Explanation inferences, are even more frequent than Results in the PDTB2. Of the 2,628 cases where S_2 describes an event that closely precedes S_1 without a marker,[2] 2,467, almost 94%, were annotated as Explanations. Together, Results and Explanations make up more than 25% of all implicitly related sentence pairs in the PDTB2.

Given that these inferences are common, and apparently present in comprehenders final interpretations of a discourse, I am naturally interested here in the timecourse of their comprehension. Inspired by the dissociation between consideration and selection observed in the previous chapter, we can ask about these processes separately here as well. On the one hand, (a) are these inferences (generated and) considered during online processing? When does that consideration begin, and can it facilitate processing of inference-compatible material like consideration of *some*? And on the other hand, (b) does selection of an interpretation featuring the inference happen during online processing? If so, when does it occur, is it postponed until offline processing like the interpretation of *some*?

The answers I will give here are roughly: (a) Yes, causal inferences rapidly influence processing of their tail in a manner consistent with online consideration; and (b) No, selection of a causal inference is not apparent in running discourse, even at a delay of a few sentences. These answers come from a growing literature on these phenomena, to which I add a series of three reading studies which look for and fail to find reanalysis costs for Explanation inferences, using the same methodologies used elsewhere in this dissertation. In Experiment 7, I examine cases where causal meaning might be anticipated, finding that readers use world knowledge and the possibility of a causal inference to anticipate likely content in the tail, but do not suffer costly reanalysis afterwards when they encounter evidence against the inference. Experiments 8 and 9 follow up by manipulating the contextual possibility of a causal interpretation for a world-knowledge-plausible

[2] That is, TEMPORAL.Asynchronous.subsequence and CONTINGENCY.Cause.reason. The latter corresponds to Explanation in the present terms. This is about 15% of the corpus.

Explanation, and probing for reanalysis within a larger time window. Results continue to reveal the absence of reanalysis effects, and also demonstrate that online expectations conditioned on anticipated causal inferences are not sensitive to context-specific information, a crucial limit on the way that these expectations arise. On the whole, I will argue that the status of these inferences in online comprehension adheres to the hypothesis of Rapid Consideration Without Selection advanced in the previous chapter.

The chapter is organized as follows: in the next section (§4.1), I will review the two most prominent approaches to these inferences in formal pragmatics, to establish some basic understanding of the ultimate goal of the comprehender. Next, in section 4.2, I review the existing literature on the timecourse of these inferences, which has frequently found evidence for online consideration but generally avoided questions of selection. I then present Experiments 7, 8, and 9, before wrapping up with a general discussion in section 4.6, detailing my conclusions and taking care of some odds and ends relating to domain-general causal attribution and the semantics of *because*.

4.1 Causal inferences in discourse

In understanding a pair of adjacent sentences, unmarked by any particular cue to their coherence, comprehenders seem obligated to see some connection between what each sentence describes. This is the broader phenomenon described in influential theories and frameworks for implicit discourse structure proposed in, e.g. Hobbs (1979), Mann and Thompson (1988), Kehler (2002), and Asher and Lascarides (2003). In these approaches, it's this mandatory connection, **discourse coherence**, that can give rise to inferences like the causal ones highlighted here.

It's apparent that the nature of the **coherence relation** most naturally arrived at, and thus the nature of the inference, is strongly controlled by the semantic content of the two segments being connected, call them the **head** and the **tail**. We can see this by comparing minimally different segment pairs with similar structure. Example (27) gives two pairs of transitive sentences in the English simple past tense: in each case, the subject of the tail is a pronoun referring back to indefinite object of the head. Nevertheless, the discourses yield opposite causal inferences. In (27a), the closing of the factories is most naturally understood as a result of the activist's purchase, while in (27b), the threatening of the CEO seems to describe the reason why the manager was fired. I will use the symbol

➻ below discourses to mark natural inferences.

(27) **Implicit intersentential causal inferences**

a. The activist bought$_1$ a cosmetics company. It closed$_2$ several factories.

➻ Result: e_2 because e_1

b. The company fired$_1$ a manager. He threatened$_2$ the CEO.

➻ Explanation: e_1 because e_2

Theoretical approaches to the nature of discourse coherence in semantics and pragmatics fall into two basic camps. On one, which owes principally to Hobbs (1979), coherence is simply a process of inferencing over possible relations between adjacent segments. Types of coherence relations and the inferences they lead to are not taken to be linguistic categories, but general patterns that arise from the nature of communication and constraints on relevance. Kehler's (2002) notable elaboration and formalization of this proposal demonstrates how joint inferencing over possible coherence relations and ambiguous linguistic content like VP ellipsis, pronominal reference, and temporal interpretation can derive a number of desirable patterns. These insights have been extended by continued work in this domain in the decades since (Kehler et al., 2007; Kehler & Rohde, 2013, 2017, 2019).

The other cluster of work sees coherence as a particularized grammar of relations used to generate a binary-branching, recursive hierarchical structure, one which is directly constrained by and directly constrains other parts of linguistic meaning. Key early work laying the foundation for such accounts comes from the proposals of Reichman (1978), Polanyi (1985, 1988), and Grosz and Sidner (1986), who discuss how hierarchical models of discourse that apply coherence relations recursively can derive apparent constraints on pronominal reference and further coherence. The core generalization is today called the Right Frontier Constraint (RFC): as a new segment enters into the discourse, only material in "accessible" nodes in the hierarchical representation of discourse can be referred to with discourse anaphora, or can serve as the head in a coherence relation with the new segment.

Asher and Lascarides (2003), who coined the RFC name for the phenomenon, take this as evidence that coherence is principally language-internal. While they agree with others that the library of coherence relations available to comprehenders is in some sense derived from general principles, in the account they propose, Segmented Discourse

Representation Theory (SDRT), the relations are linguistic objects with parameters that determine how they participate in constructing the hierarchical discourse. In particular, in SDRT, in order to generate the distinctions in "accessibility" for the RFC, discourse relations must be specified for whether they (i) subordinate their tail, thus preserving the accessibility of their head, or (ii) coordinate, thus rendering their head inaccessible. Moreover, in SDRT, the relations are in turn part of the semantic content of lexical items, part of the entailments of discourse markers like *however* or *so*. The entire procedure for coherence, then, lives in an extended system of (dynamic) compositional formal semantics. Note that these proposals, like Hobbs (1979) and Kehler (2002), have to appeal to some kind of inferencing for the resolution of coherence in cases like (27) where the linguistic content does not fully constrain possible parses, though SDRT provides a particular logic for that inferencing.

These two main approaches to coherence, which I'll call SHALLOW INFERENCING and HIERARCHICAL STRUCTURE-BUILDING, do not strictly require that the head and tail of a coherence relation must be independent sentences. In fact, both have frequently extended to cover relationships between roughly clausal segments within a single sentence. In particular, these are usually segments whose grammatically-specified relation is either minimally constraining (e.g. coordination structures; 28a) or equivalent to the inference that would arise from a coherence relation (e.g. *because*, *so*, *then*; 28b).

(28) **Run-of-the-mill intrasentential causal inferences**

a. The activist bought$_1$ a cosmetics company and it closed$_2$ several factories.

➻ Result: e_2 because e_1

b. The company fired$_1$ a manager because he threatened$_2$ the CEO.

➻ Explanation: e_1 because e_2

Such cases fit in naturally with the SHALLOW INFERENCING approach: either the tail already specifies the way in which it is coherent with the head, or it remains vague and the comprehender will have to generate a connection just as they do across sentences. HIERARCHICAL STRUCTURE-BUILDING theories also need not say much here: these segments do not relate to each other in any way besides meanings that already must be captured in the system of discourse coherence, so it is plausible that they, like sentences, are separate discourse segments.

But there are other instances where discourse coherence-like inferences seem to arise intra-sententially beyond these basic cases. Eventualities described in restrictive modifiers of nouns, like restrictive relative clauses (RRCs), seem to give rise to such inferences as well (Cohen & Kehler, 2021). We can construct examples parallel to (27) where the tail is not a following sentence but an RRC modifying an indefinite object in the head, and at least the Explanation inference is still apparent. Because such cases are among the phenomena examined in the experiments to follow, I will linger on how we might derive the apparent inferences here.

(29) **Intrasentential causal inferences with RRCs**

a. The activist bought$_1$ a cosmetics company that closed$_2$ several factories.

➻ ??Result: e_2 because e_1

b. The company fired$_1$ a manager that threatened$_2$ the CEO.

➻ Explanation: e_1 because e_2

These inferences are a robust feature of running discourse much like prototypical cases of Result and Explanation. In a pilot investigation of restrictive relative clauses in the Wall Street Journal subset of the Penn Treebank (Marcus et al., 1999), I find that roughly 20% yield Explanation-like inferences (30).[3]

(30) **Explanations in naturally-occuring RRCs**

a. Two years ago, the Rev. Jeremy Hummerstone, vicar of Great Torrington, Devon, got so fed up$_1$ with ringers who didn't attend$_2$ service he sacked the entire band; the ringers promptly set up a picket line in protest.

➻ Explanation: e_1 because e_2

b. Stockholders who took the hint and sold$_2$ shares escaped$_1$ the October debacle.

➻ Explanation: e_1 because e_2

RRCs are not the only way of instantiating this kind of inference between the way an individual is described and an event in which they are a participant: participial modifiers

[3] No clear cases of a Result-like inference were found, where the event of the RRC occurred because of the matrix event. (Note that this isn't a matter of temporal interpretation alone—there were still several cases where RRCs described events which temporally preceded the matrix event.) This may be evidence for a structural constraint on the availability of sub-sentential discourse inferences, but this topic is not relevant to the discussion of this chapter.

(31a) and deverbal nouns (31b) can give rise to the same inferences (Webber, 1991 *apud* Rohde et al., 2017; Hobbs, 1990; Cohen and Kehler, 2021).

(31) **Other intrasentential causal inferences within nominal descriptions**

a. The manager Sheila lectured$_1$ the texting$_2$ intern Kevin.

➻ Explanation: e_1 because e_2 (Rohde et al., 2017)

b. A car hit$_1$ a jogger$_2$ in Palo Alto last night.

➻ Explanation: e_1 because e_2 (Hobbs, 1990)

A SHALLOW INFERENCING approach to coherence is capable of handling these cases similarly to others, with one small adjustment. Unlike the other examples reviewed so far, inferences arising from eventualities within nominal descriptions cannot be the result of mandatory coherence. Very frequently, constructions like RRCs are used simply to establish reference; there is no need for them to be integrated in any other way in the discourse (32) (Hoek et al., 2021a).

(32) **RRCs without inference**

a. Last year, two cosmetics companies folded: one went bankrupt, the other merged with a competitor. An activist bought$_1$ the cosmetics company that went$_2$ bankrupt.

➻ ~~Result: e_2 because e_1~~

b. For a long time, Rebecca was happy at her job. But then, the company fired$_1$ the manager that supervised$_2$ her, because he threatened the CEO.

➻ ~~Explanation: e_1 because e_2~~

This seems to be one reason these approaches characterize such inferences as general "pragmatic enrichment" (Cohen & Kehler, 2021); while the same kind of inferences are generated, the process must be driven by a more optional desire to connect ideas.

In contrast, HIERARCHICAL STRUCTURE-BUILDING approaches, here exemplified by SDRT, are constrained in their ability to account for these kinds of inferences. Restrictive modifiers like RRCs serve a critical semantic function determining the individuals that their containing nominal expression picks out. In SDRT, there is no way for external material to feed into nominal semantics—the latter is strictly prior to the representations that participate in coherence. This means RRCs, so long as they have their typical semantic function, cannot be visible for the linguistic system of discourse coherence. In the face of cases like (29), this approach would either have to give up the idea of strict layering between semantic and discourse composition, or appeal to a non-linguistic source for the inference.[4]

To summarize, theories of causal inferences generally fold them into the process of coherence-building in discourse. This process is often modeled as selection from a library of possible schemata for the structure of a discourse; when comprehenders relate input to certain schema, like Result or Explanation, they effectively enrich the meaning of the input. Some theories treat coherence resolution as generalized pragmatic inference over potential communicative choices, while others cast coherence as a process of hierarchical structure-building, to capture apparent relationships between coherence inferences and patterns of coreference and potential continuations. The prototypical data explained by these theories is typically the interpretation of adjacent sentences or coordinated clauses, although similar inferences seem to arise between sub-clausal elements of a discourse; there may be a stronger case that these sub-clausal inferences belong to the domain of astructural pragmatic inference.

4.2 The processing of causal relations

I will take as given that causal inferences between the events described in two adjacent units of discourse are part of the output of the interpretive process. I will also follow the formal pragmatics literature in assuming that the decision in this case is more complicated than a simple binary decision to enrich: other readings for a pair of sentences can in principle be generated by the grammar, and so comprehenders may be entertaining alternatives with various enrichments before apparently settling on something causal. In the context of this dissertation, I am of course interested, then, in when during online comprehension the processor begins considering these various alternatives, and when, if ever, they make a selection.

There has been a sizable literature which can shed some light on some aspects of this timecourse, principally reporting data from self-paced reading and eyetracking while reading. In this section, I will step through first the evidence for the timecourse of Result

[4] Or, perhaps one might try to deny the premise that these are really RRCs, resolving that, despite their surface characteristics, they ultimately have the semantic contribution of non-restrictive, appositive relative clauses, which have been argued to contribute separate discourse segments and participate in discourse relations (Burton-Roberts, 1999; Koev, 2013; Jasinskaja, 2016). This does not seem very plausible.

interpretations, and second, the timecourse of Explanation interpretations. In both cases, we will observe evidence that causal relations can drive expectations for associated lexical material and reference patterns, evidence which is quite similar to the evidence for rapid generation of scalar implicatures. But as regards selection, there has been no good evidence for difficult selection or reanalysis at any timescale. This will motivate a trio of experiments aiming to probe for reanalysis costs.

Before I move on, in the interest of appropriately relating this literature to other findings, I would like to highlight the ways in which a causal inference is functionally similar and dissimilar from an enriched meaning for a weak scalar item, for the comprehender. On the one hand both the quantifier *some* and the relevant implicit relations discussed here express a relation between two arguments, where a given interpretation may lead to expectations for the latter argument(s) given context, or conversely, where the choice of interpretation may benefit from observing the latter argument(s). The main difference here is just that *some* relates sets of individuals, while discourse relations relate something like propositions. But on the other hand, in this case there is no contentful unit at which the presence of the ambiguity comes into focus. The uncertain meaning of a statement containing a weak scalar item enters a comprehenders' awareness when they encounter the weak scalar item itself; the potential for a causal inference between two discourse segments, I suppose, enters a comprehenders' awareness when they encounter the absence of any explicit discourse connective on a potential tail. As mentioned above, another key difference is that a causal enrichment is one of many other possible interpretations, rather than a binary case of deciding between enrichment or non-enrichment, although below I will often discuss it like this to simplify things.

With all of this in mind, some avenues of comparison are clear. For instance, patterns of comprehension behavior on the potential tail of a relation are an informative place to test for incremental expectations, comparable to the critical arguments of *some* investigated in studies like Hunt et al. (2013). However, note that many studies of *some* probed for insight into the immediate consequences of potential pragmatic enrichment by investigating behavior directly on *some*; there is of course no comparable position to examine here. We will see, at least, that reading behavior on the potential tail of a causal inference demonstrates some of the same expectation effects as observed in Chapter 3.

4.2.1 Expecting Results

In a small literature on the processing of narrative, it has been often found that Results are expected, at least sometimes (Sanders, 2005; Mulder, 2008). As an exemplary case, consider an EEG experiment reported in Kuperberg et al. (2011). Subjects read multi-sentence narratives in three conditions, exemplified in (33). ERPs were measured relative to the onset of a critical word in the final sentence, which was equally probable in each condition as measured by lexical co-occurrence measures. In "Highly related" and "Intermediately related" conditions, the target word was additionally predictable due to world-knowledge given the first sentence. Critically, in only "Highly related" conditions, the second sentence of the discourse was designed to make a Result inference between S2 and S3 predictable. Because the target word was always Result-related, the authors predict that if comprehenders exploit that predictability and generate expectations about particular words on the basis of it, they will have a higher expectation for the target word in "Highly related" vs. "Intermediately related".

(33) a. **Highly related**: Jill had very fair skin. She forgot to put sunscreen on.

b. **Intermediately related**: Jill had very fair skin. She usually remembered to wear sunscreen.

c. **Unrelated**: Jill's skin always tanned well. She always put on sunscreen.

Target: She had **sunburn** on Monday.

Indeed, the authors report differences in N400 effects among the three conditions, such that the N400 on the target word is lowest in "Highly related" conditions. This would suggest that in at least these particular contexts, comprehenders used a prediction of a Result to anticipate certain lexical content in the upcoming discourse unit.

This predictive facilitation is consistent with a repeatedly-attested facilitation effect in sentence-by-sentence self-paced-reading. In experiments that measure reading times on a sentence in various contexts, reading times decrease as a function of the degree to which they could be expected as a result of the context (Keenan et al., 1984; Myers et al., 1987; Wolfe et al., 2005). I.e., reading times were fastest in contexts where the target sentence was a very predictable result (34a) and slowest where the target sentence was not a particularly predictable result (though not necessarily incoherent) (34d).

(34) A stimulus set from Keenan et al. (1984):

a. Joey's big brother punched him again and again.

b. Racing down the hill, Joey fell off his bike.

c. Joey's crazy mother became furiously angry with him.

d. Joey went to a neighbor's house to play.

Target: The next day his body was covered with bruises.

These results are usually taken to indicate support for the online expectation that the next sentence is going to be a result. As the story goes, comprehenders anticipate the particular content of a likely result of S1, and S2 is read faster when it is in line with the anticipated content.

But other findings suggest that results are not particularly privileged in comprehenders' expectations. A study reported in Murray (1997) compared reading times on S2 where S1 made S2 a plausible result (35a) to Contrast conditions where S2 provided information that would be unexpected given S1 (35b), and Elaboration cases where S2 merely added information about the event in S1 (35c). Unsurprisingly, Result conditions were read faster than Contrast conditions, but there was no reliable difference between Result and Elaboration differences. It would seem that Result tails are not more expected than other kinds of predictable continuations.

(35) A stimulus set from Murray (1997):

a. **Result**: Manny needed to publicize the garage sale.

b. **Contrast**: Manny forgot to publicize the garage sale in the paper.

c. **Elaboration**: Manny informed his staff about the garage sale.

Target: He arranged for flyers to be made.

Likewise, Mulder (2008, Ch. 5) reports that in the reading of more naturalistic narratives in Dutch, Result tails (36) seem to be no more expected than elaborations. That study further provides a case where another kind of information was more expected than a result: sentences describing a solution to a problem given in context (37)[5] were read more quickly than either of the other conditions. Mulder takes this as evidence that these reading time differences are driven more by more flexible mechanisms of contextual predictability, rather than a rigid default for Results.

[5] See Sanders and Noordman (2000) for more description of these kinds of examples.

(36) **Result:** The City Council of Maasbracht announced yesterday that the sewer system under the Main Street will be renewed. The work will start at the end of this year. Local residents of the Main Street will have to reckon with worse accessibility of their houses. Because of the sewer work the street will have to be partly broken open. **As of December first, the Main Street will be closed for motor vehicles**. Bicyclists and Pedestrians will still be able to use the road.

(37) **Solution**: The Main Street in Maasbracht has been a popular traffic route for years. Not surprisingly, local residents have filed a lot of complaints about discomfort brought about by stench and noise. Also, heavy traffic in the Main Street has led to numerous dangerous situations. According to the City council this situation has gone too far. **As of December first, the Main Street will be closed for motor vehicles**. The residents have reacted positively on this news.

Comprehenders seem to entertain the notion that a head and a tail might adhere to a Result schema during online comprehension. Specifically, the evidence discussed here suggests that this consideration must be early and substantial enough to drive some expectations about the content of the tail.

Occasionally, researchers like Sanders (2005) have advanced the proposal that this rapid consideration could reflect a default for causal attribution in narrative processing. This offers a possible solution to what Sanders calls the "paradox of causal complexity": comprehension of Result-compatible narratives is faster than less coherent narratives, but it also fuels better memory representations (Black & Bern, 1981; Trabasso & van den Broek, 1985; Sanders & Noordman, 2000), contradicting a general expectation that more enriched representations require costly processing. But the failure to find Result-specific facilitation effects in Murray (1997) and Mulder (2008) suggests that the relevant reading time effects here should be attributed to general coherence rather than causal attribution alone, undermining the main evidence for any default causal inferencing.

Also notice that the evidence offered from the existing studies is not strong enough to argue for a default meaning: we would expect to perhaps see evidence that selection of the non-default non-causal meaning in supporting contexts was associated with difficulty and effort, and that in neutral contexts, late disambiguation to a non-causal meaning will require costly reanalysis. To my knowledge, no such effects have been demonstrated in the literature with Results.

The picture, so far, at least partially resembles the evidence for rapid consideration of enriched meanings for *some* reviewed in Chapter 3. The possibility of a certain contextually-likely enrichment can be used to expect certain meaning and material in the portions of the text which will combine with the enrichable meaning.

4.2.2 Expecting Explanations

A larger literature on the processing of texts adhering to an Explanation schema corroborates evidence for rapid anticipation of causal relations in a text, and offers more opportunities to test whether any causal enrichments are selected during online comprehension. The relative richness of the Explanation literature is perhaps owed to a long line of work beginning with Garvey and Caramazza (1974) demonstrating how Explanations can be influenced by Implicit Causality (IC) verbs in a head. Because of the hypothesized connections between this initial verb and the prediction of an Explanation tail, expectations can be manipulated more straightforwardly than the Result cases reviewed above, and researchers have been able to advance and test more fine-grained processing predictions. The overall picture is one where Explanation tails are anticipated early, which leads to faster and easier reading of the tail in certain cases, but does not engender the selection of an enriched meaning.

I'll first review studies of overall reading time. Mak and Sanders (2013) report that in an eyetracking-while-reading experiment in Dutch, tails given by *when* clauses[6] were read faster following an IC verb (38). Note that this methodology allows some finer localization of the effect: In addition to faster first-pass times at the pre-verbal object, readers made fewer and shorter regressions from post-verbal regions in the IC conditions. In these conditions, the IC verb is taken to have aided integration of the second clause by cuing an expectation for a particular coherence relation.

(38) (Glosses of) a stimulus set from Mak and Sanders (2013):

a. **IC**: The protester got a fine from the policemen...

b. **No IC**: The protester spoke with the policemen...

Target: when he the rules broke during the demonstration.

[6] Incidentally, *when* clauses are a source of potential dispute among theories of discourse coherence: they are canonically understood to be restrictive relative clauses specifying the time of their heads, and so would not be expected to participate in discourse relations in an unmodified SDRT.

Evidence from similar studies reported in Cozijn (2000) helps to support the conclusion that this facilitation comes from causal reasoning in particular. Cozijn, reasoning that Explanation inferences should be more accessible in scenarios where comprehenders were highly familiar with the causal mechanism by which S2 could explain S1, constructed and normed narratives which might feature more familiar causal mechanisms and less familiar causal mechanisms (39). Across several self-paced reading and eyetracking experiments, he observed that the typical Explanation-contingent speedup was reduced when the causal mechanism was less familiar. Truth-value judgments on a statement of the causal mechanism were also delivered faster in the familiar condition, taken to diagnose activation of the mechanism for causal reasoning.[7]

(39) (Translations of) an excerpted stimulus set from Cozijn (2000):

They decided, however, not to buy the house in the city.

a. **Familiar**: It was actually very expensive for them.

b. **Unfamiliar**: They would have to deal with a tenant.

A similar study reported in Hoek et al. (2021a) compared reading times on RRCs following various IC verbs (40) using self-paced reading. The RCs were read faster if the IC verb described an event they could plausibly explain, compared to cases where the IC verb described an event which would be unexpected given the RC, but also compared to a neutral control condition. This supports the idea that IC verbs facilitate integration by driving expectations for particular, meaning-congruent lexical material.

(40) A stimulus set from Hoek et al. (2021a):

a. **Explanation**: She **praised** the guy...

b. **Unexpected**: She **fired** the guy...

c. **Neutral**: She **joked with** the guy...

Target RC: who made a lot of money for the company.

A large literature has investigated how these expected Explanation relations drive patterns in the interpretation of ambiguous pronouns; see e.g. Koornneef and Sanders (2013) for an exhaustive review. Though there has been some debate, the modern consensus has been

[7] See Singer and Halldorson (1996) and Halldorson and Singer (2002) for the use of similar priming logic to argue for the activation of the causal mechanism in the case of Result processing.

that Explanation-consistent reference patterns in the tail are predicted online in the immediate wake of the IC verb in the head. As recent eyetracking evidence, Mak and Sanders (2013) find that tail subject pronouns referring back to the head object have slower first-pass reading times after a subject-biasing IC verb. The same effect shows up in the number of regressions out of the verb region. It would seem that, after IC verbs, comprehenders develop expectations for pronominal reference that are active already from the initial stages of reading the first words of a potential tail. If these effects are indeed mediated by expectations for particular discourse relations, as in the Hobbsian tradition (Hobbs, 1979; Kehler, 2002), then they show that causal connections between a head and a tail are anticipated before the content of the tail is encountered.

The same conclusion is supported by evidence from Rohde et al. (2011) via an effect on syntactic parsing. The authors argue that the same parts of Explanation meaning that lead to biases in pronominal reference—that is, anticipated discussion of the most causally-relevant entity—could derive biases in syntactic parsing as well. In particular, possible Explanation tails provided by RRCs with temporarily ambiguous attachment would be preferentially interpreted as modifying the causally-relevant entity. As a result, while comprehenders will follow the typical English low-attachment bias in (41a), after an IC verb as in (41b) they will be more likely to choose a high attachment parse. self-paced reading reading times can be used on a disambiguating verb in the relative clause to diagnose what parse was chosen: if more high parses were adopted, reading times on a verb consistent only with low attachment will be longer on average than reading times on a verb consistent only with high attachment, because the first cases will be more likely to prompt costly reanalysis. This is indeed what Rohde et al. (2011) find, and it serves as additional evidence that expectations for Explanations are formed early enough to influence the low-level parsing of their tails.

(41) A stimulus set from Rohde et al. (2011):

a. **No IC**: John babysits the children of the musician…

b. **IC**: John detests the children of the musician…

Target: who {**low**: is | **high**: are} generally arrogant and rude.

Convergent evidence of the predictability of Explanations following IC verbs comes from an implicit learning task conducted by Rohde and Horton (2014) using a visual world paradigm. Participants were shown animations where the point of emergence of a ball from a forked tube was (implicitly) conditioned on whether a discourse excerpt presented simultaneously over headphones featured an Explanation relation or an Occasion (a.k.a. Narration) relation. They were instructed to learn the relationship between the sentences and the ball's motion. In a second, test block, participants had to anticipate where the ball would emerge based on just the head of the relation, and click on it before the tail would be played. Their gaze was tracked during the test block, and results reveal that when hearing heads with IC verbs, participants were more likely to direct their gaze predictively towards the Explanation position, beginning around 1000 ms after the verb offset.[8] If the ability to exploit these cues in this task reflects a general ability to anticipate causal connections between sentences based on the properties of their heads, again we see here evidence that this anticipation is robust and early.

In all, the literature on the processing of Explanations matches the literature on Results: when an Explanation is predictable, comprehenders expect Explanation-consistent content in the first possible tail, and exhibit facilitation when they get it. But do these expectations come from the early selection of an Explanation, or mere consideration? First, a few pieces of evidence suggest that these expectations can be quickly cancelled in the light of cues that they will not be fulfilled. For instance, Koornneef and Sanders (2013) find that the usually-reliable post-IC reading time patterns for pronouns are absent for tails that are explicitly marked as non-explanatory through discourse connectives like *and* or *but*. Similarly, Hoek et al. (2021b) show that while explanation-inconsistent connectives are usually read more slowly than explanation-consistent connectives in the next independent clause following an IC verb, when the IC clause is followed by a brief RRC that could offer an explanation, that connective preference effect is eliminated. If these effects were the consequence of a selected enriched meaning, we might expect that they would not disappear so easily.

Elsewhere, studies comparing implicit Explanations to explicit Explanations marked with *because* largely find that facilitation effects on the tail are larger when explicitly marked (Millis & Just, 1994; Cozijn, 2000). If an Explanation was selected with commitment in advance of the tail, we might expect that facilitation should be equivalent. On the contrary, the fact that we observe a gradient of facilitation here is more compatible with the idea that comprehenders are spreading expectation across multiple candidate

[8] They similarly learned to predict Narration relations from a transfer-of-possession verb.

interpretations in the implicit case, and only commit in the presence of *because*.[9] [10]

Finally, most convincingly, additional evidence suggests that there is no reanalysis effect following an IC verb if the next clause cannot provide a suitable explanation. Recall the Mak and Sanders (2013) study discussed above, where IC verbs, compared to heads without IC verbs, speed up the reading of a plausibly explanatory *when* clause. The same study examined the reading of *when* clauses that are implausible explanations. If we thought that IC verbs could lead participant to commit to an Explanation inference in advance of reading the tail, we would expect that implausible explanations following an IC verb would exhibit costly reanalysis. But Mak and Sanders find no such difference: these *when* clauses were read the same, regardless of the presence of a preceding IC verb.

In sum, it would seem that IC verbs help listeners expect an Explanation, and thus help them expect various connectives, reference patterns, event descriptions, and so on. Nevertheless, these are flexible predictions, and we don't have any evidence that they ever become firm commitments.

4.2.3 Interim summary

Across a collection of studies, mostly investigating narrative reading behavior, we have seen convergent evidence that causal relations are considered, and drive top-down expectations during the comprehension of their projected tail, in a variety of contexts. This holds for Results (e.g. Keenan et al., 1984; Kuperberg et al., 2011), and certainly for Explanations (e.g. Mak and Sanders, 2013; Hoek et al., 2021a, 2021b).

We might expect to observe a concomitant generation cost when these enriched meanings are considered, following the same logic that expected costly generation for enriched meanings of *some*. Such a cost is not immediately apparent, but it is also unclear exactly how we would observe it. While e.g. Breheny et al. (2006) hypothesized that reading times on *some* should reflect additional processing costs when context supports enrichment, there is no token which introduces the possibility for multiple meanings in this case. We must imagine such costs would materialize before the consequences of a hypothesized causal meaning are observed, i.e. before the tail, perhaps associated with material

[9] Millis and Just (1994) and Cozijn (2000) also observe a specific tail-final reading penalty in their *because* conditions. We might attribute this to the costs of generating a causal meaning. See also evidence from Cozijn (2000) and Millis et al. (1995) for stronger priming of the causal mechanism with *because*.

[10] Of course, this argument hinges on the assumption that these middling means are the result of reduced facilitation in individual trials; if response times are bimodal, this might be explained away as a higher proportion of individuals who simply do not consider the Explanation interpretation in the implicit case.

in the head which raises the likelihood of a causal relation. To my knowledge, such effects have not been observed: e.g. in Hoek et al. (2021a), self-paced reading latencies showed no particular costs for IC verbs compared to verbs chosen to set up neutral expectations. Perhaps we should not be surprised here, as reading time studies have largely failed to find generation costs at *some* (Politzer-Ahles and Fiorentino, 2013; S. Lewis, 2013; Hartshorne and Snedeker, 2014; see also Chapter 4).

The major remaining question pertains to the timing of selection of a causal inference. I have laid out some arguments above that this selection can't precede the contents of the tail, mostly based on evidence for fast, cost-free abandonment of Explanation expectations during the reading of the tail. But what is the status of the causal meanings considered during the tail as the comprehender moves on to later discourse material? Millis and Just (1994) and Cozijn (2000) suggest that causal integration proceeds during clausal wrap-up processing at the end of the tail, which accords with findings about some other pragmatic enrichment (e.g. Foraker and Murphy, 2012 on sense selection for polysemy). Both of those studies indeed observed an increase in processing costs there with *because*, but given that *because* is an unambiguous marker, we can't attribute this slowdown to the selection of causal meaning over some alternatives. Instead, we might tie this effect to the construction of causal meaning, which indeed might engender some baseline difficulty.

In the absence of selection costs during reading, critical evidence for online selection would come from observations of costly reanalysis. To my knowledge, the only study that examined the reading of material after the hypothetical tail of a causal relation is Hoek et al. (2021b), who report eyetracking data on a clause following an object IC verb and an object-modifying relative clause. When the relative clause did not provide a suitable reason for the matrix event, eye movements reflected shorter reading times on the connective *because* vs. alternative *and so*, taken to diagnose expectations for a subsequent Expectation. But when the relative clause did provide a suitable reason, this effect was somewhat reduced, in particular associated with longer reading times on *because* than in the other condition, although *and so* was still associated with slower reading. There are two potential analyses here. If *because* clauses are strictly incompatible with a second, implicit Explanation from the same head, this could be evidence for costly reanalysis away from the causal inference, cued by *because*. Alternately, *because* could have been facilitated due to the IC verb while participants were still expecting explanatory material, and this facilitation in turn could have been dampened given the causal potential of the relative clause, without concluding that a causal meaning for the relative clause was selected at the clause boundary. The critical condition to compare would have been a neutral one, where we expect less facilitation of *because*. If *because* following a plausibly explanatory relative clause is read more slowly than this control, we could be certain that this was a reanalysis effect rather than a reduction of facilitation. Without this condition, evidence for reanalysis remains elusive.

To sum up, we find ourselves in a similar place as in the literature on scalar implicature. Convergent evidence across multiple paradigms suggests that this type of pragmatic enrichment is considered during online comprehension, increasing expectations for enrichment-congruent material. This consideration seems to come about without a costly process of generation. We know less here about the selection of this enriched meaning, in the absence of truth-value judgment studies which depend on the enriched meaning,[11] and in the absence of attempts to directly probe cancellation at a delay. Having determined at least that the enrichment is considered online and not selected before the offset of the tail, what remains is to determine whether this enrichment is selected during comprehension at all. If we expect that Rapid Consideration Without Selection is a property of all pragmatic enrichments, we should expect that selection only occurs post-hoc.

In the remainder of this chapter, I will present three studies that probe for online selection by examining reanalysis costs in the Maze task and self-paced reading. They concur with the existing available evidence that while comprehenders do exploit the possibility of an Explanation inference to develop expectations for the possible tail, they do not select an Explanation-enriched meaning, even after several sentences. I begin with a Maze task experiment testing for reanalysis triggered by a *because* clause after a plausibly-explanatory relative clause.

4.3 Experiment 7: Causal inferences from relative clauses in the Maze

In order to test more explicitly for reanalysis effects associated with the late cancellation of Explanation inferences, I adapt here the classic moveable-disambiguator paradigm of Frazier and Rayner (1990), used in Experiments 1 and 2 in Chapter 2. By probing the processing of material which disambiguates against a plausible causal infer-

[11] It's hard to imagine what these would look like.

ence before and after the tail of that inference, I can observe whether comprehenders have made a decision during the processing of the tail that necessitates costly reanalysis to retract.

Following the concept of Hoek et al. (2021b), the critical disambiguator in this case will be a *because* clause. I, for now, adopt the assumption that explicit *because* clauses and implicit Explanations compete for the same interpretive slot. Consider example (42a), where an RRC offers a plausible explanation. If the same statement is followed with an explicit *because* clause (42b), it seems to rule out an implicit Explanation inference entirely. This assumption will be revisited and discussed more extensively in §4.6.

(42) a. The company fired$_1$ a manager that threatened$_2$ the CEO.

➻ Explanation: e_2 causes e_1

b. The company fired$_1$ a manager that threatened$_2$ the CEO because they were caught$_3$ stealing office supplies.

➻ Explanation: ~~e_2~~ e_3 causes e_1

In this experiment, I will also follow Hoek et al. (2021b) in using relative clauses as the hypothetical tails of the critical Explanation inferences. They are a nice place to begin our investigation, because they have a relatively salient non-causal interpretation that is regularly available in most contexts, acting solely as a restrictor for the purposes of establishing semantic reference. Unlike pairs of adjacent sentences, no theory of discourse structure expects that a relative clause must instantiate some contentful discourse relation with its matrix clause. Nevertheless, the availability of explanatory readings is by now well-attested (Rohde et al., 2011; Hoek et al., 2021a, 2021b).

Finally, note that this experiment will use the Maze paradigm. The Maze, as established in Chapter 2, can provide fairly precise reading time data, and encourages earlier commitment to at least some types of comprehension decisions. It provides what may be the best opportunity to observe early decision-making and its consequences.

These features come together in a 2×2 design, crossing the Position of the disambiguating *because* clause (Early vs. Late) and the explanatory Plausibility of the relative clause (Plausible vs. Implausible). This design is demonstrated for a sample item in Table 4.1. More discussion of the materials and design will follow, but before moving on, I will highlight the relevant theoretical questions this experiment can speak to.

Table 4.1: A sample item from Experiment 7.

Sally lives in a small city, where recently there was a citywide election for a new mayor with several candidates, and she had to decide among them on her mail-in ballot.		
Plausibility	Early Disambiguation	Late Disambiguation
Plausible	Last week, because his name is first on this year's ballot, she voted for the candidate that has a progressive platform, Pat Mirabella.	Last week, she voted for the candidate that has a progressive platform, Pat Mirabella, because his name is first on this year's ballot.
Implausible	Last week, because his name is first on this year's ballot, she voted for the candidate that has a big mustache, Pat Mirabella.	Last week, she voted for the candidate that has a big mustache, Pat Mirabella, because his name is first on this year's ballot.

In the first place, I am interested in how quickly and flexibly readers consider causal interpretations of a relative clause. To the extent that comprehenders entertain the causal interpretation, I should observe facilitation for causally-relevant lexical content within the first few words of the relative clause, as seen in many previous studies. In the present experiment, the design permits a novel comparison between Late disambiguation, where a locally-salient causal interpretation is coherent with preceding material, and Early disambiguation, where the locally-salient causal interpretation would not be coherent with preceding material. I expect to observe the standard facilitation effects in response latencies within causally-Plausible relative clauses with Late disambiguation. But if comprehenders' considerations of causal interpretations are subject to rapid influence from the global context, I should observe a reduction or indeed the absence of this facilitation effect in the Early disambiguation condition, driving an interaction of Position with Plausibility in the relative clause region.

Next, I am interested in whether causal interpretations are selected by the end of a relative clause. If comprehenders select the causal interpretation, I should observe reanalysis costs during the reading of material which disambiguates against a causal interpretation after the relative clause. On the assumption that *because* clauses are indeed incompatible with an implicit causal interpretation of a relative clause, they should feature slower response latencies only when Late after causally-Plausible relative clauses, driving an interaction of Position with Plausibility also within the *because* clause region.

Finally, it remains an open question whether rapid consideration of the causal interpretation of a relative clause emerges from a costly generation process. In these items, I assume the decision introduced in the matrix clause provides the main pre-relative-clause evidence to the comprehender that an Explanation may be forthcoming. If I observe evidence that a causal interpretation is under consideration in the early positions of the relative clause, the matrix is plausibly the point when generation must have begun. Under the hypothesis that generation is costly and gated by context, I would thus expect an increase in response latencies in the matrix clause in Late disambiguation conditions, compared to Early disambiguation conditions where an Explanation has already been provided.

4.3.1 Methods

The design, norming procedure, and analysis plan for this experiment were were preregistered (https://osf.io/u36ey). All materials, data, and analysis scripts are available for review in an OSF repository (https://osf.io/gf64q/?view_only=7ee3e5bf20354ad394342f65d51bdf72).

4.3.1.1 Participants

128 native English speakers participated in the experiment on Prolific in late 2022, compensated according to a $12 hourly wage. All participants had US nationality, at least the equivalent of a high school degree, and a minimum of 20 prior submissions with an acceptance rate of 90% on the platform. Ages were within the range of 18 to 40 (with a mean of 31).

4.3.1.2 Materials and Norming

To instantiate the design discussed above, I constructed 80 two-sentence narratives. In each case, an initial context sentence set up a decision problem for a protagonist, and the target sentence described their choice. All target sentences began with a temporal adverbial (e.g. *in the end*, *ultimately*, *last week*), and featured a matrix clause with the protagonist as the pronominal subject, and the chosen entity as a definite description following the verb.

These object definite descriptions contained one of two relative clauses, designed to provide unsurprising properties of the chosen entity which were either a plausible or implausible reason for the protagonist's choice. Implausibly-explanatory properties always described dimensions of the chosen entity which were less salient for the type of choice in context (e.g. voting based on a politician's platform vs. appearance, choosing a dentist based on their reviews vs. hobbies, or watching a show based on its actors vs. setting), rather than undesirable values on a salient dimension. So that a merely restrictive, non-causal reading was always available, head nouns of object descriptions were also selected to be non-unique in the contexts described (e.g. *the candidate*, *the dentist*, *the show*), such that the further description was required to satisfy the canonical requirements of the definite determiner. Both relative clauses featured gaps in subject position. All relative clauses were followed by a nominal appositive identifying the individual by a name or other salient property.

In addition to the potentially-explanatory relative clause, all items featured an invariant *because* clause attributing the protagonist's choice to a property of the chosen entity along a different dimension (e.g. a politician's position on the ballot, a dentist's accepted insurance policies, a show's suitability as background noise). This clause always featured pronominal reference within the subject to the chosen entity. In Late conditions, the *because* clause came sentence-finally, after the nominal appositive, and thus the pronoun could refer anaphorically to the chosen entity just described. In Early conditions, the clause fell between the initial temporal adjunct and the matrix clause, and thus the pronoun was cataphoric.

The length of the matrix clause, including the relative clause and nominal appositive was standardized: relative clauses were always five words in length, the nominal appositive was always two words in length, and the preceding matrix clause, including the definite determiner and nominal head of the relative clause, was always five words in length. Initial temporal adjuncts and the *because* clause were of variable lengths.

In order to validate that Plausible relative clauses were highly plausible, and Implausible relative clauses were highly implausible, item sets were normed in a likelihood judgment task. 96 participants were recruited on Prolific with the same demographic restrictions as the intended experiment. For each of the 80 narratives, they read the context sentence setting up the decision problem, followed by a list of two additional facts: that the protagonist chose a certain entity X, and that X had a property P, either the intended Plausible or Implausible explanation. They then were instructed to rate the likelihood that the property P was the reason why a protagonist chose X, on a four-point scale where "1" was "unlikely" and "4" was "likely". Conditions were Latin-squared so that each participant saw every narrative, and never saw the same narrative in multiple conditions.

(43) **Sample judgment trial from the norming task for Experiment 7**

Sally lives in a small city, where recently there was a citywide election for a new mayor with several candidates, and she had to decide among them on her mail-in ballot.

We know two more things:

- She voted for Pat Mirabella.
- **Plausible**: He has a progressive platform.
  Implausible: He has a big mustache.

How likely do you think it is that she voted for Pat Mirabella because {he has a progressive platform, he has a big mustache}? [1 2 3 4]

Ratings were subjected to a Bayesian mixed-effects ordinal regression analysis. From this model, I extracted estimates on the underlying likelihood scale and compared them to the model's threshholds for the endpoint responses in order to determine whether each item met preregistered criteria for quality. Items were considered high-quality if the 95% highest posterior-density interval around the expected value for their Implausible condition was entirely below the threshold for "1" responses, and the 95% highest posterior-density interval around the expected value for their Plausible condition was entirely above the threshold for "4" responses. In an initial round of norming, too few items met these criteria, particularly failing to receive high-enough ratings for Plausible conditions. After adjustments to contexts and properties were made to increase the salience of the potential explanation, the norming task was repeated with another 80 participants. In this second round, a sufficient number of items met the criteria in order to proceed. 40 item sets met both conditions, and all other items show robust effects of plausibility as well, with most estimates lying in the extreme parts of the scale. Overall response distributions are summarized in 4.1, Table 4.2 reports the parameters of the model fit to the responses, featuring a large and highly-credible effect of Plausibility, $\hat{\beta}$ = 3.62, 95% CRI = (3.35, 3.89), and the itemwise estimates are depicted in 4.2. From these results, I selected 64 critical items for analysis in the Maze study, consisting of the 40 item sets which met both conditions, and

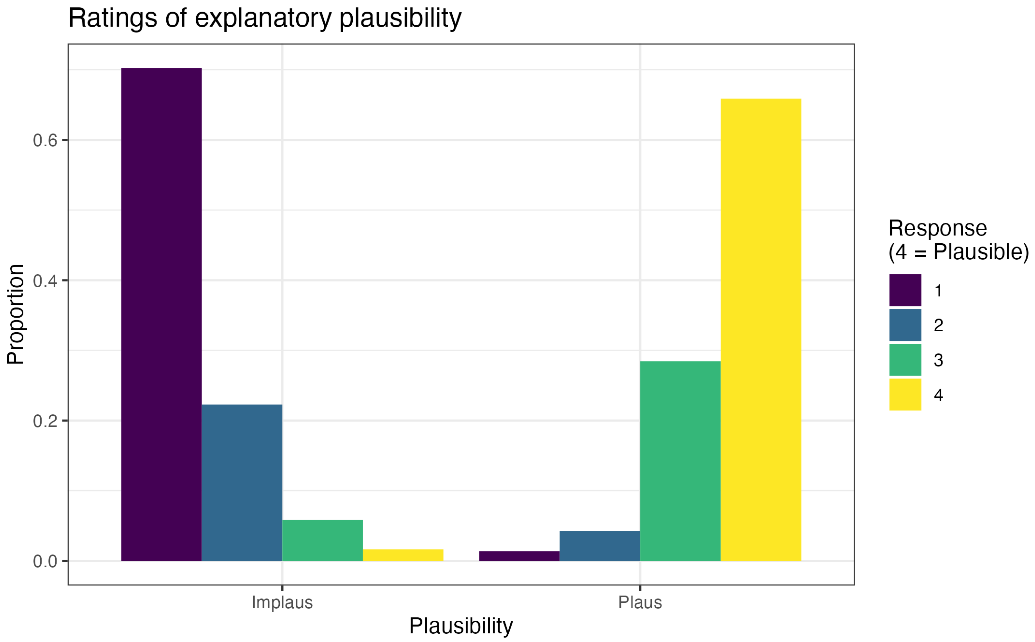

Figure 4.1: Causal plausibility ratings from the norming study for Experiment 7, by condition.

Table 4.2: Bayesian ordinal mixed-effects model fit to 4-point plausibility responses in the norming task for Experiment 7. Factor levels in parentheses were coded as positive.

Effect	Posterior $\hat{\beta}$	Posterior σ_β	95% CRI Lower	95% CRI Upper
Threshold 1|2	0.66	0.10	0.46	0.86
Threshold 2|3	1.77	0.11	1.56	1.98
Threshold 3|4	3.12	0.11	2.90	3.34
Plausibility (Neut)	3.62	0.14	3.35	3.89

the 24 next-best items, where quality was evaluated by the total posterior density for that item set which lies on the intended side of the threshold for the intended low- and high-likelihood responses. To exemplify, the lowest-quality item in this set of 24 edge cases was item 62, which featured a high-quality implausible condition, and a plausible condition with an estimated likelihood which was over the threshold for “4”, but with only 56% of its posterior density. The subset of 40 high-quality items were reserved for a secondary analysis in the Maze experiment which follows; in the end this smaller sample reflected the same patterns as the full sample, and so I report the analyses performed on the full sample.

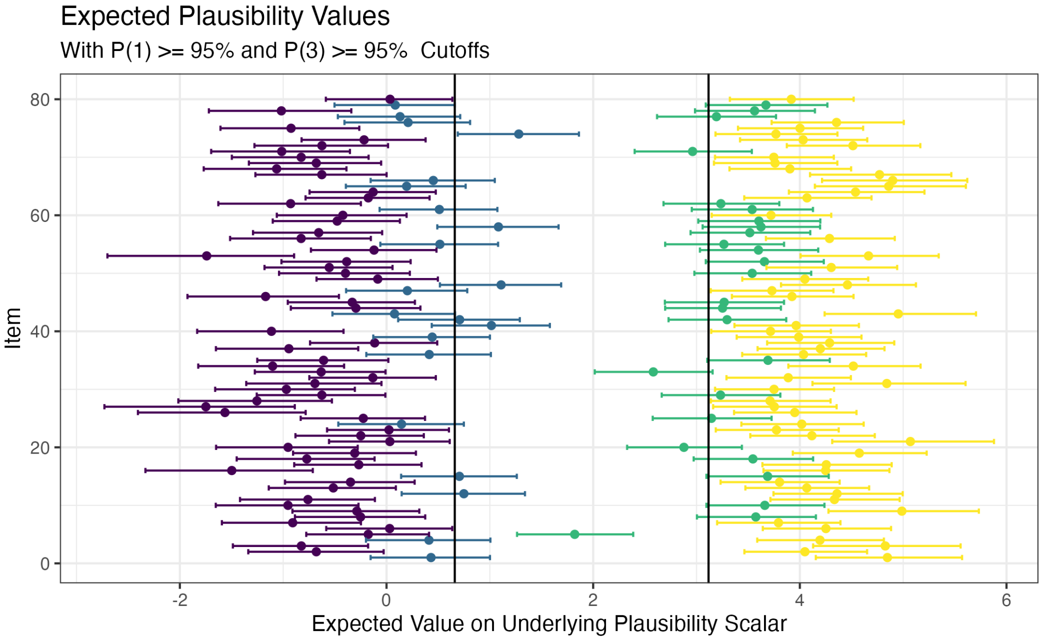

Figure 4.2: Itemwise causal plausibility values extracted from the model fit to norming responses for Experiment 7. Vertical lines indicate thresholds between “1” (“unlikely”) and “2”, and “3” and “4” (“likely”). Points indicate expected values, and error bars indicate 95% highest posterior-density intervals. Purple (darkest) points represent stimuli in Implausible conditions where the 95% HPDI was entirely below the 1|2 threshold, while blue (next lightest) points represent stimuli in the same condition where this cutoff was not met. Yellow (lightest) points represent stimuli in Plausible conditions where the 95% HPDI was entirely above the 3|4 threshold, while green (next darkest) points represent stimuli in the same condition where this cutoff was not met.

Table 4.3: Foil strings from Experiment 7 corresponding to the target sentences in Table 4.1.

Early Disambiguation	Late Disambiguation
x-x-x add, fancies and why wild basis site hall headed going, why idiot camp if alongside shirk damn did happen commanded, Xxx Xxxxxxxxx.	x-x-x add, why idiot camp if alongside shirk damn did happen commanded, Xxx Xxxxxxxxx, fancies and why wild basis site hall headed going.

4.3.1.3 Procedure

The experiment was prepared in Ibex (Drummond, 2010), and deployed on PCIbexFarm. For each item, participants read a context sentence presented all at once, followed by the critical sentence presented in a Maze task.

Maze foils were sampled from length-matched high-surprisal words with reference to the language model of Gulordava et al. (2018) (Boyce et al., 2020). Foils were generated separately for the *because* clause, and the matrix clause with the relative clause(s), both in contexts following the temporal adverbial associated with that item set. Identical foils were generated for the two relative clauses. The particular string of foils for a given condition was then determined according to the Early or Late position of the *because* clause; i.e. the foils generated for the *because* clause were always slotted in at the position of the *because* clause in the target string. Foils for the two-word nominal appositive were always same-length, case-matched sequences of ”x”, to avoid typical high error rates on named characters and place names. An example set of foil strings is given in Table 4.3.

As in Experiments 2, 4, and 6, in order to encourage accurate performance of the Maze task, participants saw a counter at the top of the screen during each Maze decision measuring how many targets they had chosen correctly without an error. This number reset to 0 when participants chose a foil, and the ongoing sentence was immediately terminated, moving prematurely to the comprehension question.

All items were followed by binary forced-choice comprehension questions. These probed information presented in various parts of the narrative, and included both yes/no and content questions.

The 64 test items were presented in pseudo-randomized order across four Latin-squared forms, balanced across our final sample of 128 participants. Participants also saw 80 filler items from various sources, including the 16 narratives which were not carried over from the norming stage, and 16 additional two-sentence narratives constructed to resemble the critical items. All of these 32 similarly-structured fillers featured somewhat plausibly-explanatory relative clauses, without a later *because* clause—the intention being to avoid the possibility of participants learning to avoid all causal inferences. The other 48 filler items were unrelated three-sentence narratives, presented as a sequence of two self-paced chunks followed by a single sentence in the Maze. All fillers were also followed by comprehension questions. Two of the same-design fillers and six of the generic narrative fillers were presented to participants as practice items, and another two plus six were reserved as "burn-in" items, presented at the beginning of the main body of the experiment before participants were shown the first critical item.

After completing all 144 trials, participants completed short exit questionnaires on their experience in the study, and demographic information regarding their language history, before receiving compensation. The procedure in entirety was estimated to take about 55 minutes.

4.3.1.4 Analysis

2296 test trials in which a participant either failed to complete the entire Maze stimulus or responded incorrectly to a comprehension question were excluded from analysis. The remaining sample includes data from 5896 critical trials.

Two measures of word-by-word response latencies were computed for the analysis of test items. All measures relied on residual log response latencies, derived from a linear mixed-effects model fit using the `lme4` package in `R` (R Core Team, 2016; Bates et al., 2015) to log response latencies for words in all unexcluded trials, with fixed slopes for number of characters and position in the sentence, and random participant intercepts. The two critical measures were: (i) summed log RTs across the five-word relative clause region, e.g. *that has a big mustache*, and (ii) summed log RTs within the first four positions of the *because* clause, e.g. *because his name is.* As a post-hoc addition, I also examine summed residual response latencies in the five-word matrix region, e.g. *she voted for the candidate.*

For analysis, Bayesian linear mixed-effects models were fitted to these measures with Stan (Stan Development Team, 2019) using the `brms` package in `R` (Bürkner, 2017, 2018) with principled weakly-informative priors, maximal random effects structures, and treatment-coded predictors. Late *because* clauses and causally-Plausible relative clauses were coded as treatment levels. Weakly-informative non-default priors were adopted for fixed effects and the intercept, $\mathcal{N}(0,1)$. Models were fit on 6 chains of 10,000 iterations (including 2,000 warmup interations), with all other `brms` parameters left to their defaults. I take model parameters whose 95% credible intervals (CRIs) do not contain 0 to indicate noteworthy effects. All models reported feature $\hat{R} = 1.00$ for the parameters of interest.

Maze decision errors were analyzed separately as a secondary measure of incremental difficulty, using Bayesian logistic mixed-effects models fit in `brms`. Principled weakly-informative priors were adopted for the intercept, $\mathcal{N}(-2,1)$, and other fixed effects, $\mathcal{N}(0,1)$, reflecting the expectation that error probability will remain between 0% and 30%, with little likelihood around the edges—people are generally very good at finishing Maze task sentences. Models were otherwise fit and interpreted as above.

4.3.2 Results

All 128 participants retained for analysis answered comprehension questions with accuracy of greater than 75% and an average Maze depth of greater than 50%, which is to say that more often than not, they successfully made it past the halfway point of the Maze stimuli. Other participants who had been recruited were excluded from analysis to ensure a base level of attentive comprehension. Mean comprehension accuracy in the final sample was 89%, Maze completion rate was 86%, and average Maze depth was 92%. In this section, I report the response latencies and error rates in the various regions of interest. A summary of the residual log response latencies across all regions is presented in Figure 4.3.

4.3.2.1 Relative Clause

Residual log response latencies summed across the full relative clause region are presented in Figure 4.4, and broken down word by word in Figure 4.5. Posterior values for $\hat{\beta}$ and σ_β from the linear mixed-effects model along with 95% CRIs are provided for fixed parameters of interest in Table 4.5. We observe a simple effect of plausibility, such that when sentences feature Early disambiguation from *because*, plausibly explanatory relative clauses were read faster than implausibly explanatory ones, $\hat{\beta}$ = -0.38, 95% CRI = (-0.49, -0.27). The absence of a credibly non-zero interaction, $\hat{\beta}$ = -0.02, 95% CRI = (-0.09, 0.05), suggests that this difference is not modulated by the position of *because*: indeed,

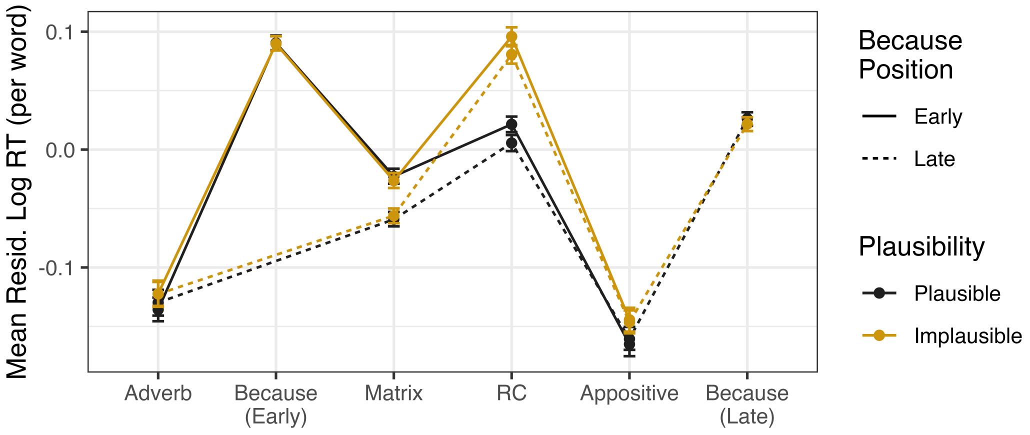

Figure 4.3: Average residual log response latencies in various regions in Experiment 7, by condition.

Table 4.4: Conditional means and measures of spread for the relative clause region in Experiment 7. Standard errors are reported over the sum of raw response latencies in the region, and bootstrapped 95% confidence intervals are reported over the critical measure, summed residualized log response latencies.

Plausibility	Position	Sum RT	SE	Sum Resid. Log RT	95% CI
Implaus	Early	3895	53	0.60	(0.55, 0.63)
Implaus	Late	3735	38	0.56	(0.52, 0.60)
Plaus	Early	3534	28	0.23	(0.19, 0.26)
Plaus	Late	3492	125	0.17	(0.13, 0.21)

marginal comparisons reveal facilitation of a similar size for plausibly explanatory relative clauses with Late disambiguation from *because*, $\hat{\delta}$ = -0.40, $P(\delta < 0)$ = 0.99. Inspection of this difference over the course of the relative clause (see Figure 4.5) suggests that it develops quickly: although there is no visible difference at the first distinct word, the verb, the relative speed of plausibly-explanatory relative clauses holds by the first word of the object onward.

Bayes factor analysis used a model fit with strong priors informed by the results of the relatively-comparable self-paced reading study reported as Experiment 5 in Cozijn (2000), where latencies within a potential tail were measured in conditions where the tail was more or less plausible, and a facilitation on the log scale of about -0.08 was observed.

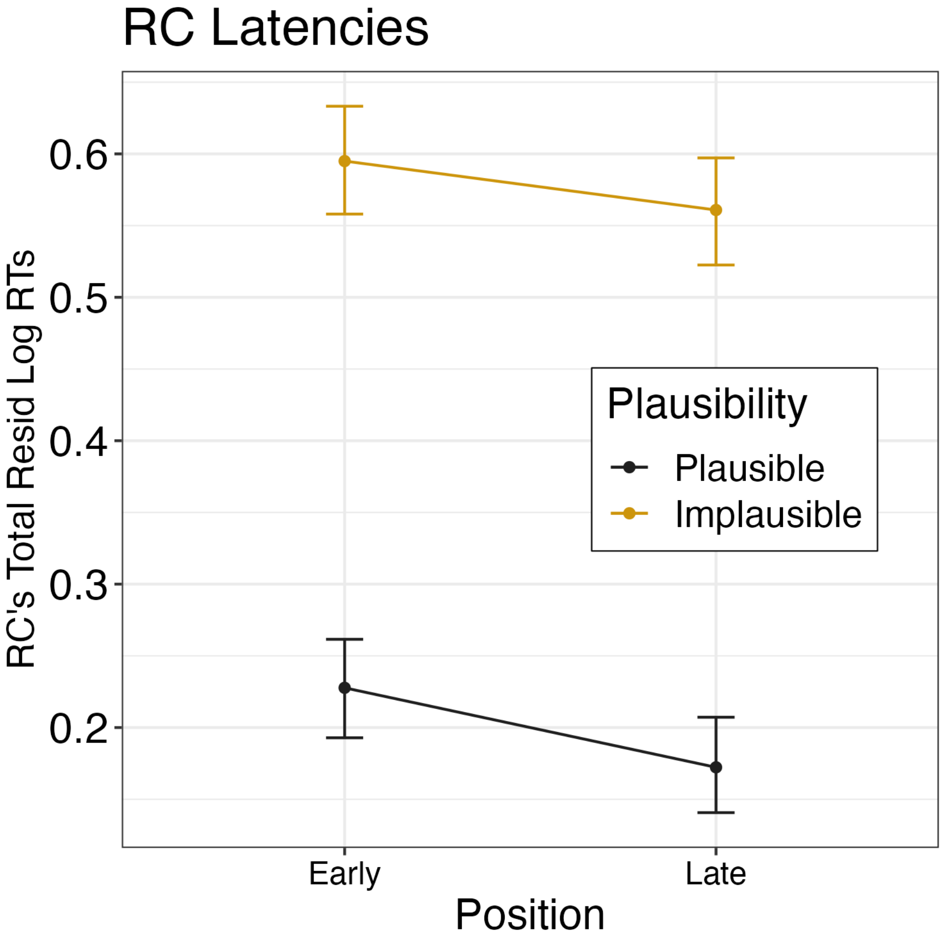

Figure 4.4: Sum residual log response latencies in the relative clause region in Experiment 7, by condition.

Table 4.5: Bayesian linear mixed-effects models fit to summed residual log response latencies in the relative clause in Experiment 7. Factor levels in parentheses were coded as the treatment.

Effect	Posterior $\hat{\beta}$	Posterior σ_β	95% CRI
Intercept	0.61	0.04	(0.53, 0.70)
Position (Late)	-0.03	0.03	(-0.08, 0.02)
Plausibility (Plaus.)	-0.38	0.06	(-0.49, -0.27)
Pos × Plaus	-0.02	0.03	(-0.09, 0.05)

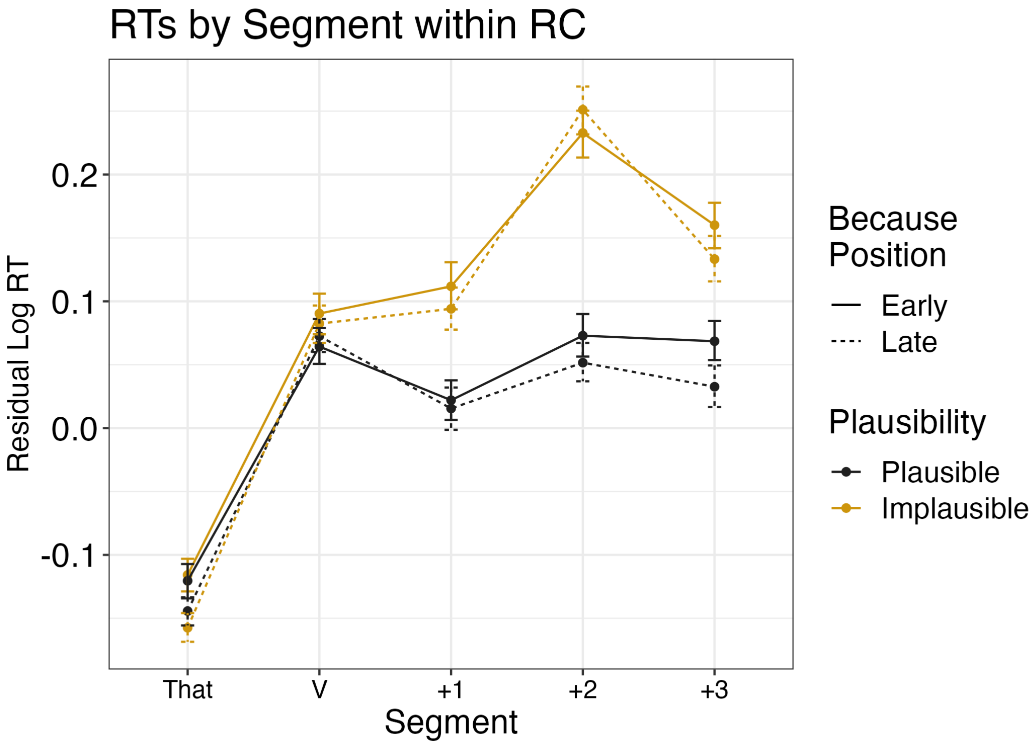

Figure 4.5: Residual log response latencies at each position within the relative clause region in Experiment 7, by condition.

Table 4.6: Informative priors used for Bayes factor analysis of relative clause latencies in Experiment 7, derived from Cozijn (2000).

Effect	Distribution
Intercept	$\mathcal{N}(0.0, 0.10)$
Position (Late)	$\mathcal{N}(0.96, 0.10)$
Plausibility (Plaus.)	$\mathcal{N}(0.0, 0.10)$
Pos × Plaus	$\mathcal{N}(-0.24, 0.10)$

Table 4.7: Bayesian logistic mixed-effects model fit to error rates in the relative clause in Experiment 7. Factor levels in parentheses were coded as the treatment.

Effect	Posterior $\hat{\beta}$	Posterior σ_β	95% CRI
Intercept	-3.21	0.19	(-3.60, -2.84)
Position (Late)	0.03	0.17	(-0.30, 0.36)
Plausibility (Plaus.)	-0.46	0.23	(-0.92, -0.01)
Pos × Plaus	-0.14	0.26	(-0.66, 0.36)

In line with the observation that Maze effect sizes are about three times as large as those observed in self-paced reading (Witzel et al., 2012), priors were constructed to represent the expectation of a -0.24 log ms facilitation for plausibly-explanatory relative clauses in the late disambiguation condition only (Table 4.6). Under these priors, I observe moderate evidence for the presence of an interaction (BF_{10} = 8.77). That is, there is some support for a small reduction in this plausibility effect in the presence of a preceding *because* clause.

Cumulative Maze error rates within the relative clause are presented in Figure 4.6. Posterior values for $\hat{\beta}$ and σ_β from the logistic mixed-effects model along with 95% CRIs are provided for fixed parameters of interest in Table 4.7. We observe here also a simple effect of plausibility, such that when sentences feature Early disambiguation from *because*, participants were less likely to select a foil within an plausibly explanatory relative clause than an implausibly explanatory one, $\hat{\beta}$ = -0.46, 95% CRI = (-0.92, -0.01), moving from an error rate of about 6% to about 4%. The absence of a credibly non-zero interaction, $\hat{\beta}$ = -0.14, 95% CRI = (-0.66, 0.36), suggests that this penalty is not modulated by the position of *because*, although marginal comparisons reveal a numerically larger increase in errors for implausibly explanatory relative clauses with Late disambiguation, $\hat{\delta}$ = -0.60, $P(\delta < 0)$ = 0.99.

Note that these conditions were associated with distinct lexical content inside

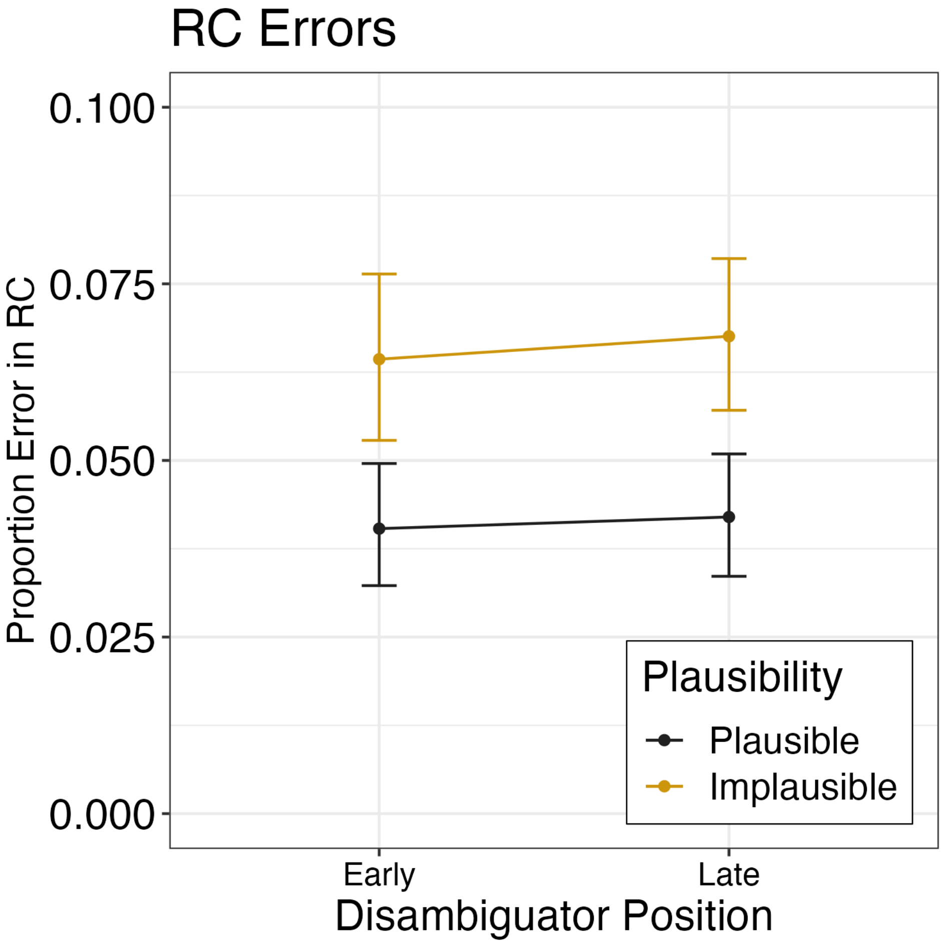

Figure 4.6: Proportion of trials featuring the selection of a foil within the relative clause in Experiment 7, by condition.

Table 4.8: Conditional means and measures of spread for the disambiguator region in Experiment 7. Standard errors are reported over the sum of raw response latencies in the region, and bootstrapped 95% confidence intervals are reported over the critical measure, summed residualized log response latencies.

Plausibility	Position	Sum RT	SE	Sum Resid. Log RT	95% CI
Implaus	Early	3483	30	0.45	(0.42, 0.49)
Implaus	Late	3347	28	0.11	(0.07, 0.15)
Plaus	Early	3481	28	0.45	(0.41, 0.49)
Plaus	Late	3373	28	0.10	(0.07, 0.14)

the relative clause. Further analysis will be necessary to determine whether slower latencies and increased error rates can be attributed to anything more than low-level lexical processing; see §4.3.3.

4.3.2.2 Because Clause

Residual log response latencies summed across the first four words of the *because* clause are presented in Figure 4.7 and Table 4.8, and broken down word by word in Figure 4.8. Posterior values for $\hat{\beta}$ and σ_{β} from the linear mixed-effects model along with 95% CRIs are provided for fixed parameters of interest in Table 4.9. We observe a simple effect of the position of the *because* clause, such that in conditions with implausibly-explanatory relative clauses, Late *because* clauses were read faster than Early, $\hat{\beta}$ = -0.35, 95% CRI = (-0.44, -0.27). Inspection of the position effect over the course of the *because* clause (see Figure 4.8) suggests that it is strongest for the first two words, *because* and the pronoun which always began the clause: after this point the differences begin to diminish. From this distribution, together with the observation that the effect of position has survived residualization for typical position effects, we might conclude that participants struggle with fronted adjuncts more than other sentence-initial material.

The absence of a credibly non-zero interaction, $\hat{\beta}$ = 0.01, 95% CRI = (-0.06, 0.08), suggests that this difference is not modulated by the causal plausibility of the relative clause: indeed, marginal comparisons reveal a similar effect of position in sentences with plausibly-explanatory relative clauses, $\hat{\delta}$ = -0.34, $P(\delta < 0)$ = 0.99. That is, we observe no particular cost for late *because clauses* when they follow a plausible relative clause, as would be predicted if this condition involved the selection and later reanalysis of a causal

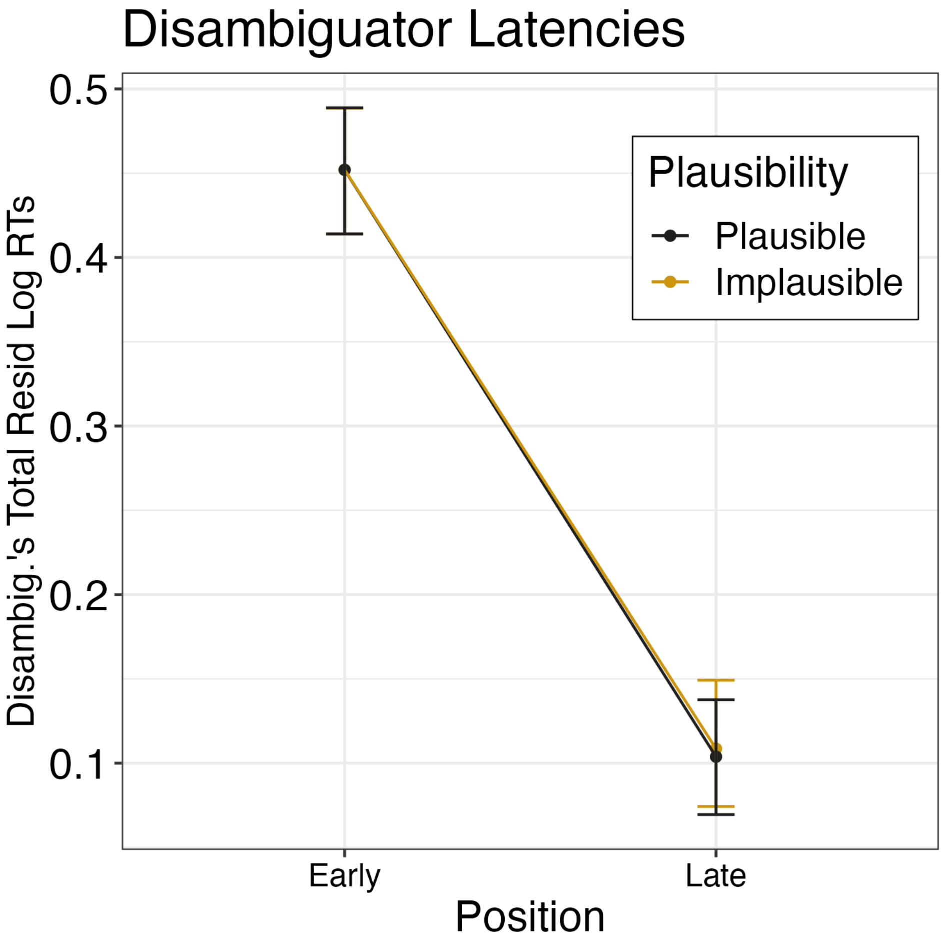

Figure 4.7: Sum residual log response latencies in the disambiguator region (the first four positions within the *because* clause) in Experiment 7, by condition.

Table 4.9: Bayesian linear mixed-effects model fit to summed residual log response latencies in the disambiguator region (the first four positions within the *because* clause) in Experiment 7. Factor levels in parentheses were coded as the treatment.

Effect	Posterior $\hat{\beta}$	Posterior σ_{β}	95% CRI
Intercept	0.46	0.04	(0.38, 0.54)
Position (Late)	-0.35	0.04	(-0.44, -0.27)
Plausibility (Plaus.)	-0.01	0.03	(-0.06, 0.04)
Pos × Plaus	0.01	0.04	(-0.06, 0.08)

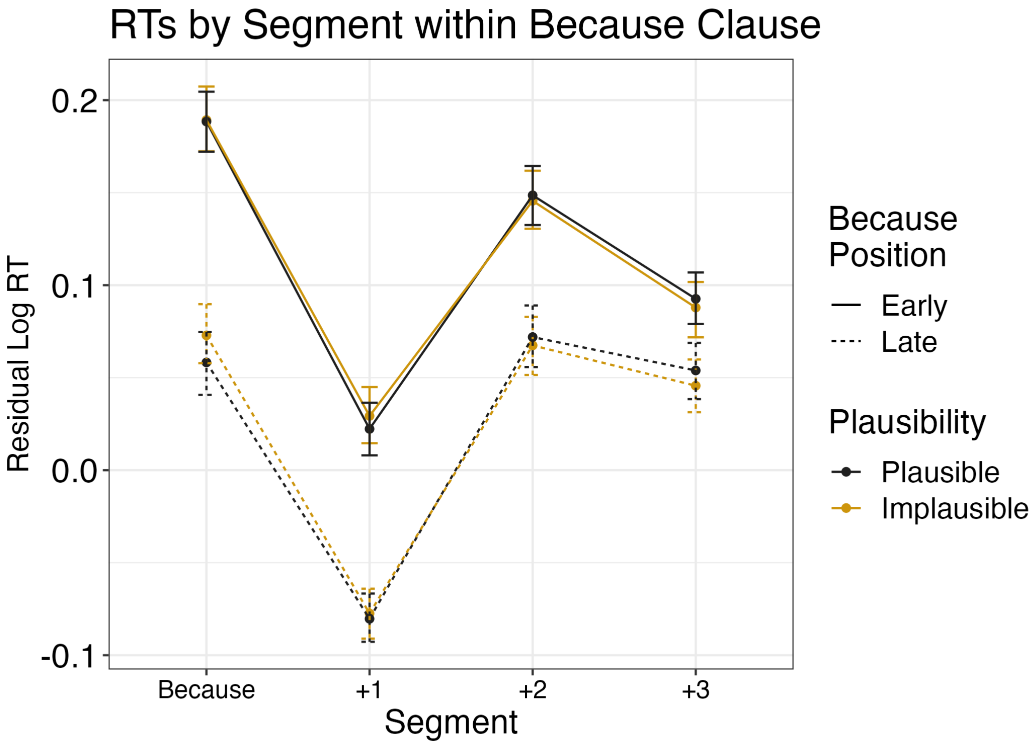

Figure 4.8: Residual log response latencies at each position within the relative clause region in Experiment 7, by condition.

Table 4.10: Informative priors used for Bayes factor analysis of *because* clause latencies in Experiment 7, derived from reanalysis costs observed for polysemes in the Maze in Experiment 2.

Effect	Distribution
Intercept	$\mathcal{N}(0.37, 0.10)$
Position (Late)	$\mathcal{N}(-0.72, 0.11)$
Plausibility (Plaus.)	$\mathcal{N}(0, 0.13)$
Pos × Plaus	$\mathcal{N}(0.33, 0.14)$

Table 4.11: Bayesian logistic mixed-effects model fit to error rates in the *because* clause in Experiment 7. Factor levels in parentheses were coded as the treatment.

Effect	Posterior $\hat{\beta}$	Posterior σ_β	95% CRI
Intercept	-3.39	0.17	(-3.75, -3.06)
Position (Late)	-0.08	0.19	(-0.45, 0.30)
Plausibility (Plaus.)	0.06	0.17	(-0.26, 0.39)
Pos × Plaus	-0.07	0.25	(-0.56, 0.42)

interpretation of the relative clause. I again computed BF_{10} Bayes factors to determine the strength of evidence for or against the expected interaction, here, in the absence of known pragmatic reanalysis effects, specifying priors based on the polysemy reanalysis effects observed in the Maze in Experiment 2 (Table 4.10). The resulting analysis indicates strong evidence against the expected interaction (BF_{10} = 0.05).

Cumulative Maze error rates within the *because* clause are presented in Figure 4.9. Posterior values for $\hat{\beta}$ and σ_β from the logistic mixed-effects model along with 95% CRIs are provided for fixed parameters of interest in Table 4.11. We observe no credibly non-zero effects of interest here, including position, $\hat{\beta}$ = -0.08, 95% CRI = (-0.45, 0.30). Whatever is driving the increased response latencies in early *because* clauses does not fuel increased errors as well, in contrast with the generalized effect of plausibility in the relative clause region.

4.3.2.3 Matrix Clause

As a post-hoc addition to the analysis, I also examined residual log response latencies summed across the pre-relative clause matrix region, presented in Table 4.12 and Figure 4.10. Posterior values for $\hat{\beta}$ and σ_β from the linear mixed-effects model along with

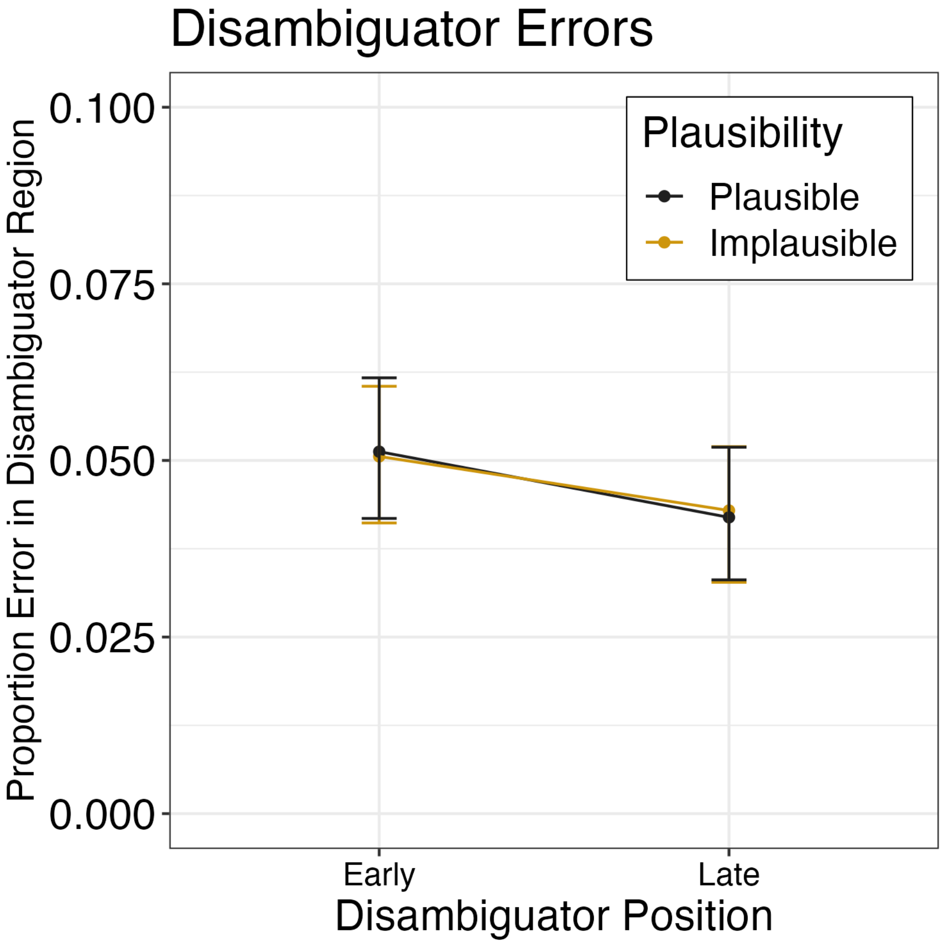

Figure 4.9: Proportion of trials featuring the selection of a foil within the relative clause in Experiment 7, by condition.

Table 4.12: Conditional means and measures of spread for the matrix region in Experiment 7. Standard errors are reported over the sum of raw response latencies in the region, and bootstrapped 95% confidence intervals are reported over the critical measure, summed residualized log response latencies.

Plausibility	Position	Sum RT	SE	Sum Resid. Log RT	95% CI
Implaus	Early	3778	27	-0.21	(-0.24, -0.17)
Implaus	Late	3576	25	-0.31	(-0.34, -0.27)
Plaus	Early	3786	25	-0.21	(-0.24, -0.17)
Plaus	Late	3607	27	-0.32	(-0.36, -0.29)

Table 4.13: Bayesian linear mixed-effects model fit to summed residual log response latencies in the matrix region in Experiment 7. Factor levels in parentheses were coded as the treatment.

Effect	Posterior $\hat{\beta}$	Posterior σ_β	95% CRI
Intercept	-0.20	0.03	(-0.26, -0.13)
Position (Late)	-0.11	0.02	(-0.15, -0.07)

95% CRIs are provided for fixed parameters of interest in Table 4.13. We observe a simple effect of the position of the *because* clause, $\hat{\beta}$ = -0.11, 95% CRI = (-0.15, 0.07), such that response latencies are faster at the matrix when the *because* clause comes sentence-finally. This effect has the opposite direction from what we would expect if the matrix clause triggered a context-gated costly generation process to anticipate a potentially-causal relative clause. Together with the difficulty observed within Early *because* clauses, this might diagnose general difficulty with sentences featuring fronted causal adjuncts, or else perhaps it reflects a process of causal integration with the preceding clause.

4.3.3 Discussion

The patterns in Maze performance presented above yield three main generalizations: (i) plausibly-explanatory relative causes in a decision narrative are read with greater ease than implausibly-explanatory relative clauses regardless of any preceding explanation, (ii) final causal adjuncts do not exhibit reanalysis effects following plausibly-explanatory relative clauses, and (iii) fronted causal adjuncts are associated with some baseline difficulty in comprehension. Generalization (iii) is not particularly of interest to the current study, but I will lay out in this section the potential significance of general-

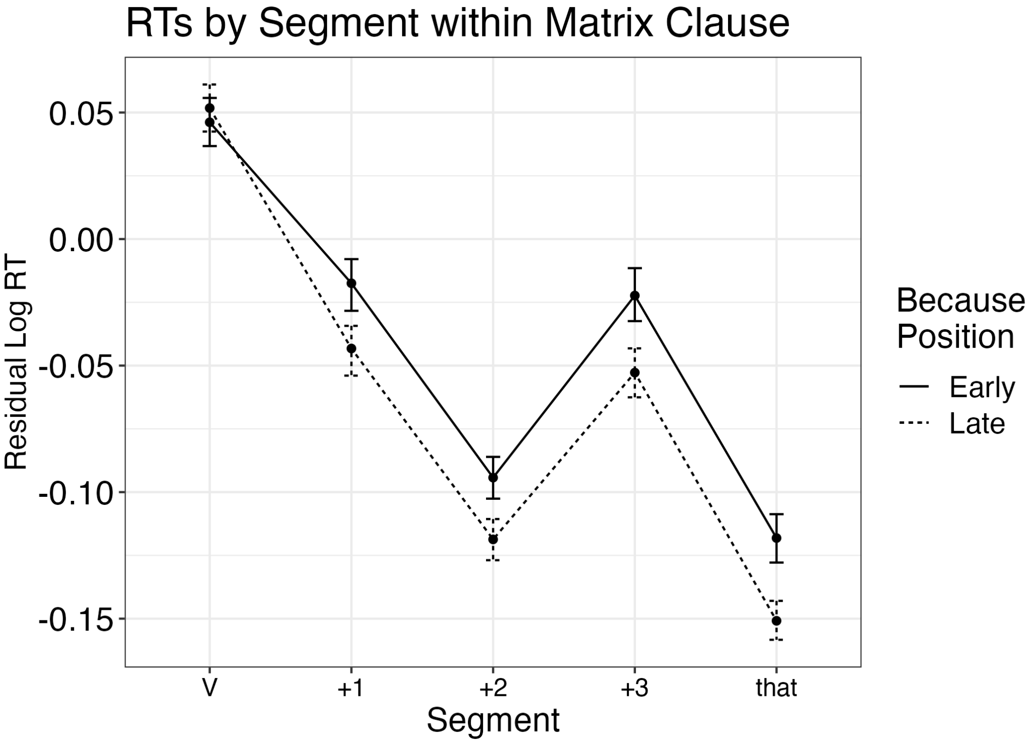

Figure 4.10: Residual log response latencies at each position within the relative clause region in Experiment 7, by condition.

izations (i) and (ii), with particular attention to dismissing an alternative interpretation of (i).

Generalization (i), that plausibly-explanatory relative clauses were facilitated, is supported by a decrease in response latencies and in error rates within those relative clauses. We might take this as partial evidence for facilitation of causally-relevant lexical material, akin to e.g. relative clause reading time facilitation provoked by explanation expectations in Hoek et al. (2021a), or the reductions in N400 responses driven by result expectations in Kuperberg et al. (2011). But we see the same facilitation (perhaps slightly less) in a case where explanatory content would, by assumption, be less likely inside the relative clause, when an explicit explanation has already been given. If this facilitation were derived from expectations for a causal reading of the relative clause, we must conclude that these expectations are not sensitive to context. Is that a reasonable conclusion? In a few cases, we observe that expectations for an explanation can be reduced given the presence of an explicit cue: e.g. the finding in Koornneef and Sanders (2013) that explanation reference patterns are not anticipated when a clause is introduced by a non-causal connective. Nevertheless, it could very well be the case that clear simple cues like connective choice are rapidly incorporated into comprehenders' discourse predictions, while factors of global coherence like a pre-existing alternative explanation enter into consideration later.

On the other hand, drawing a meaningful conclusion about causal processing from these differences alone is irresponsible. Because the plausible and implausible relative clauses varied substantially in their lexical content, it would be more parsimonious to attribute the differences to low-level lexical processing. For instance, it is well-known that shorter, simpler, and more predictable words are processed more easily than words which are longer, more complex, and less predictable (e.g. Erlich and Rayner, 1981; Levy, 2008).[12] The residualization procedures employed as part of the analysis here should have helped control for differences in word length between the critical relative clauses, but differences in predictability independent from the causal reasoning being tested here might have been enough to drive the effect. For instance, it could be the case that plausibly-explanatory relative clauses featured words which are more closely related to the meaning of the head noun, or words which are more likely to figure in choice contexts. It would only be appropriate to attribute the differences in response latency to causal expectations if these words

[12] I'll leave aside whether contextual predictability might be able to subsume the other effects.

were specifically more predictable in the context of an explanation of a choice, but they may well have been more predictable simply in the context of a choice.

In order to compare the evidence for lower-level predictability effects vs. predictability-by-way-of-explanation-expectation effects here, I will present here the results of a toy model comparison, where GPT-2 surprisal values in various contexts (Radford et al., 2019) were extracted to stand in for the predictability of the key content given various types of information available in context.[13] For a demonstration that GPT-2 surprisal serves as a strong linear predictor for Maze response latencies just as for self-paced reading response latencies, see Boyce and Levy (2023).

Surprisal values for the critical relative clause materials were calculated given four different preambles. To approximate predictability of the relative clause content simply as a property of the given head noun, surprisals were extracted following a preamble which merely set up a relative clause following the given head noun (44a). To approximate predictability of the relative clause content as a property of an individual which was chosen within the given choice context, surprisals were extracted following a preamble consisting of the core matrix clause from the target sentence (44b). Finally, to approximate predictability of the relative clause as part of an explanation of the protagonist's choice, surprisals were extracted following a preamble setting up the critical clause as part of a *because* clause explaining the protagonist's choice (44c). Models will be fit using each of these surprisal values to predict the response latencies I observed within the relative clause, and compared using Bayes factors (BF_{12}). To validate this method and ensure that any potential differences among the different surprisals are not driven merely by the length of the preamble, I also compare surprisals extracted following a preamble setting up the critical clause as part of a concessive *even though* clause signalling that the protagonist's choice would not be expected based on the property (44d).

(44) **Preambles used for surprisal extraction**

a. **Property-based surprisal**:
That's the candidate that has a progressive platform.

b. **Choice-based surprisal**:
She voted for the candidate that has a progressive platform.

[13] GPT-2 surprisals were extracted using a helpful Python script by Yizhi Tang and Lisa Levinson at the University of Michigan WordLab, which makes use of the `minicons` package (Misra, 2022), available at https://github.com/UMWordLab/surprisal_with_minicons.

c. **Explanation-based surprisal**:

She voted for that candidate because he has a progressive platform.

d. **Concessive-based surprisal**:

She voted for that candidate even though he has a progressive platform.

For each set of surprisals, I fit a Bayesian linear mixed-effects model using `brms` to predict residual log latencies from surprisal values on each word in the relative clause. To assess whether the effect of surprisal was able to subsume the observed plausibility effect, I extracted the residuals from this model, and fit over them a final linear mixed-effects model using the binary plausibility manipulation as a predictor. Parameters for the effects of surprisal and plausibility in these analyses are reported in Table 4.14. Each surprisal model yielded a credible positive effect of approximately the same size, $\hat{\beta}$ = 0.02–0.03, and in all cases, a credible negative effect of plausibility remained. Nevertheless, effect size for the residual plausibility effect varied, indicating that the predictability of the content in a choice and explanation contexts was able to subsume somewhat more of the plausibility effect than predictability in the radically-different concessive context. Bayes factors indicate that explanation-based surprisal was the best predictor of latencies, compared to surprisals from choice contexts (BF_{12} > 100) and concessive contexts (BF_{12} > 100). Note that concessive-based surprisal also served as a better predictor than choice contexts (BF_{12} > 100), indicating that the predictive power of the *because* context might come from the utility of any connective highlighting causal relevance, as well as the explicit affirmation of a positive causal relationship.

Table 4.14: Coefficients for surprisal predictors across four models using surprisal values extracted from GPT-2 using different preambles, and the corresponding coefficients for binary plausibility predictors on the residuals.

Surprisal Model	Surprisal		Plausibility (Residual)	
	$\hat{\beta}$	95% CRI	$\hat{\beta}$	95% CRI
Property	0.02	(0.02, 0.02)	-0.35	(-0.44, -0.25)
Choice	0.03	(0.03, 0.03)	-0.32	(-0.42, -0.23)
Explanation	0.02	(0.02, 0.02)	-0.32	(-0.41, -0.22)
Concessive	0.02	(0.02, 0.03)	-0.36	(-0.46, -0.27)

This analytical method, attempting to estimate the predictive gain associated with different kinds of contextual information, is admittedly an ad-hoc innovation, with unverified validity. However, if it is to be trusted, it yields two key findings. First, local lexical predictability does not not subsume the plausibility effect observed in response latencies. This was not guaranteed; see Delogu et al. (2017) and L. Levinson (2023) for cases where surprisal was able to account in full for ERP and reading effects attributed to other sources of cognitive difficulty. The relative robustness of the plausibility manipulation here may be due to the use of preambles which were shorter than the full contexts given to participants in the experiment, or else it may reflect the existence of a cost above and beyond low-level lexical predictability. Second, predictability tracks more closely with latencies when it is conditioned on the presence of a causal relationship. This is expected if comprehenders were indeed reasoning about upcoming content under the expectation that an explanation might be coming.

On the whole, I conclude that low-level differences in predictability cannot convincingly account for the differences in latencies across the two plausibility conditions. The slower latencies for implausibly-explanatory relative clauses seem better attributable to a process of explanation expectation, where crucially, expectation is not reduced by the presence of another explanation.

I move on now to generalization (ii). Under a model where a causal meaning for the relative clause, if plausible, is selected before a comprehender exits the clause, likelihood of selection at this point will be highest in the plausibly-explanatory case. Given the further assumption that the causal adjunct is incompatible with the selected implicit meaning, that model expects slower latencies in causal adjuncts following plausibly-explanatory relative clauses, as comprehenders complete reanalysis. Recall that Hoek et al. (2021b), in an eyetracking investigation using a similar design, observed a difference consistent with the predicted reanalysis effect in the position after *because*: a 29ms increase in total fixation duration when the *because* clause followed a plausibly-explanatory relative clause. The movable-adjunct design of the present experiment would allow me to weigh whether this effect should indeed be taken as reanalysis triggered by *because* when the relative clause was plausibly explanatory, or lingering facilitation for *because* when no plausibly-explanatory information had yet been encountered. However, I fail to replicate the basic effect here: at no position within the *because* clause do we observe costs conditional on the causal plausibility of the relative clause.

Why the inconsistency? One possibility is that the design and method of the present study simply lacked the precision and power to observe the Hoek et al. (2021b) effect. Although I will not report a full power simulation here, I am not inclined to believe this is the case. Forster et al. (2009) and Witzel et al. (2012) observe that Maze studies yield comparable effect sizes to eyetracking in most cases, unlike self-paced reading. Given this, power should be somewhat larger in the present study, with 128 participants measured on 64 critical stimuli across 4 conditions yielding 2048 observations per condition (before trial exclusions) compared to the 75 participants, 32 critical stimuli, 4 conditions, and thus 600 observations per condition of Hoek et al. (2021b). Nevertheless, I will note that the effect observed by Hoek and colleagues is ultimately rather small, a 1% increase in log reading times. Typical power in a sentence processing experiment is rarely sufficient to detect such small effects reliably, or to provide convincing evidence of their existence. Beyond power concerns, there are several small differences of design and stimulus between the two studies—the nature of the plausibility manipulation, the presence of a spillover region between the relative clause and *because*, etc. None of these seem particularly obvious sources of the variation observed. Between these two experiments then, evidence is more or less equivocal for a reanalysis effect triggered by *because*. Evidence regarding this potential effect from further experiments with different designs will be useful for settling this question more conclusively.

4.4 Experiment 8: Causal inferences from sentence sequences in self-paced reading

In the remaining two experiments in this chapter, continuing to pursue the same questions of the timecourse of consideration and selection of causal interpretation, I conduct a self-paced reading/Maze task comparison, using a design with three key differences from that of Experiment 7. First, I will replace the plausibility manipulation in Experiment 7 with a more subtle manipulation of plausibility in context. Observing facilitation for plausibly-explanatory content in a potential tail provided evidence that comprehenders were to some extent expecting an explanation. This expectation seems to fuel anticipation of content which could provide a likely explanation, given world knowledge. At the same time, the persistence of this facilitation in the face of a prior explanation suggested that comprehenders may not be taking context into account as they adopt these local expectations.

Are comprehenders' pragmatic expectations rapidly sensitive to subtle features of the context? In the reading studies on scalar implicature reviewed in Chapter 3, there are a few instances that demonstrate comprehenders can take context into account somewhat quickly in their anticipation of enriched scalar meaning, e.g. when an implicit contextual question makes an upper-bound meaning for *some* more likely, comprehenders seem to anticipate reference to a complement set (Breheny et al., 2006; Politzer-Ahles & Fiorentino, 2013; S. Lewis, 2013; Hartshorne & Snedeker, 2014). In the literature on causal inferences, there is less evidence of such subtle dependencies, with experimenters generally exploiting the simple, local cue of an IC verb to manipulate whether context makes an upcoming explanation predictable.

To probe whether comprehenders are rapidly sensitive to more nuanced features of the context in the nature of their expectations, I will adapt a manipulation of protagonist knowledgeability from Bergen and Grodner (2012) and the similar experiments in Chapter 3. The key assumption is as follows: in intentional decision-making contexts, in order for a property of a chosen entity to be an explanation for the protagonist's choice, the protagonist must have been aware that the chosen entity possessed the property.[14] To the extent that comprehenders are expecting **contextually-suitable** explanations, then in contexts where a protagonist is described as ignorant about a certain class of properties, the comprehender should reduce their expectation of content related to those properties within a potential tail.

The next major change from the design of Experiment 7 is a shift from relative clause tails to sequences of sentences. This choice was made simply in order to explore a larger portion of the space of potential causal inferences. One important benefit is that it offers a chance to explore selection of inferences over sentence pairs, where Experiment 7 and the previous study by Hoek et al. (2021b) have only done so with relative clauses. Nevertheless, the evidence from work on sentence pairs reviewed above (e.g. Mak and Sanders, 2013, Koornneef and Sanders, 2013) does not suggest any effects which should be substantially different between sentence pairs and relative clauses.

Finally, rather than probe reanalysis effects within a clause immediately following the potential tail of the Explanation, in these experiments I seperate the tail and the sentence hosting the crucial *because* clause with a full sentence of buffer. It is possible

[14] It's perhaps possible to imagine counter-examples where the explanation is indirect: perhaps I can say that *Sally voted for Mirabella because of his platform* in the case where Mirabella's platform directly brought about an endorsement from a musician Sally trusts unconditionally, and this endorsement guaranteed Sally's vote even though she remained ignorant of Mirabella's political views. These cases seem suitably obscure that we can rule them out as plausible online inferences.

Table 4.15: A sample item from Experiment 8.

Sentence	Knowledgable	Ignorant
C1	Sally lives in a small city, where recently there was a citywide election for a new mayor with several candidates, and she had to decide among them on her mail-in ballot.	
C2	She spent some time reading everything she could about the candidates before mailing in her ballot.	She didn't have any time to read anything about the candidates before mailing in her ballot.
S1	In the end, she voted for Pat Mirabella.	
S2	He has the most progressive platform in the race.	
S3	He's from a very socio-economically diverse area, and has always championed public programs.	
S4	She voted for him because his name was first on the ballot.	

that the selection process for causal inference arises at a relatively slow time-scale, so that even if it is initiated at the end of a potential explanation, selection has not been finalized within the span of a few following words. Alternatively, even if selection can occur fast, perhaps comprehenders generally wait for discourse to continue unfolding before making a firm decision online. In either of these scenarios, we might not observe reanalysis effects when disambiguation against the inference is presented soon after the potential tail, as in Experiment 7, but they would become apparent if disambiguation against the inference was presented at a longer delay. By postponing *because* for a few sentences, we can test the predictions of such proposals.

The result is a simplified single-factor design manipulating protagonist Knowledge of the property described in the potential tail (Knowledgable vs. Ignorant). This design is demonstrated for a sample item in Table 4.15. Again, before moving on to detailed discussion of the methods, I will highlight the relevant theoretical questions this experiment can speak to.

As described above, I expect latencies in the tail of a potential Explanation to reflect whether comprehenders' expectations for the content of a likely explanation are moderated by contextual cues which make typical predictable content unlikely. The potential tail here is given in S2. Comprehenders exhibiting this sensitivity should demonstrate slower response latencies in S2 when the protagonist was described as ignorant to the property it describes.

As for the ongoing question of reanalysis, assuming that comprehenders will be sensitive to the contextual manipulation here, and that the explicit *because* clause in S4 is incompatible with an explanatory reading of S2, I expect latencies in the *because* clause to reflect whether comprehenders engage in selection of causal meaning at some point before S4. That is, comprehenders engaging in online selection of causal meaning should demonstrate slower response latencies in the *because* clause when the protagonist was described as knowledgeable about the property in S2, because this is an instance where context supports the causal inference.

Whether comprehenders take context carefully into account as they anticipate upcoming content, and whether they engage in online selection of pragmatic meaning may be modulated by task demands. I begin here with a self-paced reading investigation, to establish a baseline in a low-demand reading task.

4.4.1 Methods

The design, norming procedure, and analysis plan for Experiments 8 and 9 were preregistered (https://doi.org/10.17605/OSF.IO/52HB7).[15] All materials, data, and analysis scripts are available for review in an OSF repository (https://osf.io/a4vx6/?view_only=9aaaa9ab11254efba98d08deb40008a1).

4.4.1.1 Participants

80 native English speakers participated in the experiment on Prolific in early 2023, compensated according to a $12 hourly wage. All participants had US nationality, at least the equivalent of a high school degree, and a minimum of 20 prior submissions with an acceptance rate of 90% on the platform. Ages were within the range of 18 to 40 (with a mean of 31). This was the same sample who completed Experiment 5.

[15] As in Experiments 5 and 6, this study was planned first as a Maze experiment. After observation of the data, the self-paced reading experiment was carried out for the purposes of task comparison. For ease of comparison to prior work, I present the self-paced reading experiment first before moving on to the originally-planned Maze experiment.

4.4.1.2 Materials and Norming

To instantiate the design discussed above, I constructed 40 six-sentence narratives. The narratives were adapted from the Plausible condition of the 40 narratives used in Experiment 7 which were revealed in norming to exhibit the most consistent high and low judgments of causal likelihood. The first of two context sentences was adopted directly from the original narrative, and a second context sentence was added in which the knowledge state of the protagonist at the time of decision was described, either Knowledgeable or Ignorant about a critical property of the entities they were considering for their choice. Ignorant sentences were often, but not always, constructed by negating assertions about knowledge acquisition made in the corresponding Knowledgeable sentence. The number of words in C2 was always matched across conditions.

Four shorter sentences followed describing the protagonist's choice. S1 contained only the initial temporal adverbial and the matrix clause from the original narrative, with the non-unique definite description substituted for a name or unique description, usually the one provided in the nominal appositive in the original narrative. S2 then contained the critical property, always based on the lexical content of the plausibly-explanatory relative clause from the previous experiment. The exact wording was adjusted to ensure that the relevant property would be interpreted as unique to the chosen entity; in the previous experiment this was ensured by virtue of the relative clause's presence within a definite description. E.g. *the candidate that has a progressive platform* becomes *he has the most progressive platform in the race*). These adjustments led S2 to be of variable length across different narratives. S3 followed, providing more information on the property described in S2, so that S1 remained accessible for further elaboration. In S4, the matrix content of S1 was repeated, followed by the causal adjunct used in the previous experiment.

In order to validate that the context manipulation affected whether comprehenders would assume the protagonist was aware of the critical property in S2, item sets were normed in a likelihood judgment task. Sixty participants were recruited on Prolific with the same demographic restrictions as the intended experiment. This was the same sample as the participants who completed the norming task reported in Chapter 3. For each of the 40 narratives, they saw the first three sentences of the narrative in a cumulative self-paced format, with either the Knowledgeable or Ignorant C2. The nature of presentation of the critical S2 property was manipulated across participants. One group saw S2 presented as part of the narrative, while the other only saw it as part of the question. This was done in order to explore a secondary research question, whether causal anticipation in narrative comprehension might itself contribute to the knowledge attribution task, leading narrative-internal statements that could explain a protagonist's decision to be assumed as part of their knowledge more than the narrative-external control. All participants were instructed to rate the likelihood that the protagonist was aware of the property on a four-point scale where "1" was "unlikely" and "4" was "likely". Conditions were Latin-squared so that each participant saw every narrative, and never saw the same narrative in both Knowledgeability conditions.

(45) **Sample judgment trials from the norming task for Experiment 8**

a. **Narrative-internal property**

Sally lives in a small city, where recently there was a citywide election for a new mayor with several candidates, and she had to decide among them on her mail-in ballot.

i. **Knowledgeable**: She spent some time reading everything she could about the candidates before mailing in her ballot.

ii. **Ignorant**: She didn't have any time to read anything about the candidates before mailing in her ballot.

In the end, she voted for Pat Mirabella.

He has the most progressive platform in the race.

How likely is it that Sally knew that Pat Mirabella has the most progressive platform in the race? [1 2 3 4]

b. **Narrative-external property**

Sally lives in a small city, where recently there was a citywide election for a new mayor with several candidates, and she had to decide among them on her mail-in ballot.

i. **Knowledgeable**: She spent some time reading everything she could about the candidates before mailing in her ballot.

ii. **Ignorant**: She didn't have any time to read anything about the candidates before mailing in her ballot.

In the end, she voted for Pat Mirabella.

> Pat Mirabella has the most progressive platform in the race. How likely is it that Sally knew that? [1 2 3 4]

Overall response distributions are summarized in 4.11, Table 4.16 reports the parameters of a Bayesian mixed-effects ordinal regression model fit to the responses. We observe a large, credibly-positive effect of Knowledgeability, $\hat{\beta}$ = 2.67, 95% CRI = (2.22, 3.11), such that protagonists were indeed judged more likely to know of the critical property in Knowledgeable contexts. The hypothesized effect of Property position was not observed, $\hat{\beta}$ = -0.09, 95% CRI = (-0.42, 0.23), nor was any credible interaction with Knowledgeability, $\hat{\beta}$ = -0.20, 95% CRI = (-0.73, 0.35).

The null finding for Property position is somewhat of interest, as it suggests that evidence about a protagonist's knowledge state from the explicit content in the narrative is far more influential than any effect driven by pragmatic reasoning in offline knowledge attributions. But as to the critical norming objective, I conclude that the intended contextual manipulation has the desired effect regarding the protagonist's knowledge of the critical property.

Planned investigation of the posterior effect estimates per-item revealed more variability than anticipated. Rather than adjust items to meet a stringent pre-registered inclusion threshold, I opted relax the inclusion criterion, and ensure simply that Knowledgeable conditions drove likelihood responses above the scale's midpoint, and Ignorant conditions below the scale's midpoint, for each item set. Three item sets did not meet this condition, all featuring Ignorant conditions which received relatively high-likelihood responses. Each case featured a critical property which was readily inferrable even for ignorant protagonists, e.g. that diet cola has very little sugar. These three item sets were adjusted to avoid this issue before the critical reading experiments.

4.4.1.3 Procedure

The experiment was prepared in Ibex (Drummond, 2010), and deployed on PCIbexFarm. For each item, participants read the two context sentences (C1, C2) presented in cumulative, moving window self-paced reading, followed by the four critical sentences (S1–S4) presented non-cumulatively in fixed-window, word-by-word self-paced reading. These presentation choices were adopted to reduce the fatigue and time required to read the whole six-sentence narrative, and allow the most minimal comparison with the Maze

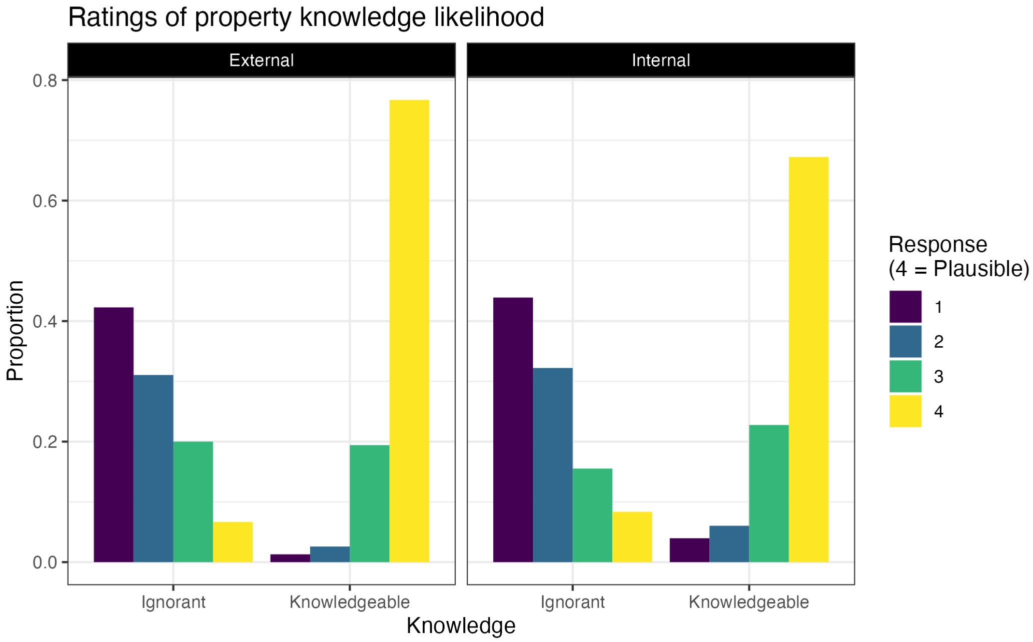

Figure 4.11: Causal plausibility ratings from the norming study for Experiments 8 and 9, by condition.

Table 4.16: Bayesian ordinal mixed-effects model fit to 4-point likelihood responses in the norming task for Experiments 8 and 9. Factor levels in parentheses were coded as treatments.

Effect	Posterior $\hat{\beta}$	Posterior σ_β	95% CRI Lower	95% CRI Upper
Threshold 1|2	-0.30	0.15	-0.60	0.00
Threshold 2|3	0.71	0.15	0.41	1.01
Threshold 3|4	1.81	0.16	1.51	2.12
Knowledge (Knowl.)	2.67	0.23	2.22	3.11
Property (Intern.)	-0.09	0.16	-0.42	0.23
Know × Prop	-0.20	0.28	-0.73	0.35

task of Experiment 6.

Items were followed by binary forced-choice comprehension questions. The topics of the questions were evenly distributed across material in each of the six sentences. For instance, one in six comprehension questions asked a yes/no question checking participants' sensitivity to the knowledge state described in C2 (46a). Another one in six comprehension questions asked participants to indicate the reason for the protagonist's choice, as disambiguated in S4 (46b).

(46) **Example narratives with comprehension questions from Experiment 8**

a. Patricia had a toothache last month, but she had only lived in town for a few months, so she had to find a local dentist to see. She spent some time gathering information about all of the dentists in her area. Ultimately, she went to Doctor Graziano. He gets the best reviews. He has a no-nonsense attitude, and gets his work done very quickly. She went to him because his office is close to her work.

Did Patricia research dentists before her appointment? {Yes, No}

b. Chef Mira has been trying to work out which vinegar to use in the glaze for her restaurant's new asparagus dish, among the four that she had in her kitchen. She tasted each of them as she decided. Last week, she decided on the balsamic. It has the richest taste. It was aged for fifty years in oak barrels. She picked it because it comes from a famous region in France.

Why did Mira pick the balsamic? {Its origin, Its taste}

Accuracy on each type of comprehension question for the 80 participants in the final sample is summarized in Table 4.17. A post-hoc analysis comparing accuracy across the question types finds a few notable effects, including lower accuracy for questions probing the explanation for the protagonists' choice when compared to the other question types, $\hat{\beta}$ = -0.31, 95% CRI = (-0.55, -0.09), higher accuracy for questions about S1 than about either context sentence, $\hat{\beta}$ = 1.17, 95% CRI = (0.59, 1.84), and trends that were not credibly non-zero indicating lower accuracy for C2 questions, $\hat{\beta}$ = -0.57, 95% CRI = (-1.21, 0.07), driven by particular low accuracy in Ignorant conditions, $\hat{\delta}_{Context}$ = -0.68, $P(\delta < 0)$ = 0.91. The latter is likely attributable to a simple "Yes" bias, while the other effects indicate that comprehenders may have struggled somewhat in encoding and retaining information about the context and in arriving at a final decision about the correct explanation for the protagonist's choice.

Table 4.17: Accuracy on comprehension questions in Experiment 8. Standard errors are given in parentheses.

Question Target	% Correct	
	Igno.	Knowl.
C1	93% (2%)	94% (1%)
C2	73% (3%)	90% (2%)
S1	99% (1%)	100% (0%)
S2	96% (1%)	94% (1%)
S3	98% (1%)	95% (1%)
S4	83% (2%)	85% (2%)

Participants completed this experiment at the same time as Experiment 5. Elements of the procedure are briefly reiterated here. Test items were presented in pseudo-randomized order across eight Latin-squared forms, balanced across our final sample of 80 participants. Test items were mixed with 70 filler items from various sources, including the 40 three-sentence narratives from Experiment 5, and 10 additional three-sentence narratives constructed to resemble those item sets. A final 20 six-sentence filler items were constructed to camouflage the test items from this experiment. Crucially, these fillers always left a potential causal inference , and never included an explicit *because* clause to cancel it. All fillers were also followed by comprehension questions. Three of the six-sentence fillers were presented to participants as practice items, and four six-sentence fillers and two three-sentence fillers were reserved as "burn-in" items, presented at the beginning of the main body of the experiment before participants were shown the first critical item.

After completing all 110 trials, participants completed short exit questionnaires on their experience in the study, and demographic information regarding their language history, before receiving compensation. The procedure in entirety was estimated to take about 35 minutes.

4.4.1.4 Analysis

268 test trials in which a participant responded incorrectly to a comprehension question were excluded from reading time analysis. The remaining sample includes reading time data from 2932 critical trials.

Two measures of word-by-word response latencies were computed for the analysis of test items, log response latencies averaged across all positions in S2 (which varied in length) and averaged across the first four positions of the *because* clause in S4. As

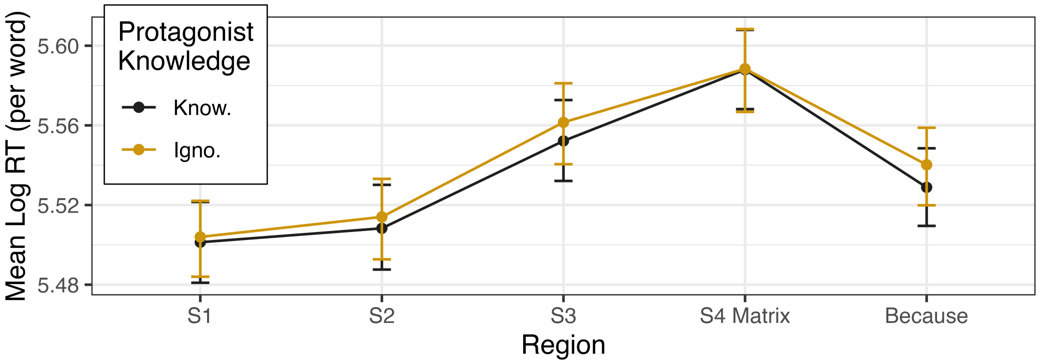

Figure 4.12: Average residual log response latencies in various regions in Experiment 8, by condition.

a post-hoc addition, I also examine average residual response latencies in other regions to develop a sense of behavior across the entire stimulus. For analysis, Bayesian linear mixed-effects models were fitted to these measures using `brms` as described in §3.4.1.3, with Knowledgeable-protagonist contexts coded as the treatment level.

4.4.2 Results

All 80 participants retained for analysis answered comprehension questions with accuracy of greater than 75%, and recorded average response latencies of 1000ms or more on the context sentence. Other participants who had been recruited were excluded from analysis to ensure a base level of attentive comprehension. Mean accuracy in the final sample was 93%. In this section, I report the response latencies in the various regions of interest. A summary of the residual log response latencies across all regions is presented in Figure 4.12 and Table 4.18.

4.4.2.1 S2

Residual log response latencies are broken down word-by-word for the first five positions in S2 in Figure 4.13. Posterior values for $\hat{\beta}$ and σ_β from the linear mixed-effects model over average log latencies in the entire S2 are provided in Table 4.19. I observe a near-credible effect of protagonist knowledge, such that latencies were faster after contexts which presented a protagonist as knowledgeable, $\hat{\beta}$ = -0.02, 95% CRI = (-0.04, 0.00).

Bayes factor analysis used a model fit with strong priors informed by the results of the relatively-comparable self-paced reading study reported as Experiment 5 in Cozijn

Table 4.18: Conditional means and measures of spread for various regions in Experiment 8. Standard errors are reported over the sum of raw response latencies in the region, and bootstrapped 95% confidence intervals are reported over the critical measure, average per-word log response latencies.

	Knowledgeability	Sum RT	SE	Avg. Log RT	95% CI
S1	Igno	2201	35	5.50	(5.48, 5.52)
	Know	2187	34	5.50	(5.48, 5.52)
S2	Igno	1826	27	5.51	(5.49, 5.54)
	Know	1836	28	5.51	(5.49, 5.53)
S3	Igno	2990	53	5.56	(5.54, 5.58)
	Know	3213	264	5.55	(5.53, 5.57)
S4 M.	Igno	1090	60	5.59	(5.57, 5.61)
	Know	1048	24	5.59	(5.57, 5.61)
Bec.	Igno	1109	12	5.54	(5.52, 5.56)
	Know	1107	13	5.53	(5.51, 5.55)

Table 4.19: Bayesian linear mixed-effects models fit to average log response latencies in S2 in Experiment 8. Factor levels in parentheses were coded as the treatment.

Effect	Posterior $\hat{\beta}$	Posterior σ_{β}	95% CRI
Intercept	5.52	0.04	(5.44, 5.59)
Knowledge (Knowl.)	-0.02	0.01	(-0.04, 0.00)

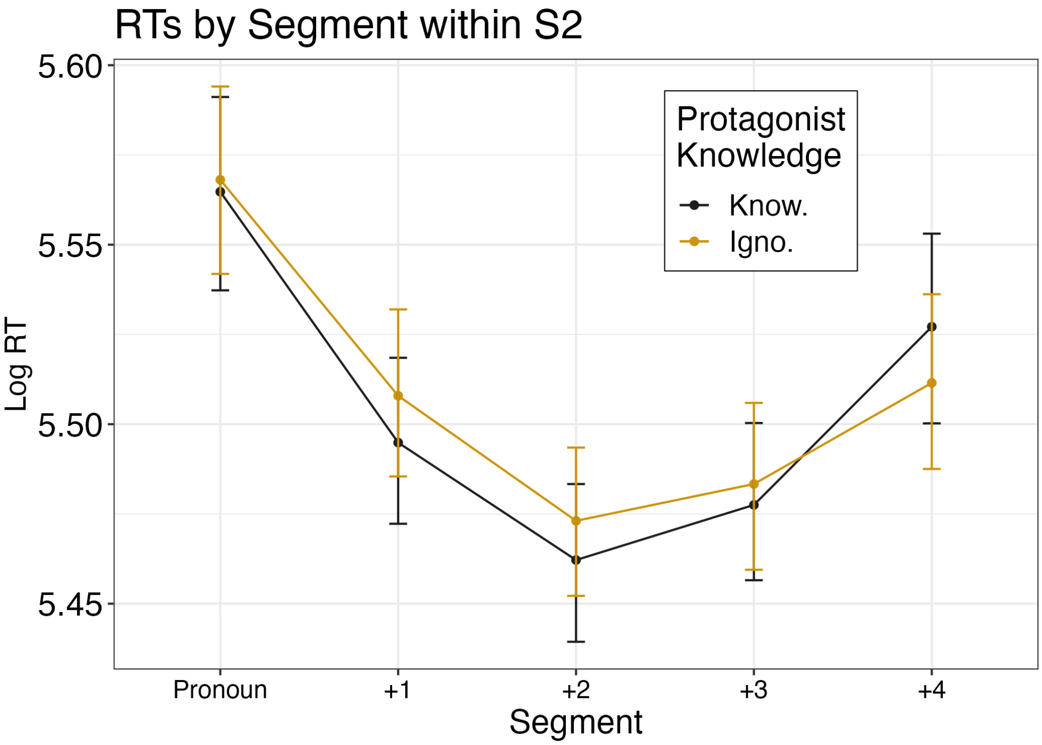

Figure 4.13: Residual log response latencies at each position within the relative clause region in Experiment 8, by condition.

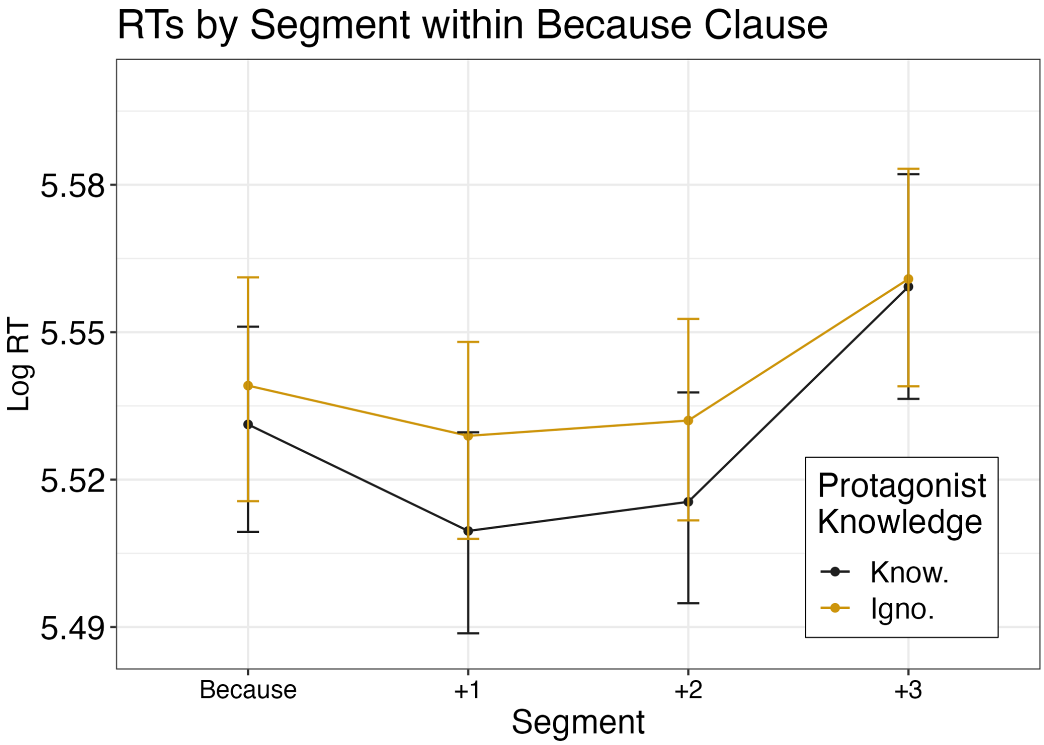

Figure 4.14: Residual log response latencies at each position within the *because* clause region in Experiment 8, by condition.

(2000), where latencies within a potential tail were measured in conditions where the tail was more or less plausible, and a facilitation on the log scale of about -0.08 was observed. Under these priors, I observe moderate evidence for the absence of an effect (BF_{10} = 0.26).

4.4.2.2 S4 Because Clause

Residual log response latencies are broken down word-by-word within the first four positions of the *because* clause in Figure 4.14. Posterior values for $\hat{\beta}$ and σ_{β} from the linear mixed-effects model over average log latencies for these positions are provided in Table 4.20. I observe a just-credible effect of protagonist knowledge, such that latencies were faster after contexts which presented a protagonist as knowledgeable, $\hat{\beta}$ = -0.02, 95% CRI = (-0.04, -0.00), continuing the pattern in S2. Note that this contrasts with the predictions of an account which expects costly reanalysis here more in Knowledgeable conditions.

Bayes factor analysis used a model fit with strong priors informed by the same

Table 4.20: Bayesian linear mixed-effects models fit to average log response latencies in the disambiguator region (the first four positions in the *because* clause) in Experiment 8. Factor levels in parentheses were coded as the treatment.

Effect	Posterior $\hat{\beta}$	Posterior σ_β	95% CRI
Intercept	5.54	0.04	(5.47, 5.61)
Knowledge (Knowl.)	-0.02	0.01	(-0.04, -0.00)

self-paced reading studies of reanalysis that informed priors in disambiguating regions in Experiment 1, where costs on the log scale of about 0.02 were observed in conditions which were more likely to require reanalysis. Under these priors, I observe anecdotal evidence for the presence of an effect (BF_{10} = 1.59).

4.4.2.3 Other Regions

The consistency of the reduction in latencies for narratives in Knowledgable-speaker contexts motivated post-hoc analysis of other regions of the narrative, beginning with the context sentence that introduced the difference in protagonist knowledgeability (C2) and continuing through the core of the narrative, including the matrix clause of S4. Posterior values for $\hat{\beta}$ and σ_β from the linear mixed-effects models over average log latencies within these regions are provided in Table 4.21. No credible effect is observed on the critical context sentence, $\hat{\beta}$ = -0.00, 95% CRI = (-0.01, 0.01), though already by S1 an effect may have been growing, $\hat{\beta}$ = -0.01, 95% CRI = (-0.03, 0.01), consistent with the trend observed in S2. A just-credible effect is then observed in S3 in the same direction, $\hat{\beta}$ = -0.02, 95% CRI = (-0.04, -0.00), and again fails to meet the credibility threshold in the matrix of S4, $\hat{\beta}$ = -0.01, 95% CRI = (-0.03, 0.01), before emerging as credible again in the disambiguation region as noted above. This seems to suggest that if anything, a small and consistent difficulty emerged for the entire narrative when the protagonist was introduced as ignorant, not necessarily tied to their specific ignorance of the property described in S2.

4.4.2.4 Within-Trial Correlation

As an additional, post-hoc examination of the data, inspired by the correlational analyses reported in Chapter 3, I examined the trial-by-trial relationship between response latencies for S2 to response latencies for the various regions of S4. I first fit simple linear

Table 4.21: Bayesian linear mixed-effects models fit to summed log response latencies in other regions of Experiment 8. Factor levels in parentheses were coded as the treatment.

	Effect	Posterior $\hat{\beta}$	Posterior σ_β	95% CRI
C2	Intercept	0.63	0.02	(0.59, 0.67)
	Knowledge (Knowl.)	-0.00	0.00	(-0.01, 0.01)
S1	Intercept	5.51	0.04	(5.44, 5.58)
	Knowledge (Knowl.)	-0.01	0.01	(-0.03, 0.01)
S3	Intercept	5.56	0.04	(5.49, 5.64)
	Knowledge (Knowl.)	-0.02	0.01	(-0.04, -0.00)
S4M	Intercept	5.59	0.04	(5.52, 5.66)
	Knowledge (Knowl.)	-0.01	0.01	(-0.03, 0.01)

Table 4.22: Bayesian linear mixed-effects models fit to residual response latencies in the critical regions of S4, using S2 latencies as a predictor, in Experiment 8.

	Effect	Posterior $\hat{\beta}$	Posterior σ_β	95% CRI
Mat.	Intercept	0.15	0.10	(-0.04, 0.35)
	S2 Latency	-0.03	0.02	(-0.06, 0.01)
Bec.	Intercept	0.14	0.08	(-0.01, 0.29)
	S2 Latency	-0.03	0.01	(-0.05, -0.00)

mixed-effects models predicting average latencies in S4 from average latencies in S2 within the Ignorant conditions, where that dependency is presumed to resemble the standard dependency between any two regions within the same trial (i.e. by assumption there is no causal inferencing going on). These models recorded the expected positive relationships for the matrix clause, $\hat{\beta}$ = 0.75, 95% CRI = (0.70, 0.79) and for the beginning of the *because* clause, $\hat{\beta}$ = 0.73, 95% CRI = (0.69, 0.77). These models were then used to generate predicted S4 latencies from S2 latencies in the Knowledgeable conditions, and residuals were extracted by comparing these predictions to the actual values. Final, more complex models were fit to predict these residuals based on the S2 latencies (Table 4.22). In the models over the residuals in Knowledgeable conditions (Table 4.22), I observe a slight negative relationship, near-credible at the S4 matrix, $\hat{\beta}$ = -0.03, 95% CRI = (-0.06, 0.01), and just-credible at the beginnnig of the *because* clause, $\hat{\beta}$ = -0.03, 95% CRI = (-0.05, -0.00), such that longer latencies in S2 were predictive of shorter latencies in that region.

As in §3.5.2.4, it's useful to know whether this apparent facilitation arises from standard comprehension principles, or experience with the stimuli. A secondary analysis was conducted for the matrix clause, adding exposure counts as a predictor for residual

Table 4.23: Supplementary Bayesian linear mixed-effects model fit to residual response latencies in the S4 matrix, using S2 latencies and exposure counts as predictors, in Experiment 8.

Effect	Posterior $\hat{\beta}$	Posterior σ_β	95% CRI
Intercept	1.00	0.17	(0.67, 1.34)
S2 Latency	-0.17	0.03	(-0.23, -0.11)
Exposure	-0.06	0.01	(-0.08, -0.04)
S2 $\times$ Exp.	0.01	0.00	(0.01, 0.01)

latency (Table 4.23). This model captures a small near-credible first-order interaction between the effect of S2 latencies and the number of exposures, $\hat{\beta}$ = 0.01, 95% CRI = (0.01, 0.01), such that as exposures increase, the strength of this facilitatory effect decreases. Marginal comparisons reveal a large, credible facilitatory relationship at first exposure, $\hat{\delta}$ = -0.16, $P(\delta < 0)$ = 0.99, which is eliminated by the twentieth and final exposure, $\hat{\delta}$ = 0.03, $P(\delta < 0)$ = 0.15. The same pattern is observed at the *because* clause. It would seem that the facilitatory effect here is mainly driven by responses early in participants' experience in the experiment, and reduces over time.

4.4.3 Discussion

Rather than facilitation specific to content which can serve as a predictable explanation, in this experiment we observe if anything a generic difficulty with narratives that describe a decision made by an ignorant protagonist. It has been previously observed that performance on some tasks suffers when participants represent the incomplete knowledge states of other agents (e.g. Samson et al., 2010; Johnson and Keil, 2014), as comprehenders must be doing in Ignorant conditions. I take the overall effect here to be a signature of some general cost like this. The critical alternative analysis is that this is somehow a consequence of the differences in predictions hypothesized about above: that is, in the Knowledgeable condition, comprehenders anticipated the content of S2 as a likely explanation, whereas in the Ignorant condition, explanation-based anticipations were reduced or directed towards properties the protagonist knew of. This does match with the direction of the effect, but not its timecourse. Because trends in this direction emerge as early as S1 and persist through nearly the entire narrative, I find a general cost to be a more likely story.

This apparent pervasive, but small, difficulty aside, we observe no effects of the critical manipulation on reading behavior in any region. This crucially includes the disambiguating *because* region, where we would have expected to see evidence of costly reanalysis processes if comprehenders had selected a causal interpretation of S2 at any point before *because* in S4. The evidence here for the absence of costly reanalysis concurs and strengthens the finding of absence in Experiment 5. It would appear that the absence of reanalysis there was neither due to the atypical construction examined (relative clauses) or to proximity between the potential tail and the disambiguating information.

These results continue to support the conclusion that comprehenders do not select causal enrichments of passages of discourse during online reading, and thus the generalizability of a hypothesis like Rapid Consideration Without Selection for various pragmatic enrichments. However, given the absence of any apparent facilitation effects in S2 conditioned on context in this experiment, we must apparently acknowledge some limits to the factors which can influence early consideration of enrichments.

4.5 Experiment 9: Causal inferences from sentence sequences in the Maze

Experiment 8 followed Experiment 7 in failing to observe evidence for the reanalysis effects predicted by models where causal enrichments are selected online. It also, however demonstrated a case where apparently, a subtle contextual manipulation didn't affect comprehenders' expectations about the content of a potential explanation. In Chapter 3 I observed some evidence that online expectations conditioned on possible pragmatic enrichment were strengthened in the Maze task compared to self-paced reading. In this section, I report a Maze task with the same materials as Experiment 8, providing continued evidence for the absence of reanalysis and the absence of subtle contextual manipulations on expectations in this domain.

4.5.1 Methods

All materials, data, and analysis scripts are available for review in an OSF repository (https://osf.io/a4vx6/?view_only=9aaaa9ab11254efba98d08deb40008a1).

4.5.1.1 Participants

80 native English speakers participated in the experiment on Prolific in 2023, compensated according to a $12 hourly wage. All participants had US nationality, at least the equivalent of a high school degree, and a minimum of 20 prior submissions with an acceptance rate of 90% on the platform. Ages were within the range of 18 to 40 (with a mean of 31). This was the same sample who completed Experiment 6.

4.5.1.2 Materials

The final four sentences of the same 40 test items used in Experiment 8 served as the target sentences in the Maze task. Following Boyce et al. (2020), Maze foils were sampled from length-matched high-surprisal words with reference to the language model of Gulordava et al. (2018). Foils that were too repetitive or were judged as plausible continuations were then replaced by hand. As in Experiments 6 and 7, foils for proper nouns were always same-length, case-matched sequences of "x", to avoid typical high error rates. An example set of foil strings for the target sentences in table 4.15 is given in (47).

(47) **Example foils for Maze portions of Experiment 9**

a. In the end, she voted for Pat Mirabella.

x-x-x lose whom, ago worse ride Xxx Xxxxxxxxx.

b. He has the most progressive platform in the race.

Kid guy go knew catastrophize grateful sick miss glad.

c. He's from a very socio-economically diverse area, and has always championed public programs.

Easy yeah than ones environmentalists permits send, ifs ton forgo mattresses unless appeared.

d. She voted for him because his name was first on the ballot.

Trap smell hill seem weather hear seen trip worry eat buy tonnes.

4.5.1.3 Procedure

The experiment was prepared in Ibex (Drummond, 2010), and deployed on PCIbexFarm. For each item, participants read the two context sentences (C1, C2) presented one at a time in cumulative moving-window self-paced reading, followed by the four

Table 4.24: Accuracy on comprehension questions in Experiment 9. Standard errors are given in parentheses.

Question Target	% Correct	
	Igno.	Knowl.
C1	94% (2%)	91% (2%)
C2	63% (3%)	90% (2%)
S1	100% (0%)	100% (0%)
S2	96% (1%)	100% (0%)
S3	98% (1%)	100% (0%)
S4	93% (2%)	87% (3%)

critical sentences (S1-S4) presented in a Maze task. As in previous Maze experiments reported here, choice of a foil terminated the trial and reset a running score counter.

The same comprehension questions were used as in Experiment 8, targeting various components of the narrative. Accuracy on each type of comprehension question for the 80 participants in the final sample is summarized in Table 4.24. A post-hoc analysis comparing accuracy across the question types finds some of the same patterns as in Experiment 8, including higher accuracy for questions about S1 than about either context sentence, $\hat{\beta}$ = 1.81, 95% CRI = (0.96, 2.78), and trends indicating lower accuracy for C2 questions, $\hat{\beta}$ = -0.74, 95% CRI = (-1.51, 0.02), seemingly driven by particular low accuracy in Ignorant conditions, $\hat{\delta}_{Context}$ = 0.20, $P(\delta > 0)$ = 0.61. As before, difficulty with C2 questions in Ignorant contexts would be attributable to a "Yes" bias to questions asking about protagonist knowledge, while the other effect diagnoses sharper memory for critical portions of the narrative vs. the preceding context In Experiment 8, we also observed that questions probing the explanation for the protagonists' choice (in S4) received lower accuracy than other question types; that is not credibly the case here, $\hat{\beta}$ = -0.23, 95% CRI = (-0.63, 0.17). There is a small depression in accuracy particular to items with Knowledgeable protagonists, however, $\hat{\delta}_{Context}$ = -0.74, $P(\delta < 0)$ = 0.72, which may diagnose some difficulty in narratives where a causal inference was plausible but accompanied later by a different explicit causal statement.

Practice procedure, randomization, filler items, and exit questionnaires were as in Experiment 8, with data collected simultaneously as Experiment 6. This procedure was estimated to take about 60 minutes.

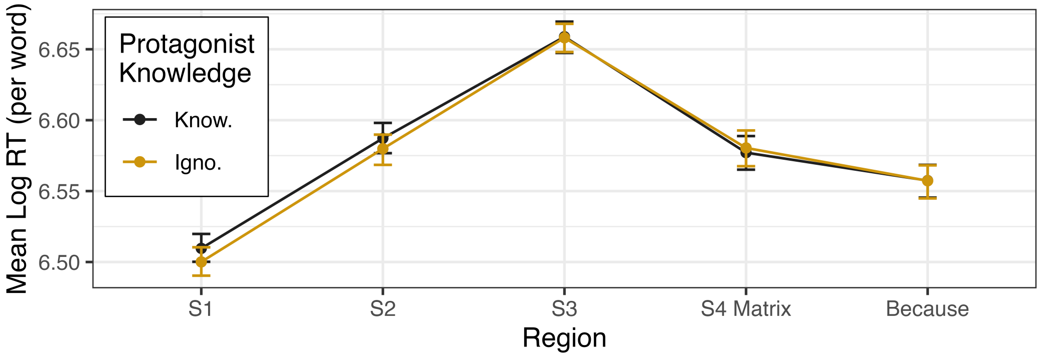

Figure 4.15: Average residual log response latencies in various regions in Experiment 9, by condition.

4.5.1.4 Analysis

1074 test trials in which a participant responded incorrectly to a Maze decision or a comprehension question were excluded from analysis of response latencies. The remaining sample includes data from 2126 test trials. Maze decision errors were analyzed separately as a secondary measure of incremental difficulty, but no patterns of interest were observed. Critical response latency measures and their analysis were computed using the same procedures as in Experiment 8.

4.5.2 Results

All 80 participants retained for analysis answered comprehension questions with accuracy of greater than 75% and an average Maze depth of greater than 50%. Other participants who had been recruited were excluded from analysis to ensure a base level of attentive comprehension. Mean comprehension accuracy in the final sample was 89%, Maze completion rate was 75%, and average Maze depth was 86%. In this section, I report the response latencies in the various regions of interest. A summary of the residual log response latencies across all regions is presented in Figure 4.15 and Table 4.25.

4.5.2.1 S2

Residual log response latencies are broken down word-by-word for the first five positions in S2 in Figure 4.16. Posterior values for $\hat{\beta}$ and σ_β from the linear mixed-effects model over average log latencies in the entire S2 are provided in Table 4.26. I do not ob-

Table 4.25: Conditional means and measures of spread for various regions in Experiment 9. Standard errors are reported over the sum of raw response latencies in the region, and bootstrapped 95% confidence intervals are reported over the critical measure, average per-word log response latencies.

	Knowledgeability	Sum RT	SE	Avg. Log RT	95% CI
S1	Igno	5687	165	6.50	(6.49, 6.51)
	Know	5702	90	6.51	(6.50, 6.52)
S2	Igno	4894	48	6.58	(6.57, 6.59)
	Know	4969	60	6.59	(6.58, 6.60)
S3	Igno	8280	87	6.66	(6.65, 6.67)
	Know	8326	88	6.66	(6.65, 6.67)
S4 M.	Igno	2657	33	6.58	(6.57, 6.59)
	Know	2668	36	6.58	(6.57, 6.59)
Bec.	Igno	3004	24	6.56	(6.55, 6.57)
	Know	2991	22	6.56	(6.55, 6.57)

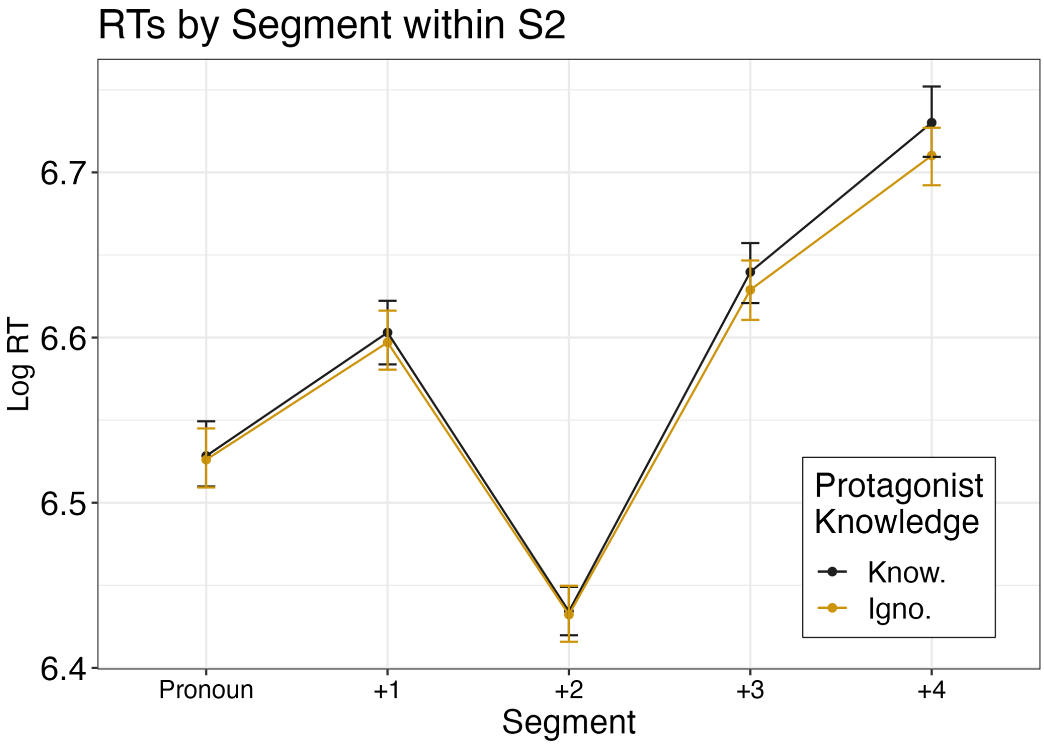

Figure 4.16: Residual log response latencies at each position within the relative clause region in Experiment 9, by condition.

Table 4.26: Bayesian linear mixed-effects models fit to average log response latencies in S2 in Experiment 9. Factor levels in parentheses were coded as the treatment.

Effect	Posterior $\hat{\beta}$	Posterior σ_β	95% CRI
Intercept	6.59	0.02	(6.55, 6.62)
Knowledge (Knowl.)	-0.00	0.01	(-0.02, 0.01)

serve the predicted effect of protagonist knowledge; latencies were not faster after contexts which presented a protagonist as knowledgeable, $\hat{\beta}$ = -0.00, 95% CRI = (-0.02, 0.01), as would be expected if comprehenders rapidly adjusted their expectations for explanations based on features of the context.

Bayes factor analysis used a model fit with strong priors informed by the self-paced reading facilitation effect observed in Cozijn (2000) (see §4.3.2), adjusted by a factor of 3 given the larger effect sizes expected in the Maze, thus centered around an effect of -0.24. Under these priors, I observe extreme evidence for the absence of an effect (BF_{10} < 0.001).

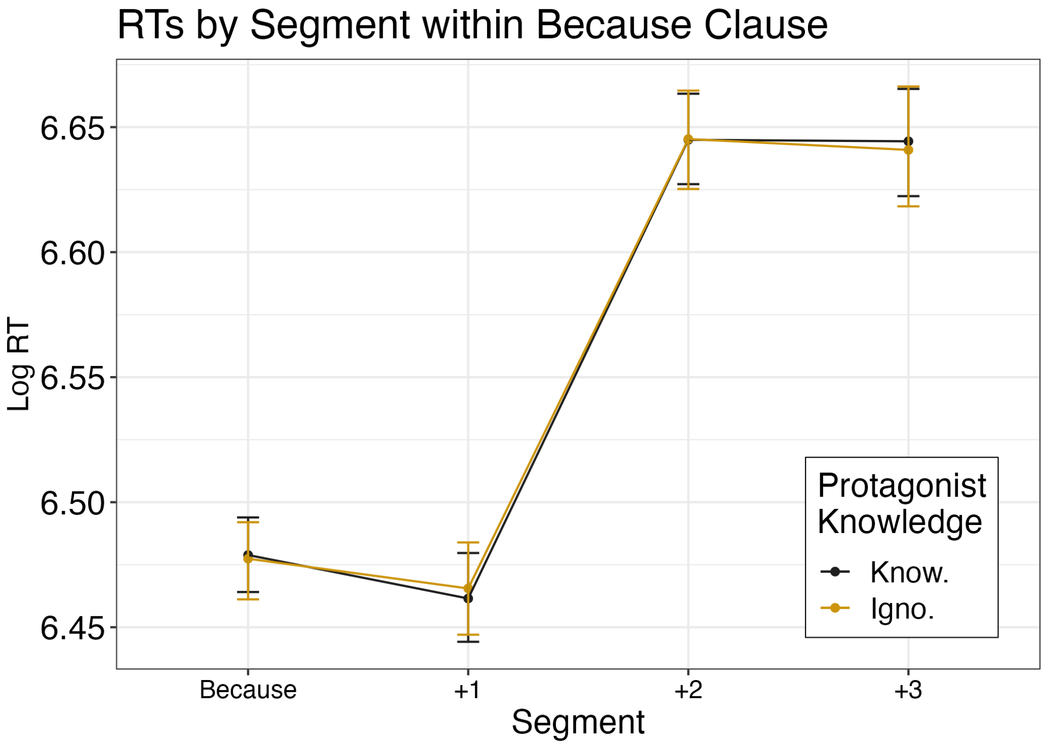

Figure 4.17: Residual log response latencies at each position within the *because* clause region in Experiment 9, by condition.

4.5.2.2 S4 Because Clause

Residual log response latencies are broken down word-by-word within the first four positions of the *because* clause in Figure 4.17. Posterior values for $\hat{\beta}$ and σ_β from the linear mixed-effects model over average log latencies for these positions are provided in Table 4.27. I do not observe the predicted effect of protagonist knowledge; latencies were not slower after contexts which presented a protagonist as knowledgeable, $\hat{\beta}$ = -0.00, 95% CRI = (-0.02, 0.01), as would be expected if comprehenders suffered reanalysis costs in that condition.

Bayes factor analysis used a model fit with strong priors informed by reanalysis effects in self-paced reading, adjusted by a factor of 3 given the larger effect sizes expected in the Maze, thus centered around an effect of 0.06. Under these priors, I observe strong evidence for the absence of the predicted effect (BF_{10} = 0.07).

Table 4.27: Bayesian linear mixed-effects models fit to average log response latencies in the disambiguator region (the first four positions in the *because* clause) in Experiment 9. Factor levels in parentheses were coded as the treatment.

Effect	Posterior $\hat{\beta}$	Posterior σ_β	95% CRI
Intercept	6.56	0.02	(6.52, 6.60)
Knowledge (Knowl.)	-0.00	0.01	(-0.02, 0.01)

Table 4.28: Bayesian linear mixed-effects models fit to summed log response latencies in other regions of Experiment 9. Factor levels in parentheses were coded as the treatment.

	Effect	Posterior $\hat{\beta}$	Posterior σ_β	95% CRI
C2	Intercept	0.65	0.02	(0.61, 0.69)
	Knowledge (Knowl.)	-0.00	0.00	(-0.01, 0.00)
S1	Intercept	6.51	0.02	(6.47, 6.54)
	Knowledge (Knowl.)	-0.00	0.01	(-0.01, 0.01)
S3	Intercept	6.66	0.02	(6.63, 6.69)
	Knowledge (Knowl.)	-0.00	0.01	(-0.01, 0.01)
S4M	Intercept	6.58	0.02	(6.55, 6.62)
	Knowledge (Knowl.)	-0.01	0.01	(-0.02, 0.01)

4.5.2.3 Other Regions

Post-hoc analysis of other regions of the narrative (Table 4.21) finds no evidence for conditional differences elsewhere in the narrative.

4.5.2.4 Within-Trial Correlation

As in Experiment 8, I also examined the trial-by-trial relationship between response latencies for S2 to response latencies for the various regions of S4 in the Knowledgeable conditions, after residualizing based on patterns in the Ignorant conditions. In Ignorant conditions, we observe standard positive relationships between S2 latencies and the S4 matrix clause, $\hat{\beta}$ = 0.52, 95% CRI = (0.44, 0.60) and the beginning of the *because* clause, $\hat{\beta}$ = 0.71, 95% CRI = (0.67, 0.75). In the models over the residuals in Knowledgeable conditions (Table 4.22), I observe a more negative relationship, just-credible at the S4 matrix, such that longer latencies in S2 predicted shorter latencies in that region, $\hat{\beta}$ = -0.06, 95% CRI = (-0.12, -0.00).

A secondary analysis was conducted for the matrix clause, adding exposure counts as a predictor for residual latency in order to examine how this effect was re-

Table 4.29: Bayesian linear mixed-effects models fit to residual response latencies in the critical regions of S4, using S2 latencies as a predictor, in Experiment 9.

	Effect	Posterior $\hat{\beta}$	Posterior σ_{β}	95% CRI
Mat.	Intercept	0.42	0.20	(0.02, 0.82)
	S2 Latency	-0.06	0.03	(-0.12, -0.00)
Bec.	Intercept	0.11	0.18	(-0.24, 0.46)
	S2 Latency	-0.02	0.03	(-0.07, 0.04)

Table 4.30: Supplementary Bayesian linear mixed-effects model fit to residual response latencies in the S4 matrix, using S2 latencies and exposure counts as predictors, in Experiment 9.

Effect	Posterior $\hat{\beta}$	Posterior σ_{β}	95% CRI
Intercept	0.27	0.33	(-0.37, 0.92)
S2 Latency	-0.04	0.05	(-0.13, 0.06)
Exposure	0.04	0.03	(-0.01, 0.09)
S2 × Exp.	-0.01	0.00	(-0.01, 0.00)

lated to experience with the stimuli (Table 4.30). This model captures a small near-credible first-order interaction between the effect of S2 latencies and the number of exposures, $\hat{\beta}$ = -0.01, 95% CRI = (-0.01, 0.00), such that as exposures increase, so does the strength of this facilitatory effect. Marginal comparisons reveal a small, non-credible facilitatory relationship at first exposure, $\hat{\delta}$ = -0.04, $P(\delta < 0)$ = 0.83, which increases many times over by the twentieth and final exposure, $\hat{\delta}$ = -0.17, $P(\delta < 0)$ = 0.99. It would seem that the facilitatory effect here is mainly the result of experience with these stimuli.

4.5.3 Discussion

Behavior for these stimuli in the Maze task shows no dependence on the critical manipulation. We observe evidence that contextually-possible explanations were not facilitated more than contextually-impossible explanations, and evidence that contextually-possible explanations did not induce online selection that caused reanalysis costs on a later explicit *because* clause. (We also fail to observe the global costs for narratives with ignorant protagonists observed in Experiment 8.) These results continue to mismatch the predictions of any model where comprehenders are sensitive to contextual features like protagonist knowledge in the kinds of assertions they expect and accept as explanations for that protagonist's behavior.

The absence of this kind of sensitivity does not appear to be subject to task variation: whatever pressures for better expectations provoked different behavior in the Maze task in Chapter 3 did not change the basic expectation behavior here in ways that I can measure. We even observe in both tasks instances of partial facilitation between reading times early in the narrative and reading times towards the end of the narrative. One difference is how these within-trial dependencies developed over experience with the stimuli. In the Maze task, we see the effect emerges through experience with the stimuli. Again, I assume that latency-dependent facilitation like this simply demonstrates the presence of slowly-developing expectations, which are stronger when comprehenders read more slowly, and I refrain from advancing any particular explanation for the basic variance in the earlier reading latencies. In this case, the conclusion would be that by the end of the experiment, comprehenders in the Knowledgeable-protagonist condition were especially able to anticipate the content of the S4 matrix based on the content of S2. It was certainly plausible that comprehenders anticipated the S4 matrix, as it was always a repetition of the choice described in S1, but it's unclear why the Knowledgeable context in particular supported these expectations. I decline to further analyze the effect, but finding a second case of stimulus-anticipation effects in the Maze is informative to our understanding of participants' behavior in this task at a general level.

In this light, the self-paced reading effect is somewhat puzzling—it reflects an early facilitatory dependency which disappears over time. The disappearance seems related to the observation that self-paced reading participants seem particularly prone to decreased attention over time (see e.g. increased noise in the second half of Experiment 1). Whatever facilitatory process is happening there, comprehenders abandon it by the end of the experiment. This still doesn't explain the nature of the initial process. The discussion above would suggest that we should take this as a case where participants engaged useful predictions as S2 unfolded that facilitated processing in S4. But in self-paced reading, unlike the Maze, those expectations seem to have been present from the beginning of the experiment, rather than the result of experience with the stimuli.

4.6 General discussion

Across the three experiments reported in this chapter, I observe consistent evidence against any difficulty processing a *because* clause after a discourse segment which provided information that could have served as a plausible explanation. This arbitrates against a maximally-incremental model of causal inferences in discourse. On such a model, we would expect comprehenders to select a causal enrichment during the integration of the relevant tail when plausible. Under the assumption that the causal enrichment is incompatible with the presence of an explicit *because* clause, on encountering the *because* clause, the comprehender would have to abandon their enrichment, and adopt a different analysis of the discourse function of that segment. But we observe no slowdowns in reading behavior at *because* consistent with a costly reanalysis process, and so I conclude that comprehenders do not engage in online selection of causal enrichment in these experiments.

Nevertheless, other findings accord with the existing literature in demonstrating that causal enrichment was considered online, and indeed, these results help define some of the parameters of that consideration. In Experiment 7, I observe that in contexts where comprehenders might be expecting explanations for a certain action, relative clauses which could explain that action were facilitated relative to relative clauses which would not (see also Hoek et al., 2021b and Mak and Sanders, 2013). I have argued that this could be the reflex of an online consideration of an explanation schema for the discourse, which in turn drives expectations for content which would be compatible with that schema, given the action to be explained. An exploratory comparison among various surprisal correlations in §4.3.3 supports the active role of causal meaning here: the observed reading behavior for the content of these critical relative clauses is better captured by the contextual probability of the clause after *because* than the contextual probability of the clause in its actual context.

However, I observe that this hypothetical consideration must be contextually-naive, in two ways. First, in Experiment 7, the process of expectation described above seems to be active even when an explicit explanation has already been advanced by a preceding *because* clause. If comprehenders indeed disprefer multiple explanations of the same action, it must be the case that the online consideration of an implicit explanatory reading is nevertheless insensitive to the presence of an existing explanation. This is somewhat of a surprise, as this process of online consideration has been found to be neutralized rather quickly in the face of other discourse connectives which cue other interpretations (Millis et al., 1995; Koornneef & Sanders, 2013). Still, the presence of an existing explanation, as the outcome of a comprehension process itself, might reasonably be a type of information which is less available to the incremental comprehender than a bottom-up cue like a discourse connective.

Second, in Experiments 8 and 9, the degree to which content is facilitated through an expected explanation for a protagonist's actions was not contingent on the protagonist's knowledge state. Specifically, properties which might be expected as world-knowledge-plausible explanations for a protagonist's action were not facilitated any less when the protagonist was introduced as ignorant to that property. I conclude that when comprehenders consider an explanatory reading of upcoming content, their expectations for the content are sensitive to world knowledge, but not context-specific features which restrict admissible explanations.

These results arbitrate against both maximally-incremental comprehension, and a naive model of pragmatic enrichment which is strictly post-hoc. They are most compatible with the hypothesis of Rapid Consideration Without Selection advanced in Chapter 3 for the timecourse of scalar implicature, together with a constraint on the role of context in generating the second-order expectations conditioned on considered enrichments. Note, however, that unlike the literature on scalar implicature, the literature on causal enrichment in discourse has not investigated the timing or cognitive-load-contingency of enrichment-consistent responses in a truth value judgment task. For scalar implicatures, I argued that slow and resource-contingent responses reflected a process of difficult offline selection—in the absence of that kind of effect here, we have no evidence that causal enrichment is ever selected by a comprehender with full commitment. To support the argument that this model generalizes across many types of implicit pragmatic enrichments, future work should investigate comparable offline effects for this phenomenon, and others.

I will close this section by following up on two topics: first, reflecting on the (non-existent) effects of protagonist knowledge given what we know about ignorance in domain-general causal attribution, and second, revisiting the assumption that multiple explanations are avoided. The latter discussion raises important doubts about my conclusion re: selection, but I'll show that even allowing for the offline interpretability of multiple explanations, given a standard account the semantics of *because*, we still expect online difficulty when comprehenders encountered the explicit explanation.

4.6.1 Sensitivity to knowledge in the explanation of motivated behavior

In Experiments 8 and 9, comprehenders failed to use a protagonist's knowledge state to constrain expected explanations. This may not be strictly a failure of the incremental comprehension system, but indeed a vulnerability of human causal attribution more generally. Johnson and Keil (2014) report a judgment study where participants read narratives with protagonists who were knowledgeable or ignorant about some feature F which helped distinguish between choices in a decision problem. In one experiment, the authors measured the degree to which participants took the value of F as a potential explanation by asking whether they thought the protagonist's behavior required additional explanation, depending on the choice she made. The authors observed that desire for further explanation was higher when the protagonist made a choice with a non-optimal F value, even when the protagonist was explicitly ignorant of F. In another experiment, the authors asked participants to predict the protagonist's choice; likewise, here, participants indicated that the choice with the optimal F value was most likely, even when the protagonist was explicitly ignorant of F. This would suggest that comprehenders generally struggle to take into account an agent's limited knowledge state when attributing reasons for their behavior. Note, however, that participants in Johnson and Keil (2014) were somewhat able to take character knowledge into account in their offline judgments, as they were less sensitive to the value of F when the protagonist was described as ignorant. It seems likely that this kind of contextual manipulation is difficult for comprehenders to take into account during incremental comprehension, and enters consideration more fully only as comprehenders settle on a final interpretation. I assume that offline interpretations solicited for the stimuli in Experiment 8 and 9 would reflect some sensitivity to the protagonist's knowledge, even if the online measures show no differences. Comprehension questions in at least Experiment 9 did show some patterns in line with this expectation. This should be further validated in future work.

Another note on the influence of character knowledge state: in self-paced reading, I observed a global cost in narratives with ignorant protagonists, attributed to the costly maintenance of a distinct knowledge state. However, this effect was not observed in the Maze task of Experiment 9. It is possible that this could diagnose less attention to preceding context in the Maze, attributable to the clear division between directly-presented context sentences and the main body of the narrative presented in the Maze. However, it can't be so simple: comprehenders were queried about the protagonist's knowledge state in one-sixth of the critical comprehension questions. In both experiments, comprehenders did in fact do somewhat poorly on these questions, mainly due to over-attribution of knowledge in the Ignorant conditions consistent with either the results of Johnson and Keil (2014) or a classic 'yes'-bias. This pattern was more or less the same in self-paced reading and the Maze. If comprehenders were equally attentive to character knowledge state in both experiments, why did it only slow down comprehenders in self-paced reading? Perhaps the hypothetical increase in attentive comprehension in the Maze (Forster et al., 2009) was responsible, if more attentive comprehension can overcome this kind of small baseline difficulty. I leave this possibility to future work.

4.6.2 Multiple explanation

I have suggested here that reanalysis is absent because *because* clauses do not appear to trigger any especially difficult processing when they follow content that could have prompted the online selection of an implicit explanatory meaning. If we trust the logic of the design, this suggests that implicit explanatory meanings for relative clauses or S2 are not selected online, however much they may be considered online. Without selection, we don't expect costly reanalysis. This picture, of rapid consideration without online selection, is quite similar to the proposal entertained in Chapter 3 for scalar implicature.

But what if we don't trust the logic of the design? In particular, could it be possible that there is no necessary conflict between implicit explanatory readings and the presence of an explicit *because* clause? Notice that this would also explain why early *because* clauses were not used to rule out explanation expectation as diagnosed by response latencies in the relative clause region (§4.3.3).

The original assumption here was that comprehenders are unwilling to admit more than one explanation for the same eventuality, such that explicit explanations force implicit explanations to be re-analyzed as some other kind of information. However, it must be admitted that cases of multiple explanation are frequent in actual text. Consider the examples in (48), obtained from the Corpus of Contemporary American English (Davies, 2008).

(48) **Multiple explanation**

a. We are running it all each day **because we think these hearings are important** and **because we think it is important that you get a chance to see the whole**

thing and make your own judgments.

(Jim Lehrer on broadcasting the Watergate hearings, 1973)

b. **Because of the cultural element in play here**, and **because Bortus apparently does not want to press charges**, Klyden will not be prosecuted.

(from *The Orville* episode "Primal Urges", 2019)

c. Work was something I did **because I guess I was good at it**, and **because I had to earn a living to support my family**.

(interviewee, *Dateline* episode "Murder on a long, dark stretch of road," 2012)

d. … I want to learn to produce food. And that's not only **because I think the cities may be starved out by dollar hyperinflation**. It's also **because my ancestors were good, honest farmers**.

(commenter on *alt-market.com* article "Decentralization is the only plausible economic solution left," 2012)

It is clear that there must formally be some way that multiple explanations can be offered for the same actuality.

Indeed, many theories of causal meaning permit that there can be multiple true explanations of the same actuality.[16] One typical claim (see e.g. D. Lewis, 1973; Mackie, 1974; Wright, 1985) is that expressions like *P because Q* will be true when the condition Q is counterfactually necessary for a particular set of conditions to guarantee the explanandum P. Crucially, there can often be many such conditions.

For example, to adapt an example from Wright's (1985) discussion of the evaluation of causal statements in tort law, consider a case where two small fires merge and cause damage (49). Even if neither fire would have caused the damage on their own, authors largely agree that one could truthfully assert both (49a) or (49b).

(49) **Merging fires**

Bella and Mortimer each start a small accidental fire in their home. These fires then merge, and become large enough to do damage to the home of a third party on their block, Cassandra.

a. Cassandra's house was damaged because Bella started a fire.

b. Cassandra's house was damaged because Mortimer started a fire.

[16] I am grateful to Dean McHugh (p.c.) for valuable suggestions and conversations that helped sharpen my insights in the ensuing discussion.

Example (49a) is expected to be true because, although Bella's fire would not have been sufficient to cause damage on its own, it is part of a set of conditions (Bella's fire + Mortimer's fire) that was sufficient, and Bella's fire was indeed necessary for that set to be sufficient, as Mortimer's fire wouldn't have been sufficient on its own. Similar reasoning holds for (49b).

If we accept that this is at least one way for an explanation to be true, as is standard,[17] then we can indeed predict at least one globally acceptable, non-contradictory reading for a sentence like (50), or an adapted version of one of our stimuli, like (51), where both potential explanations are present.

(50) Cassandra's house was damaged because Bella started a fire, and because Mortimer started a fire.

(51) Sally voted for Pat Mirabella **because he has the most progressive platform** and **because his name was first on the ballot**.

Example (51) can be true with the above rough semantics if Mirabella's platform and ballot position together were enough to guarantee Sally's vote, and neither condition was sufficient on its own.

It is nevertheless hard, by my judgment, to access this interpretation for (50) or (51) with any clarity. Why might this be? Work on counterfactual reasoning in the semantics of natural language has long observed that there is much flexibility in how we imagine realities different from the actual (e.g. Stalnaker, 1968; Kratzer, 1981). What is key here seems to be my willingness to vary whether a known and related fact Q' is true as I counter-factually consider e.g. whether an actual condition Q completed a sufficient set of conditions. By my own judgment I can adjust my willingness to accept a causal statement like (49a) or (49b) based on contextual parameters like whether an eventuality was expected or not. For instance, consider if Mortimer's fire was an annual prescribed burn on his property, while Bella's was a complete accident. In this context the statement

[17] In fact, authors often admit causal claims with even weaker contributions, e.g. Wright (1985) describes a similar scenario that has legal precedent, where Mortimer's fire would have been sufficient on its own. Under Wright's particular construal of the condition above, called there the Necessary Element of a Sufficient Set (NESS), Bella's fire may still be judged a necessary member of a sufficient set with portions of Mortimer's fire. Essentially, Wright finds it relevant that if Mortimer's fire were existent but smaller, Bella's fire would have been necessary to bring about the damage. Nevertheless, courts have often held in such cases that the actual sufficiency of another condition on its own removes the fault of insufficient contributions like Bella's; see Wright's discussion of cases of *overwhelming force*. Per the discussion below, the possibility for variation in these cases can be attributed to variation in exactly which known facts we permit to be false as we reason counter-factually.

in (49a) seems very much true, but (49b) less so. This could be out of an asymmetry in whether I hold constant the existence of the other fire (Q'): when thinking about Bella's accidental fire in (49a), I take as given Mortimer's prescribed burn, and so find that Bella's fire will always complete a sufficient set, but when thinking about Mortimer's prescribed burn in (49b), I may find it easier to imagine cases where the other fire never occurred, as it was accidental, in which case, Mortimer's burn won't always complete a sufficient set.

I suggest that this flexibility over the presence of other contributing conditions, plus the way that we might typically resolve it when comprehending narrative, makes multiple explanation a case of incremental contradiction, although a globally consistent interpretation can ultimately be reached. Essentially, I argue that the strength of the causal meaning we typically derive from an single explanation is incompatible with the meaning that is offered by a second explanation provided in the context of the first. Given this, I still derive the prediction that late *because* clauses should have resulted in costly reanalysis in online comprehension if an implicit explanation had already been constructed and selected. In the remainder of this section, I will demonstrate how exactly this incremental contradiction arises, given a semantics for *because* like the sketch above, together with reasonable assumptions about how comprehenders construct causal backgrounds.

For precision's sake, I will use a simplified semantics for *because* adopted from McHugh (2023a). McHugh's proposal is a useful exemplar in the modern tradition of formal semantics for causation descending from the counterfactual approach outlined by D. Lewis (1973), with particular attention to the importance of counterfactual alternatives (see also McHugh, 2020, 2023b). While there are several innovations that help this proposal account for difficult edge cases, the relevant features here are largely shared among classic approaches (D. Lewis, 1973; Mackie, 1974; Wright, 1985) and other contemporary descendants (e.g. Beckers, 2016). Following McHugh, I take P *because* Q to feature two key truth conditions, a positive condition and a negative condition, as in (52).[18]

(52) **McHugh's semantics for *because***:

$$[\![P \text{ because } Q]\!] \rightsquigarrow p \wedge q \wedge \Box_{f,g}(q)(p) \wedge \neg\Box_{f,g}(\neg q)(p)$$

a. Positive condition: Given the circumstances (modal base f and ordering source

[18] Take $\Box$ to indicate counterfactual necessity. McHugh (2023a) in fact presents a simpler representation for this denotation, where the final conjunct is derived by the exhaustification of the former, see also McHugh (2023b). This is avoided here solely for reasons of clarity, as is the extension of this account to a production theory of causation.

g), the truth of q guarantees the truth of p.

(Q is sufficient for P)

b. Negative condition: Given the circumstances (modal base f and ordering source g), the falsity of q does not guarantee the truth of p.

(not-Q is not sufficient for P)

The critical assumption I will add is that in cases of multiple explanations, the condition of the first explanation typically becomes a fixed component of the modal base for the evaluation of the second explanation (53).[19] This would be a natural consequence of the assumption that all entailments from the preceding discourse dynamically enter the modal base, given that P *because* Q entails Q, but here I will simply stipulate it.

(53) **Preceding conditions enter the base**

When interpreting P *because* Q *and because* Q', if the modal base for P *because* Q is f, the modal base for P *because* Q' is $f' = f + Q$

I will now step through the incremental comprehension of example (51). First, the comprehender interprets (54). In addition to the fact of both simple propositions here, this entails, via the positive condition, that Mirabella's progressive platform (p) guaranteed Sally's vote (s), given some modal base f (54a). It also entails, via the negative condition, that in the absence of Mirabella's progressive platform, Sally's vote was not guaranteed, given f (54b). Next, the comprehender interprets (55). In addition to Mirabella's position on the ballot, this entails, via the positive condition, that Mirabella's ballot position (b) guaranteed Sally's vote (s), given the modal base f' (55b), where f' includes Mirabella's platform (55a). It also entails, via the negative condition, that in the absence of Mirabella's ballot position, Sally's vote was not guaranteed, given f' (54c).

(54) Sally voted for Pat Mirabella because he has the most progressive platform.

a. $\Box_{f,g}(p)(s)$

b. $\neg\Box_{f,g}(\neg p)(s)$

(55) ... and because his name was first on the ballot.

a. $f' = f + p$

[19] In Chapters 3 and 4, McHugh (2023a) lays out a logic for how the elements of a complex sentence fix aspects of its modal base and the counterfactual alternatives which are available (see also Ciardelli et al., 2018), but examples like this do not figure in the discussion.

b. $\Box_{f',g}(b)(s)$

c. $\neg\Box_{f',g}(\neg b)(s)$

At this point, we have a contradiction. Entailment (54a) would have it that wherever Mirabella has the most progressive platform, Sally votes for him. Entailment (55b) requires that there is at least one possibility in modal base f' where Sally does not vote for Mirabella—otherwise, her vote would be guaranteed even in the absence of his ballot position. However, f' does not admit any possibility that Sally doesn't vote for Mirabella, because f' includes the proposition that Mirabella has the most progressive platform, and by (54a), in all such cases, Sally votes for him.

This inconsistency, however, is not fatal. Because the modal base is always fixed in the pragmatics, the comprehender may return to (54) and amend f to include the fact just learned, Mirabella's ballot position. Now, (54a) is weaker: Mirabella's progressive platform only guarantees her vote among cases where he is first on the ballot. No contradiction remains, and in fact, we arrive at the intuitive reading suggested for (51) above, where each condition is a necessary component of a sufficient set which includes the other.

Under this approach, the incremental difficulty faced by the comprehender as they interpret a multiple cause sentence is much like any case where an implicit restriction is clarified by a speaker retroactively, whether in modal quantification or quantification at the level of individuals (56).

(56) **Retroactive restriction**

a. You may not be able to escape your federal tax burden, but you can lessen your state tax burden—if, that is, you can move to another state.

(from "Your state tax burden," *Consumers Research* 17(10), 1990)

b. The second thing, as far as a state to state, yeah, you can do that, but not with autonomous trucks, and that is because every state, the ones that have autonomous truck legislation, it is different.

(Patrick Penfield on C-SPAN, Oct. 2021)

c. Of course, I too have the video on my website. The first one, that is.

(comment on the *jwz* blog, Dec. 2010)

To my knowledge, this difficulty has not been measured, for multiple explanation or otherwise, but I presume it should come with a cost, to the extent that an interpretation for the retroactively restricted expression has been selected before the restriction is added. For the same reasons, in the circumstance where a comprehender has selected an implicit causal interpretation of a discourse unit before encountering a later explicit explanation, I imagine that if they choose to retain that causal interpretation, they will have to engage in a costly procedure of revising their assumed modal base.

This incremental contradiction, to be clear, is not a strictly necessary conclusion of the theory of causal meaning I have entertained here. It rests on the assumption in (53); if each of two explanations are evaluated entirely independently, so that the first explanatory condition is not held constant during evaluation of the second, comprehenders may relate the two explanations to each other as they see fit, without incremental contradiction. This does not seem likely to me, as it mismatches what seems to be the natural interpretation of naturally-occurring examples of multiple explanation. Further work examining this component of the semantics of *because* and the pragmatics of multiple explanation is necessary to be more certain here.

I will conclude this section by noting that the fact of multiple explanations, even given the possibility of incremental contradiction provided above, suggests that in the early-*because* conditions in Experiment 7, comprehenders may have been more welcome to entertain an explanatory parse for the relative clause than initially assumed. That is, it is in fact semantically possible for comprehenders to have fully interpreted the initial *because* clause and yet still expect a second explanation to be furnished in the relative clause. This would be another explanation for the continued presence of plausible-explanation facilitation there, although offline completion data from Hoek et al. (2021a) does suggest that comprehenders are not very likely to expect additional explanations for an action after one has already been given.

4.6.3 Conclusions

In this chapter, I've considered how causal inferences may relate to the theory of the incremental comprehension of scalar implicatures advanced in Chapter 3. Across three experiments, I see no evidence for the costly reanalysis predicted if causal inferences received selection online. This result, together with the common observation that potential causal inferences nevertheless drive online expectations, suggests that indeed, Rapid Consideration Without Selection may be a viable theory of incremental comprehension for many types of pragmatic enrichments. This would separate the processes of intepretation-conditioned expectation and interpretation selection for many types of uncertain meanings.

Chapter 5

Conclusion

In the three previous chapters, I have profiled the incremental comprehension of five different linguistic phenomena with uncertain meaning. Key benchmark effects thought to diagnose the selection of a single meaning, as observed in previous work and measured further through performance in novel SPR and Maze experiments, are subject to notable variation across these five cases. I attribute this variation to differences in the timing of selection.

The existing understanding of such timing differences in the field of sentence processing has largely been a simple binary, between decisions made rapidly during incremental processing, and decisions which are temporarily deferred. I have highlighted two empirical insights from the present studies which can enrich our existing understanding beyond this binary. First, evidence from Chapter 2 suggests that the comprehension mechanism's preference for deferment is flexible and capable of reversal under the right circumstances. This motivates a hypothesis that selection timing is a variable and strategic component of behavior sensitive to the utility and risk of a firm decision. Second, evidence from Chapters 3 and 4 suggests that even in cases of deferred selection, the comprehension mechanism may generate, consider, and rank the multiple possible analyses in a way that drives expectations for upcoming input. This motivates a model for incremental comprehension that takes seriously both interactive expectation-building and the selection and construction of a single interpretation as component processes.

I would like to dwell in this section on the theoretical consequences of that move, including especially the desiderata for a theory of selection timing and the relationship between expectation and comprehension. I will begin in the next section by widening the discussion to include a few more cases of uncertain meaning that can be captured within the present framework. In §5.2, I then consider the task for a functional account of selection timing given this expanded picture, suggest some lessons about the nature of the selection process, and draw comparisons with major existing accounts, before considering the limitations of my argument and future directions in §5.3.

5.1 Enriching the empirical picture

My discussion so far has centered around the timing of selection for five different constructions, which have exhibited one of three kinds of basic behavior. Homonymous nouns, as has been prototypically observed in previous work, are selected immediately, driving garden-path effects, and also exhibiting subordinate selection costs when context encourages an interpretation which is dispreferred by context-independent factors (meaning dominance). Two linguistic phenomena, polysemous nouns and distributivity, were observed to vary in the timing of their selection, driving garden-path effects most clearly in experiments where immediate selection would be more strategic. A final two phenomena, scalar implicature in the interpretation of the quantifier *some* and causal inference driven by coherence, were uniformly observed to lack garden-path effects, even at very late disambiguation, although other evidence suggested preferred interpretations still played some role in expectations for upcoming content. I have argued that these final cases should be understood as long-term deferment of selection, with rapid consideration of the possible interpretations in the meantime.

Across these five case studies, even as the presence of garden-path effects varied, I have not observed any evidence against two other empirical cornerstones of my approach, costs for selection of an interpretation which is somehow dispreferred, and evidence that the potential analyses of input with uncertain meaning are generated and receive rapid incremental consideration even in lieu of selection. The former effect, which I refer to as a subordinate selection cost, is not always obvious: in Experiments 3 and 4, for instance, distributive interpretations were dispreferred, but apparent costs associated with constructing the preferred collective interpretations obscure any ability to observe costs for selecting distributive interpretation. In other cases, experiments simply have not been constructed in a way that would observe such costs, as is the case for causal inferences. Nevertheless, in the other cases, subordinate selection costs are present wherever selection seems to occur.

In this section, I will bring in evidence from literatures on five further types of input with uncertain meaning: ambiguities of quantifier scope, temporal order, aspect, modifier attachment, and discourse anaphora. Although there are many cases here where data is contradictory, I will demonstrate how each might be understood using the mechanisms I have entertained here. Notably, if my tentative outlook in each case is correct, these five further case studies fall into the same three categories already observed: rapid selection (discourse anaphora), flexible deferment of selection (quantifier scope, aspect) and uniform deferment of selection (temporal order, modifier attachment). After this review, I will move on to profile how the entire collection of ten cases might inform a strategic theory of selection timing.

5.1.1 Revisiting lexical comprehension

Before moving on to the constructions that have not already been discussed, as an addendum to the discussion in Chapter 2, I want to add one note about the processing of homonyms and the immediacy of selection. The classic evidence for rapid selection comes from studies on homonymous nouns, but as we consider homonymous verbs and homonyms which cross syntactic category, the timecourse looks somewhat different. In a classic earlier study, Frazier and Rayner (1987) examined eye movements during the processing of approximately equi-biased noun/verb homonyms like *trains* in sentences where the syntactic context was compatible with both meanings (57). Comprehenders in such sentences seem to delay specification until at least the following word; the eye movement record shows a trade-off in the timing of a costly selection process, which occurs on the target in predisambiguated controls, but is delayed until the following disambiguation region in the conditions in (57).[1]

(57) **Cross-category homonymy** (taken from Frazier and Rayner (1987))

I know that the desert trains...

a. young people to be especially tough. (*trains* = V)

b. are especially tough on young people. (*trains* = N)

[1] This delay does not obtain when preceding syntactic context resolves the ambiguity. In that case the two meanings appear to compete immediately, using syntactic fit as a decisive source of evidence much like semantic context (Folk & Morris, 2003).

Similarly, Pickering and Frisson (2001) examined eye movements during the processing of transitive homonymous verbs, and observed that costs for subordinate selection did not emerge on the verb itself, but after the verb's object if at all. Both of these are scenarios where disambiguating information is intuitively very likely to come in the immediately following position.[2] In such cases, the observation that the comprehender is willing to temporarily postpone a decision briefly is in line with the idea that when strategic, all types of comprehension decisions may be subject to delay.

5.1.2 Quantifier scope ambiguity

Sentences featuring a universally-quantified argument and an indefinite singular argument (58) have (at least) two interpretations in English and many other languages. Formal semantics generally characterizes this as an abstract ambiguity of scope, often corresponding to two possible implicitly-distinct structures, one where the linearly-first argument takes higher scope (a "surface scope" reading) and one where the linearly-second argument takes higher scope (an "inverse scope" reading).[3] E.g. on a surface scope reading of (58), there is one salient cashier, who greeted every customer, and the discourse might continue by referencing that one cashier (58a) but not by referencing a plurality of cashiers (58b). Reference to a plurality would disambiguate towards the inverse scope reading, where for every customer, there is at least one cashier who greeted them, and so a potential plurality of cashiers.

(58) **Quantifier scope ambiguity**

A cashier greeted every customer.

a. The cashier…

b. The cashiers…

Offline data on co-argument scope ambiguities suggests that comprehenders greatly prefer surface scope (Kurtzman & MacDonald, 1993; Anderson, 2004; Dwivedi, 2013), which has led many to expect a bias for surface scope in ambiguity resolution, and corresponding difficulty accessing inverse scope. And indeed, sentences pre-disambiguated to inverse

[2] It would be beneficial to evaluate this intuition with corpus evidence.

[3] From the perspective of formal semantics, it is usually more productive to use these terms to talk about relationships between the interpretation and the hierarchical position of the quantifiers in the structure of the sentence as pronounced. I will use them in this linear way here only for simplicity's sake.

scope are sometimes costly (O. Bott and Schlotterbeck, 2015; but cf. Brasoveanu and Dotlačil, 2015). Nevertheless, attempts to demonstrate garden path effects for disambiguation to inverse scope in a following clause have been mixed, with some authors finding partial evidence in button-press paradigms (Tunstall, 1998; Anderson, 2004), but others in subsequent button-press, eyetracking, and ERP studies finding nothing at all (Filik et al., 2004; Paterson et al., 2008; Dwivedi et al., 2010; Dwivedi, 2013). In a recent review, Brasoveanu and Dotlačil (2019) make the argument that garden path effects are present in these cases, consistent with an early selection and interpretation of surface scope (see Dotlačil and Brasoveanu, 2015 and Brasoveanu and Dotlačil, 2015), but it remains unclear why many authors repeatedly fail to observe this effect.[4] Examination using the same stimuli while manipulating task demands could prove fruitful, extending the efforts already demonstrated in Dwivedi (2013).

Possible postponement of the decision here must still be associated with at least consideration of both meanings, much as argued here for scalar implicatures and causal inferences. It is a common observation (Filik et al., 2004; Paterson et al., 2008; Radó & Bott, 2012) that even when comprehenders do not exhibit difficulty with later disambiguation, they experience some slowdown at the introduction of a second quantifier, particularly in conditions where bias towards (linear) surface scope conflicts with some other bias, e.g. for indirect objects to have higher scope than direct objects.

I note that one complication here is that if there is a cost for selection of inverse scope, observing costs in later disambiguation to inverse scope may simply be attributed to this selection cost, rather than evidence of an earlier choice followed by costly revision. My studies here on scalar implicature and causal inference have been able to avoid this confound because they probed the cost of reanalysis away from a subordinate analysis which is preferred in context. Critical evidence for on-line selection would have to come from a study that showed that inverse-scope-biasing contexts increase both early selection costs (on the ambiguous sentence) and reanalysis costs on regions which obligate surface scope.

[4] Some of the nulls can be explained by the observation that singular reference to an indefinite is, strictly-speaking, compatible with wide-scope or narrow-scope readings of the indefinite (e.g. just because the truth conditions allow for every cashier to have greeted a separate customer does not mean that they did not all happen to greet the same one) (Tunstall, 1998). This would explain the failure to find evidence of reanalysis triggered by the singular in Tunstall (1998), Filik et al. (2004), Paterson et al. (2008), Dwivedi et al. (2010), and Dwivedi (2013). Nevertheless, Filik et al. (2004) and Paterson et al. (2008) also fail to replicate any cost for reanalysis triggered by the plural.

5.1.3 Temporal order in discourse

Discourses featuring a sequence of simple past utterances are ambiguous between (at least) an exceedingly common reading of narrative progression, where S2 describes an eventuality which follows the eventuality described in S1 (59a), and an infrequent "backshifted" reading, where S2 describes an eventuality which precedes the eventuality described in S1 (59b). The standard approach in modern semantics, following Partee (1984), is to treat this as a case of ambiguous temporal anaphora. Modern theories of discourse coherence have sought to account for this resolution largely as a byproduct of selecting a discourse relation (Asher & Lascarides, 2003).

(59) **Temporal order ambiguity** (after Dickey, 2001)

This year Anne celebrated her birthday by taking the day off from work.

In the morning she went shopping and went out for lunch downtown.

a. She went to a matinee with her boyfriend (on her way home).

➻ $\tau(S1) < \tau(S2)$

b. She went to a matinee with her boyfriend (on her last birthday).

➻ $\tau(S2) < \tau(S1)$

If decisions about temporal order were made during the interpretation of S2, we would expect a garden-path effect for late disambiguation to the presumably-subordinate backshifting interpretation, as in (59b). However, evidence here is mixed. Dickey (2001) reported a pair of SPR experiments using a movable-adjunct-disambiguation paradigm to probe for exactly this garden-path effect. Observing costs in the spillover following the critical adverbial in all backshifted S2s, but no cost for late disambiguation to a backshifted interpretation, he concluded that backshifted interpretations were costly to select, but there was no garden-path effect. That is, forward progression may be preferred, but it was not selected before the end of S2 in the absence of disambiguation.

In a series of follow-up studies, Sasaki (2021) failed to find any strong evidence for the penalty for backshifting. Instead, in one SPR experiment, Sasaki did find that reading times on late adverbials signalling a progression interpretation were elevated when the lexical content of S2 made a backshifting interpretation highly plausible, perhaps evidence for a garden-path in the other direction driven by plausibility considerations. Nevertheless, a follow-up Maze experiment in turn failed to replicate this or the original Dickey effect.

There is much room for further investigation here, but between Dickey (2001) and Sasaki (2021), present evidence suggests that selection of a particular temporal order is postponed at least until the end of S2. This is not surprising on a coherence-based approach to this ambiguity, given my results in Chapter 4: it would seem that both temporal order and causal inference depend on the evaluation of coherence, a process which does not generally fuel decisive incremental interpretations.

5.1.4 Aspectual ambiguity

As alluded to in Chapter 2, the aspectual interpretation of predicates has received a good amount of attention as an opportunity for incremental ambiguity. By aspect here, I mean internal aspect, what has also been called Aktionsart, the classical classification of a predicate with regards to eventivity, durativity, and boundedness (see e.g. Vendler, 1957; Dowty, 1979; Moens and Steedman, 1988). While many verbs have an apparently typical interpretation, e.g. *jump* as a non-durative description of a single hop (60a), they may yield alternative interpretations in certain contexts, e.g. *jump* as a durative description of a series of hops (60b).

(60) **Aspectual ambiguity** (after Brennan and Pylkkänen, 2008)

The clown jumped...

a. at three o'clock.

b. for three minutes.

A series of early and influential investigations in sentence processing found that cases like (60b) in English, where a modifier after the verb phrase forces a durative interpretation of a typically non-durative predicate, were associated with cognitive difficulty diagnosed by increased latencies for a simultaneous lexical decision (Piñango et al., 1999, 2006) and increased latencies in a Stops-Making-Sense task (Todorova et al., 2000).[5] However, subsequent investigation of this durative coercion effect in English (Pickering et al., 2006) and German (O. Bott, 2010) using self-paced reading and eye-tracking has not always replicated

[5] See also the Dutch ERP study reported in Baggio et al. (2008), where a late adjunct contravening the telicity of a progressive event was associated with a Sustained Anterior Negativity (SAN).

these effects.[6] Pickering et al. (2006) and Pylkkänen and McElree (2006) suggested that the tasks of Piñango, Todorova, and their colleagues may have made it strategic to settle on early specification of what may be a sense distinction of the same minor degree as that found in polysemy, but other work has found such effects with more certainty (Townsend, 2013; see also Husband et al., 2008 and discussion in Stockall et al., 2010). A similar cost has been observed on the predicate itself when durative adverbials were sentence-initial, driving slowdowns in SPR in English and pronounced negativities in the ERP record in English and in Japanese (Brennan & Pylkkänen, 2008; Paczynski et al., 2014; Yano, 2018). As a result, it's somewhat hard to distinguish between a model where the costs on late adverbials are due to reanalysis following an aspectual garden path, or merely the cost of selecting an atypical aspectual interpretation for a given verb.

A series of concerted investigations by O. Bott and colleagues (O. Bott, 2010, 2013; O. Bott & Hamm, 2014; O. Bott & Gattnar, 2015) has examined these various costs across languages and types of coercion. O. Bott (2010), in a monograph systematically comparing various types of aspectual disambiguation in German, observes three key effects (i) a cost for late telic modification of non-telic predicates indexed by increased SPR latencies and a left anterior negativity (LAN) in ERP (see also O. Bott, 2013), (ii) no cost for atelic modification of telic predicates and (iii) a cost for long durative modifiers only when modifying durative telic events (Vendlerian "accomplishments") indexed by increased SPR latencies. While result (i) is taken to merely be a cost of constructing a more complex event structure, result (iii) is of interest, argued to arise from a conflict between an initially-constructed (long) single-event reading and plausibility, leading to reanalysis towards a dispreferred multiple-event reading—this suggests an immediate selection of single-event readings here when possible. A series of later SPR experiments (O. Bott & Hamm, 2014) found that result (ii) is subject to cross-linguistic variation; while German comprehenders appear to freely permit atelic modification of telic events, it is associated with some cost in English attributable to selection of canonical telic readings and late reanalysis. The authors argue that the structure of aspectual marking across the two languages leads to differing strategies: English comprehenders, who experience reliable aspectual cues on the verb in the past tense, select telic meanings at the VP boundary, while German comprehenders,

[6] I included experiments with a reduced item set adapted from Brennan and Pylkkänen (2008) and inspired by the Pickering et al. (2006) design in the fillers of Experiments 3 and 4 from Chapter 2, and failed to find any indication of costs in the relevant condition, in line with Pickering et al. (2006), although power may not have been sufficient for conclusive interpretation.

who lack such useful cues, postpone the interpretation of telicity (see related effects in a comparison between German and Russian by O. Bott and Gattnar, 2015).

It is clear that aspectual interpretation is an area with room for much more rich work. Existing generalizations suggest that in at least some languages, in some tasks, some components of aspectual interpretation come at a cost. These costs may not necessarily be associated with costly reanalysis; further work should attempt to clarify and separate costs associated with selection and costs associated with post-selection reanalysis. Whatever the appropriate explanation is for these costs, cross-linguistic and task comparisons seem to reveal rich strategic differences here. While available evidence is not enough to fully incorporate this variation into the model I lay out in this dissertation, understanding them through the same lens would be an attainable goal for future work.

5.1.5 Modifier attachment

I touch now on an ambiguity of constituency which has received much discussion in the literature on the timing of decisions, ambiguities of modifier attachment. The syntax of English in principle allows the prepositional phrase *with the moustache* in (61) to modify either the constituent *the driver of the car* (in which case the driver has a moustache) or the embedded constituent *the car* (in which case the car has a moustache).[7]

(61) **Modifier attachment ambiguities** (from Traxler et al., 1998)

The driver of the car with the moustache was pretty cool.

The Late Closure heuristic proposed by Frazier (1978) predicts the parser to heuristically build the latter structure, modifying *the car*. This would predict a garden-path effect in cases like (61), where lexical content should force late reanalysis to the alternative parse modifying the larger constituent. However, SPR data from Carreiras and Clifton (1993) on similar ambiguities of relative clause attachment in English revealed no such effect; late disambiguation to the hypothetically dispreferred parse was no more costly than to the hypothetically preferred parse. Frazier and Clifton (1996) used this data to argue that parsing decisions for modifier attachment are atypically made via parallel consideration of possible interpretations (termed "construal"), and not by immediate heuristic construction of a single preferred parse.

[7] Traxler and colleagues intend this to be an implausible meaning, however for at least a certain point of time it was not impossible (though still somewhat remarkable) to see some ride-share cars with moustaches, an observation I owe to Brian Dillon.

Against this backdrop, Traxler et al. (1998) presented a surprising observation from a series of eye-tracking experiments: modifiers disambiguated to either interpretation featured prolonged reading times compared to modifiers which remained globally ambiguous, an effect they termed the "ambiguity advantage." One compelling explanation for these findings was advanced by van Gompel et al. (2000, 2005): *contra* the construal hypothesis, selection for attachment of an ambiguous modifier is made rapidly, but stochastically, as the outcome from a noisy decision process between the two alternatives.[8] Under this hypothesis, the costs for both disambiguated structures can be explained as reanalysis effects after all, deriving from the proportion of trials where the noisy decision process selected a parse which turned out to be unsatisfactory.[9] Under this view, resolution of this ambiguity is much like resolution of an equibiased homonym: all alternatives are momentarily considered, and one is rapidly selected (Duffy et al., 1988).

But the ambiguity advantage effect has one other influential explanation, owed to Swets et al. (2008), who argue that reading ambiguous stimuli is relatively easier because decisions about modifier attachment are regularly postponed in natural reading. They take crucial evidence for this claim from an SPR study where the advantage disappears when comprehenders learn to expect more difficult questions. The apparent cost for disambiguated stimuli in less task-oriented reading is taken to be the cost incurred by optional specification. I note two meaningful objections to their analysis from more recent work. First, Logačev and Vasishth (2016) reported modeling evidence that the variation observed by Swets and colleagues could be attributed to simpler modulations of depth of processing, rather than differences in the timing of selection.[10] Second, in a forthcoming comparison across SPR, the Maze, and eyetracking, Sloggett et al. (2023) failed to replicate consistent task-dependence in line with the predictions of Swets et al. (2008).[11]

[8] They suggest a race between multiple evidence accumulation processes terminating when one process passes a threshold for selection (termed the "Unrestricted Race Model"). It seems to me that in principle, a single-accumulator drift model would work just as well. The principle difference is that the latter would predict slower decisions given the presence of evidence for an alternative, but these models are not aiming to model the effect of evidence that comes from within the modifier; they are operating under the assumption that this parsing decision is carried out before the content of the modifier is observed. Both a race model and a drift model could derive the stochasticity that they require.

[9] Logačev and Vasishth (2016, pp. 269–271) find this version of a race model to be conceptually and empirically less preferable than an alternative which does not require reanalysis, although I am not particularly convinced by their argument.

[10] This is a good benchmark for any claim of task effects on particular processing; in principle the task dependence reported in Chapter 2 could be susceptible to a similar objection.

[11] Relatedly, in a 2021 blog post (https://vasishth-statistics.blogspot.com/2021/08/a-common-mistake-in-psychology-and.html), Shravan Vasishth demonstrated that the critical interaction effect between disambiguation and task demands in the Swets et al. (2008) data does not emerge as

There is therefore little evidence to believe that the ambiguity advantage derives from task-dependent delays in the parsing of modifiers.

To be clear, task effects can still undoubtedly influence comprehension of ambiguous modifiers. Logačev and Vasishth (2016) had German comprehenders read similar sentences in an experiment where they were encouraged to generate and interpret all possible parses for ambiguous modifiers. In this case, reading times were inflated in ambiguous conditions (an "ambiguity disadvantage"), which they take to reflect the cognitive demands of constructing and storing both parses in memory. As in Chapter 2, the demonstration of atypical behavior is a useful indication of the general flexibility of comprehension in the face of strategic demands.

A model based on task-specific delayed selection is not particularly well-supported by the data; nevertheless, two recent judgment studies offer a different, more convincing challenge to rapid stochastic selection. First, Dillon et al. (2019), probing speeded acceptability judgments on late-disambiguated modifiers, demonstrated that participants rated sentences with unambiguous attachment more slowly and less certainly than controls. The authors argue that this is more naturally explained if comprehenders were considering multiple parses at the time of judgment, and referencing all active parses to administer a judgment: judgment was slower and less certain, then, because parses were split in acceptability. Second, Logačev (2023) used a deadline procedure to investigate the availability of competing parses over time for German ambiguously-attaching relative clauses. He observed that the acceptability of ambiguously-attaching modifiers first began to influence judgments somewhat faster than the acceptability of any modifiers with disambiguated attachment. If this is a robust pattern, it is incompatible with a model of early stochastic selection, where the earliest registration of acceptability should be the same in all conditions. It is far more compatible with a parallel model, where, as noted by Dillon et al. (2019), multiple parses with equivalent acceptability could drive faster responses than parses mismatching in acceptability.

These judgment studies provide suggestive evidence for more protracted parallel consideration here, and can explain patterns of response behavior, but need additional assumptions to explain the original reading time asymmetries. Dillon et al. (2019) suggest that one potential explanation would be to expect reanalysis-like behavior in the comprehender whenever one of the multiple parses they are considering must be rejected.

significant when subjected to more stringent analysis.

One similar alternative is that before one parse is selected comprehenders are engaged in expectation-building from all parses under consideration, and thus modifiers that are e.g. lexically consistent with both attachment sites will be maximally facilitated—indeed, this is exactly the argument of Levy (2008, pp. 1153–1157). Additional assumptions would also be necessary to account for the extra costs of encouraging comprehenders to interpret multiple parses, per Logačev and Vasishth (2016); perhaps these are associated with the maintenance and full interpretation of multiple parses rather than the generation of multiple parses as originally suggested by those authors. At the moment, to me, this temporary parallelism, essentially in line with the "construal" proposal from Frazier and Clifton (1996), seems like the best candidate explanation on the market, and would fit nicely within the present framework.

5.1.6 Discourse anaphora

The last case of incremental ambiguity I will discuss here is a perennial research area for sentence processing, the interpretation of discourse anaphora, which is to say, pronouns which are used to flexibly refer to antecedents from the preceding context. In English, where pronouns distinguish only grammatical gender, animacy, and number, whenever there are multiple antecedents which match the features of a pronoun, it is properly ambiguous. E.g. the final *he* in (62) could refer to *Bill* or *John*, and subsequent text may disambiguate in either direction.[12]

(62) **Ambiguous discourse anaphora** (from Gordon and Scearce, 1995)

Bill wanted John to look over some important documents.

He had to mail him the documents.

Unfortunately, he...

a. never sent the papers. (*he* = Bill)

b. never received the papers. (*he* = John)

Psycholinguistic research on the comprehension of pronouns has generally focused on the types of information comprehenders are sensitive to in the resolution of pronominal

[12] As a consequence of discussion in Winograd (1972), the human-like interpretation of these kind of post-hoc disambiguated discourses sentences has been used as an important benchmark for human-like language comprehension (Levesque et al., 2012), one now met for the first time by modern large language models (Kocijan et al., 2023).

ambiguities; see Rohde (2019) for a review. Less work has examined the timecourse of selecting an antecedent, and the potential for garden-path effects, but for a few studies. The most straightforward evidence comes from Gordon and Scearce (1995), using phrasal SPR to examine discourses like (62). In their stimuli, one potential referent was always the matrix subject of the two previous context sentences, a structure in which that referent is very likely to be the antecedent for a following subject pronoun (Gordon et al., 1993). Indeed, comprehenders seem to have selected this interpretation for sentences rapidly, such that continuations consistent with the dispreferred interpretation (62b) were associated with slower latencies in an apparent garden-path effect. The authors also observe that selection of this dispreferred referent is independently costly when disambiguating information precedes the pronoun, but the apparent garden-path costs exceed the baseline difficulty of selecting against the subject bias, evidence that there are additional costs due to reanalysis.

The only other case I am aware of comes from the second experiment reported by Stewart et al. (2007), which used word-by-word SPR, discourses which engendered less of a bias for the pronoun, and later disambiguation, in the middle of the following sentence. Stewart and colleagues observe that late disambiguation to a subject antecedent (63a) and to an object antecedent (63b) were costly to the same degree, compared to relatively faster reading of the same region following an unambiguous pronoun.

(63) **Ambiguous discourse anaphora** (from Stewart et al., 2007)

Paul lent Rick the CD before he left for the holidays. He went to the Bahamas and…

a. sent Rick a postcard from the hotel. (*he* = Paul)

b. sent Paul a postcard from the hotel. (*he* = John)

This is essentially the pattern of Traxler et al. (1998), and affords the same possible interpretations: (i) these costs index costly reanalysis following rapid stochastic selection of an antecedent (a la van Gompel et al., 2000), or (ii) these costs index reduced expectation density as a function of the presence of an ambiguity (a la Levy, 2008). In either case, the absence of asymmetry here need not be contradictory with the clear subject bias observed in Gordon and Scearce (1995); it is likely that the reduced length of the discourse and the predominance of transfer-of-possession verbs in the Stewart et al. (2007) items contributed to balance out preferences for interpretation of the pronoun, although the authors do not report any data on these preferences.

Stewart et al. (2007) argue for a third interpretation of their results, that selection has been postponed here, and the costs observed following ambiguous pronouns are indicators of symmetric costly selection. I am less inclined to adopt this explanation because the Stewart results do not show a corresponding selection cost on the pronoun in unambiguous conditions.[13]

Why might it make sense to argue for postponed selection for discourse anaphora? In a separate but prominent line of work in sentence processing, many studies have observed task-dependent behavior in the comprehension of pronouns related to the activation of alternatives. Initial evidence from cross-modal priming studies suggested that selection of a pronominal antecedent was associated with a downgrade in activation for any potential but unselected antecedents; e.g. Gernsbacher (1989) found evidence that this occurred at sentence offset, whether pronouns were disambiguated by gender or information in context. Nevertheless, this late downgrading of alternative antecedents only occurred when task features encouraged very careful comprehension (Greene et al., 1992; Rigalleau et al., 2004), even when pronouns are unambiguous due to gender information, leading Greene et al. (1992) to argue that pronominal antecedents are generally not selected during incremental comprehension. Relatedly, Stewart et al. (2007) demonstrated in their first experiment that a potentially-related cost arises on ambiguous pronouns when task features encourage careful comprehension, a finding replicated recently in Dutch (Creemers & Meyer, 2022) (although see Grant et al., 2020). In sum, there is evidence for a task-specific costly process related to pronoun interpretation which results in downgraded activation for unselected potential antecedents.

There are two ways I can see to understand this task-specific process in tandem with the clear results of Gordon and Scearce (1995). First, we could take this costly downgrading as a signature of selection itself, in which case we must say that in many of these studies, true selection of an antecedent was not occurring, even in the absence of ambiguity. Gordon and Scearce (1995) would have to be argued to be a special scenario where selection was motivated. This seems possible, but relatively undesirable: intuitively, unambiguous pronouns are interpreted rather quickly during narrative comprehension, as a rule. More concretely, the discourses of Gordon and Scearce (1995) were most like the cases where Greene et al. (1992) failed to observe their downgrading effect, as they featured

[13] The authors in fact argue against a race-like stochastic selection process on the basis that they did not observe this selection cost, but the original van Gompel et al. (2000) Unrestricted Race Model does not predict an ambiguity advantage prior to disambiguation, while a delayed-selection approach indeed would.

questions unrelated to the identity of the pronominal antecedent—why, then, does selection seem apparent for Gordon and Scearce (1995) and not Greene et al. (1992)? In light of this, I adopt the second potential explanation: that the systematic review and downgrading of all other hypothetical antecedents is a secondary strategic comprehension process, somewhat like the one induced for modifier attachment by Logačev and Vasishth (2016).[14] One piece of evidence for dissociating this priming-diagnosed activation differential from typical interpretive properties is that implicit attention to potential referents, as measured by eye movements in visual-world paradigms, accumulates quickly (i.e. before the end of the sentence) and is sensitive to cues like gender and subject-bias (Arnold et al., 2000, *inter alia*). It would seem that rapid consideration of potential meanings is a standard function of pronoun interpretation, and thus that evidence for a late and task-dependent effect from priming must be a consequence of some other process. This way of understanding this body of work predicts that garden-path effects should be observed following pronouns with a strongly preferred antecedent regardless of task variables like comprehension questions, and even in the absence of these priming-based downgrading effects; this prediction should be examined in future work.

5.2 Towards an account of the typology

Taking into account the five cases highlighted in this dissertation, and five further cases as profiled briefly in the previous section, Table 5.1 summarizes the apparent time-course of each decision.

Three groups are apparent by virtue of variation in garden-path costs. In two cases—homonymous nouns and discourse anaphora—garden-path effects are well-attested without known task or language variation, diagnosing a typical pattern of rapid selection. The classic evidence in Frazier (1978), Frazier and Rayner (1982), and subsequent work on syntactic garden paths would hold that the parsing of most constituency ambiguities falls into this same category, although I have not discussed such cases here. At the other end

[14] As reviewed above, activation differentials are taken as evidence for the exhaustive access of potential meanings of a homonym, followed by selection (e.g. Onifer and Swinney, 1981). It may seem inconsistent to claim that lowered activation of competing meanings is merely the consequence of a secondary process for pronoun interpretation. Nevertheless, cross-modal priming studies for pronoun interpretation have used re-presentation of character names as probes for activation, while the probes in homonymy studies are traditionally associates of the target meanings. Unlike associates of a contextually-incorrect meaning, a contextually-incorrect referent for a pronoun begins important and remains important in a narrative, as a salient character. I would argue these activation effects should not be taken to be directly comparable.

Table 5.1: A rough typology of decision-making timecourses across the types of decisions described in this section, with particular attention to the presence of garden path effects, subordinate selection costs, and evidence for rapid consideration of available analyses.

Case	Summary	Garden Paths?	Sub. Sel. Costs?	Rapid Consid.?
Homonymy	Always immediately selected[a]	Y	Y	Y
Discourse anaphora	Immediately selected, task-specific ambiguity costs	Y	Y	Y
Polysemy	Postponed to sentence boundary unless motivated	%	Y	-
Distributivity	Sometimes postponed to at least VP edge unless motivated	%	-[b]	-
Quantifier scope	Sometimes postponed indefinitely unless motivated	%	Y	Y[c]
Aspect	Postponed to VP edge depending on language, sometimes further unless motivated	%	Y	-
Scalar implicature	Always postponed indefinitely, but rapid consideration of enrichment	N	Y	Y
Causal inference	Always postponed indefinitely, but rapid consideration of enrichment	N	-	Y
Temporal order	Possibly postponed indefinitely	N	-[d]	-
Modifier attachment	Postponed to at least modifier edge, equi-biased parses, task-specific ambiguity costs	N	%[e]	Y

Key “Y”: presence observed consistently; “N”: absence observed consistently; “%”: observed to vary; “-”: not enough evidence available

[a] With the exception of homonymous verbs and cross-category homonymy, which may exhibit short-term postponement.
[b] Unexplained costly generation for collective interpretations may outweigh hypothetical difficulty selecting distributive interpretation.
[c] Unlike other cases, which can be understood as adjustments to expectations, evidence here for rapid competition difficulty when cues conflict.
[d] Variation in backshifting costs across two studies remains unexplained.
[e] Variation presumably attributable to variability in the direction and strength of biases.

of the table, four cases—scalar implicature, causal inference, temporal order, and modifier attachment—consistently lack garden-path effects, again without observed variation, diagnosing deferred selection. In the middle, I have profiled another four cases—polysemy, distributivity, quantifier scope, and aspect—where garden-path effects are only sometimes observed, conditioned on differences between task variables like the presence and nature of comprehension questions, or the dynamics of the reading task itself. The presence of garden paths for aspect is also apparently conditioned by language and the exact nature of the aspectual decisions to be made. In these cases, the timing of selection seems to be governed by the particular utility of a selection relative to the comprehender's goals, and the likely position and difficulty of later disambiguation.

Notice that while garden path effects can indeed be seen to vary across the cases I have profiled, the other two classes of notable effects (subordinate selection costs and evidence of rapid consideration) are more constant, either clearly observed, or else inconclusive or unexamined. It would be valuable to probe the existing gaps further, but judging solely from the current evidence, these features may be invariant consequences of how the comprehension mechanism handles ambiguity.

In the rest of this section, I will suggest and explore the consequences of possible explanations of these three features of this observed typology—the variation in the timing of selection, the consistency of costs for subordinate selection, and the consistency of evidence for rapid consideration.

5.2.1 Predicting variation in selection timing

Under the functional approach to selection timing hypothesized in Chapter 2, the comprehension mechanism selects a single analysis of input if the net value of selection is above some threshold; that is, if the relative utility of a decision at that point is high enough compared to the risk it engenders. As I have presented it, one would want this approach to explain the observed variation in garden-path effects in Table 5.1—not only the particular between-experiment variation observed for those cases in the middle of the table, but also why some input receives immediate selection consistently, and why other input does not receive any apparent selection during online comprehension.

To examine how this might be possible, I will go into a bit more detail on the factors which might influence the net value of a decision, and then give the outlook for whether such factors could vary in the predicted ways across the ten constructions under discussion. This cannot be a rigorous test of the approach, as these factors will be numerous and, unfortunately, non-trivial to quantify. However, I hope that this discussion might aid in bringing the proposal closer to a falsifiable theory.

I'll begin by laying out how we could estimate the utility of selection. One way in which selection is useful is that it presumably reduces demand on cognitive resources compared to a deferred decision. The exact demand that deferment imposes is difficult to quantify, but it may be smaller when there is a linguistic representation which can be adopted which is compatible with any of the potential analyses, as is the case for polysemy, by virtue of an available underspecified representation (Frisson, 2009). This is also the case whenever the decision lies between two alternatives where one properly entails the other, like the meanings of *some* or perhaps some aspectual interpretations (see O. Bott, 2010 on "additive coercion"): in such cases, the stronger interpretation may be adopted later without contradicting the former.[15] Like Frazier (1999), I assume it is most difficult to maintain uncertainty over multiple candidate analyses when the continued analysis of the input depends on an analysis of the ambiguous target, although I stress that this cannot be evaluated independently of assumptions about the nature of the linguistic representations used in comprehension.[16] As I noted in Chapter 1, Frazier's division between representational and post-representational decisions can do a lot of work here, correctly (given some assumptions) accounting for at least the possibility of delaying selection for polysemy, scalar implicature, and causal inference. But it cannot straightforwardly account for the rest of the typology, assuming that the comprehender does not have access to underspecified representations for ambiguities like distributivity and quantifier scope.[17]

Another main measure which should contribute to the hypothetical utility of incremental selection is the gain in the quality of expectations obtained by adopting a single interpretation. This comes from the general assumption that comprehension is generally

[15] This does not mean selection of the weaker meaning would be necessarily indistinguishable from deferred selection: in the former case, we could expect the stronger meaning to be explicitly ruled out if selection always involves exhaustive rejection of all alternative analyses.

[16] E.g. must a pronoun's referent be selected in order for the comprehension mechanism to construct a complete semantic interpretation? Not if complete semantic interpretations are thought of just as functions from variable assignments to truth values, but yes, if they are thought of as instructions for updating a mental model.

[17] This is not a necessary assumption: it is possible to imagine a scheme of representation that permits underspecification for any ambiguity (Egg, 2010) and there is a long tradition of doing so for scope ambiguity in particular, see Ebert (2005) for extensive discussion. I have tacitly assumed that the semantic representations we compute during incremental processing generally involve something more like standard Montague semantics, as outlined in e.g. Dowty et al. (1981). One has to adopt a cut-off somewhere, as it is in principle possible to adopt underspecified formalisms for all of the types of uncertainty profiled here.

aided by rapidly-developing expectation for what will come next (see e.g. Ferreira and Chantavarin, 2018 for a recent nuanced review), and that comprehension proceeds most smoothly when input was most expected. The more substantial the shift in expectations due to a decision, the more useful that shift might be to facilitate upcoming processing. The weight of this particular measure of utility presumably should vary depending on the dynamics of a task: concrete expectations are more useful in the Maze, where they help prevent trial failure, than in typical reading, where they merely facilitate integration.

On the other side, the risk of selection should largely be a factor of the cost of selection itself, and the cost and likelihood of potential reanalysis. On the first point, evidence from recent work on sense specification for polysemy (Brocher et al., 2016, 2018) has suggested that selection is not only more difficult given evidence conflict (driving subordinate selection costs), but also given close overlap between alternative meanings. The closer the relationship between the meanings in question, then, the harder selection would be; this could make deferment a more appealing prospect for cases with closely related meanings. As for the second point, although this has not been heavily explored in the realm of incremental semantic decision-making,[18] it is possible that different types of reanalysis could require more or less cognitive resources to execute. This could also perhaps be related to the distinctiveness of the analyses under question, in terms of structural representations or interpretive consequences, although the relationship between distinctness and cost is not obvious. On one view, similar meanings may not require as much revision, making reanalysis a less risky prospect. But informed by the above results for meaning proximity and costly selection, performing the operations necessary for reanalysis may be more difficult for closely related meanings.[19] Whatever the anticipated cost of reanalysis might be, it should then presumably be modulated by the estimated likelihood of reanalysis: as comprehenders are less certain that a single interpretation selected given the current information would be correct, the greater risk reanalysis costs should pose.

I summarize these contributions to utility and risk in (64–65).

(64) **The utility of selection**

a. ... increases as the resource demands of potential deferment increase

[18] See Bader (1998) for discussion of differential reanalysis costs in parsing, which he argues could be related to the need for changes to reconstructed implicit prosody.

[19] This is particularly also the case if we model reanalysis as dependent on cue-based retrieval of material from memory (Van Dyke & Lewis, 2003), which should be subject to similarity-based encoding or retrieval interference.

b. … increases as the resulting change in expectations for upcoming input increases (modulated by the value of expectations in the task)

(65) **The risk of selection**

… increases as the anticipated cost of selection increases

… increases as the anticipated cost of reanalysis increases (modulated by the likelihood that reanalysis will be necessary)

Notice that several values in the rough relationships sketched in (64–65) might be approximated by some measure of the distinctiveness of the potential interpretations. As the interpretations get more distinct, the predictive value of a choice should increase, increasing the utility of selection (64b). At the same time, the expected costs of selection processes should decrease, decreasing the risk of selection (65a). Moreover, distinctness of meaning may also play a hand in estimating a comprehender's ability to distinguish between possible interpretations given context, decreasing the likelihood of reanalysis and thus also decreasing the risk of selection (65b). The wild card is the role of distinctness on the costs of reanalysis: very distinct meanings may entail easier or harder reanalysis, depending on linking assumptions about the reanalysis mechanism (65b). Without a particular proposal for the exact contribution of each of these measures to the overall value calculation, it's hard to be sure, but given these first three points, it seems probable that comprehension decisions between more distinct meanings should overall have higher value (due to their higher utility and perhaps lower risk).

In addition to these components, we must imagine there is some contribution of specific task obligation, such that if a task requires some aspect of the interpretation to be fixed, that aspect must be selected at some point during comprehension. It's not obvious that this should motivate far earlier selection for obligatory interpretative choices (*pace* Stewart et al., 2007; Swets et al., 2008), just that it should force a decision at some point.

Could these kind of estimates of value explain the relative behavior of the ten case studies in Table 5.1? That depends on several pairwise predictions; e.g. the theory requires that homonymy must somehow entail more costly deferment and/or have more highly distinct interpretations, than any of the cases which do not receive uniformly rapid selection (polysemy, distributivity, and so on all the way down to modifier attachment; the same goes for discourse anaphora). Some of these seem to be intuitively plausible: e.g. homonymy over polysemy, implicature and matters of discourse coherence under many of the others. But other comparisons do not seem so intuitive: the extreme value of immediate selection for discourse anaphora compared to the negligible value of immediate selection for modifier attachment, for instance.

One element of nuance here is the timeline for when selection does eventually occur when it is temporarily put off. In the overall approach sketched above, the comprehension mechanism should cease deferment once it accumulates enough evidence to be more certain in a choice, lowering the risk of reanalysis in a way which is separable from the above simplification to meaning distinctness. This evidence accumulation seems to bring the net value over the relevant threshold at different delays for different decisions. From fastest to slowest: aspect and verbal homonymy (and distributivity, perhaps) are at least sometimes delayed just until the verb and its arguments have all been encountered, a delay of a few words; modifier attachment is uniformly delayed at least until the entire modifier has been encountered, a delay of as much as a full relative clause; polysemy is often delayed until the sentence boundary; and quantifier scope, scalar implicature, causal inference, and temporal order seem to frequently or always remain delayed for larger periods of the discourse. These seem to line up reasonably with where important information may fall: verbal meaning in general is heavily dependent on its arguments; the likelihood of various modifier attachments is heavily dependent on the meaning of the modifier; the exact sense of a polyseme is almost always directly recoverable from its immediate context; whereas the remaining four types of meaning plausibly have only more diffuse evidence as to their resolution.

In general, to more exactly predict or explain the timing of selection, we would need a non-circular way of estimating the key contributing factors in (64–65), plus some quantification of subsequent evidence which might be accumulated during deferment. Many of these factors could be estimated using modern distributional and information-theoretic approaches to modeling the distribution of meaning throughout the linguistic signal, and the distances between different meanings: see e.g. Gibson et al. (2019) for a recent review of some prominent applications of these methods. For instance, (64b) should in part be estimable by the conditional entropy for the next material after the ambiguous input given a selected analysis of the input vs. the disjunction of all possible analyses of the input, perhaps averaged across all possible selected analyses.[20] That itself is not

[20] Given my discussion of the pre-selection stage of comprehension here, this is an oversimplification, because different analyses may have more or less weight even before selection.

directly estimateable using typical information-theoretic models which rely on surface forms, but approximations substituting synonymous unambiguous (or less ambiguous) input would be appropriate. Likewise, probing the surprisal of synonymous unambiguous tokens substituted for the ambiguous input, given various preceding contexts, would be a useful approximation for the quality of (at least distributional) evidence among multiple possible analyses. Other parameters cannot be so easily estimated pre-theoretically, and would need to be selected by hand, I presume: e.g. the theoretically-dependent choice for the difficulty of representing the input while deferring interpretation. So far in my work on this topic, I have not yet been able to take the necessary steps to implement a model of this calculus for selection value along these lines, but it should be possible, and it would bring the proposal to the point where it can be directly tested. I hope to move this work in this direction in the future.

Moreover, if the timing of selection is indeed a rational solution as part of a learned procedure for comprehension, we should expect to be able to model it using domain-general models of sequential decision-making like reinforcement learning. That is, I would predict that a simulated agent should reach this pattern of behavior as they try and optimize their performance given experience with outcomes of rapid and delayed selection strategies. Evaluating these predictions on a complete model of discourse comprehension would be a massive undertaking, but perhaps initial progress can be made by constructing representative toy simulations.

5.2.2 Explaining subordinate selection costs and rapid consideration

Even as garden-path effects vary across the constructions in Table 5.1, the presence of subordinate selection costs and evidence for rapid consideration remain the same, at least as far as current data can tell. This too has lessons for a theory of comprehension decision-making, in the first place showing us that it is feasible to treat that process as somehow unified across different kinds of decisions, and moreover highlighting elements of that process that are apparently inevitable.

I take as given that the proper way to treat subordinate selection costs is as a difficult implicit decision where contextual evidence is pushing against some source of bias. It is a truism in the cognitive psychology of decision-making that a decision between two choices should be harder and slower when the net strength of evidence is weak (Luce, 1986), or when evidence runs counter to an active bias (Ratcliff, 1985; Ratcliff &

McKoon, 2008). The foundational metaphor for almost all modern models of continuous decision-making is the accumulation of relevant evidence, either via continuous perception of a stimulus or continuous sampling from memory of a stimulus. These models differ in whether evidence for *n* alternatives is thought to be recorded in a series of *n* accumulators, until one accumulator crosses a threshold (a *race* model, e.g. Rouder et al., 2015) or a single *n*-dimensional accumulator, until the accumulator reaches a threshold in one of the dimensions (a *diffusion* model, e.g. Ratcliff and McKoon, 2008; Ratcliff et al., 2016). They generally model these types of evidence strength and bias effects on response time as differences in the rate of accumulation, or the starting point of accumulation relative to the threshold.

Suggesting that interpretive decisions should be subject to this general truism of difficulty is desirable in its parsimony, and explains why all such decisions could be subject to subordinate selection costs, but does not on its own explain why they always seem to be. I'll highlight for a moment some reasonable alternative patterns I could have expected. Consider the way the interactive-activation model of MacDonald et al. (1994) accounts for garden-path effects: the comprehender's resolution of ambiguity is the result of parallel consideration of the possible interpretations, aiming to select one based on evaluation of several possible sources of evidence. As in the domain-general models mentioned above, when evidence from contextual fit is in conflict with a bias from frequency, the system will take longer to converge on a single high-activation alternative, explaining subordinate selection costs. Garden-path effects are in turn explained as costly re-ranking given new data, after a single alternative receives high activation. If we augmented this type of theory with the ability to, when strategic, permit continued comprehension without waiting for a single high-activation alternative to emerge, we could explain the absence of garden-path effects with e.g. scalar implicature. At the time later disambiguating material is encountered, contextual evidence may not have meaningfully affected activation of the possible interpretations, and so no noticeable re-ranking costs would be evoked. Such a model would also predict the pattern of off-line subordinate selection costs described in Chapter 3: if evidence from context for the possible upper-bound meaning never gets a chance to fully overcome evidence for the heuristically-preferred lower-bound meaning, a final decision process still should take longer to converge on upper-bound meanings.[21]

[21] I note nevertheless that this does not seem incredibly likely, and would possibly require the full freeze of activations over the alternatives to achieve in practice, given that activation from context must generally accumulate very quickly in such models. Indeed, the rapid influence of context was the very point of these

However, by imagining that evidence from context has basically had no chance to accumulate at all in these cases, we expect no effect of a preferred analysis on reading at all in the interim. This fails to match the available evidence for rapid consideration: if you recall, a wide range of reading studies (although not my own) replicate an effect whereby the contextual salience of an upper-bound meaning facilitates reference to the complement this meaning would entail (Breheny et al., 2006; Bergen & Grodner, 2012; Politzer-Ahles & Fiorentino, 2013; S. Lewis, 2013; Hartshorne & Snedeker, 2014).

I basically return to the surprising conclusion that there must be some meaningful separation between the incremental consideration of meaning that is driving expectations for upcoming content, and a process of decisive selection. Moreover, these two processes must be somewhat separated. To demonstrate, I will imagine a model where the selection process could be fed by the same incremental considerations fueling the comprehender's expectations. As already sketched in Chapter 3, facilitation effects could be captured using a system of interactive processing and expectation akin to e.g. Levy (2008), where incoming material can be facilitated based on expectations conditioned on analysis of existing input, which may involve multiple gradiently-considered alternatives given more or less credence based on their own likelihood in context. (Some adjustments, detailed there, would be necessary to derive the particular pattern where cues for higher likelihood of a given alternative analysis do not make content incompatible with that analysis more difficult, a pattern I called a "No Worries If Not" effect.[22]) I have motivated the idea that there must be then a second process of selection; to model this we might turn to a race or diffusion model over alternative analyses which outputs a single, selected analysis. Adjustments over the likelihood of possible analyses as used to generate expectations must be cost-free, in contrast it is only abandonment of a selected analysis that generates garden-path effects. In the cases of deferred selection, at the time of selection, the comprehension

models' innovation in the first place.

[22] I have not discussed "No Worries If Not" against the broader literature on lexical predictions during comprehension. There too, there has been a related debate about whether the strength of predictions for a subset of lexical items modulates the difficulty integrating an item outside of that subset. Evidence here is somewhat split. E.g., the most influential source of evidence for lexical expectations, the N400 signal in the ERP record, is not sensitive to constraint in this way—however, authors have often observed that constraint affects a variety of later positive deflections (Federmeier et al., 2007; Kuperberg et al., 2020; Stone et al., 2023). As I understand this literature, there is no consensus for how to interpret this later deflection. See also evidence from eye-tracking corpora, which has also been taken to support the idea that expectation strength does not penalize other alternatives (Luke & Christianson, 2016). I'm not certain what relationship these findings have with the effects I have focused on here (especially because they generally use cloze probability to measure expectations, which might already systematically flatten the estimates for continuations outside of the most likely), but there is at least some support for this general thesis.

mechanism has already accumulated significant evidence about the likelihood of various analyses, used to generate expectations. It would be reasonable for this evidence to be the starting point for subsequent selection of an analysis, predicting that selection costs for dispreferred meanings should be reduced when selection is deferred, via the mechanisms that developed an preference for a dispreferred analysis during earlier interpretation. But this is demonstrably not the case; again, sticking with the exemplar of scalar implicature, Degen and Tanenhaus (2016) observe in their visual world experiments that comprehenders demonstrate a subordinate selection cost for dispreferred upper-bound meanings for *some* even after developing rapid expectations conditioned on an upper-bound meaning on the same trial.

The consistency of a selection cost thus pushes the present proposal towards a somewhat modular view of the relationship between expectations and decisive comprehension. The processes which develop incremental expectations are rapid, and can exploit the contextual likelihood of various analyses of indeterminate input. But the selection of one analysis for the purposes of interpretation is driven by an independent decision process, which we can treat using a domain-general model of decision-making, sensitive to bias, and drawing on noisy sampling of relevant evidence from a memory representation of the available input. Presumably, when selection is carried out online, the selected meaning may feed back into the incremental expectation algorithm, but critically the levels of credence in that expectation algorithm do not determine the evidence for subsequent selection. Under this view, automatic interactive expectations over possible analyses and possible upcoming content benefit from decisive interpretation, and facilitate further comprehension processes at a low level, but are not themselves part of the mechanism for decisive interpretation.

5.2.3 Comparing with another two-stage model

In the discussion above, I have highlighted again why it seems that we need a model of comprehension with two stages, ending in a single interpretation for ambiguous input. A fully-interactive parallel model has desirable properties for its ability to model the rapid and gradient influence of multiple possible meanings on expectations for upcoming input, but conflates the facilitatory effects of this influence with the costs of expectation revision, whereas I have argued that the former must be possible without the latter. We arrive, then, at a two-stage model of comprehension where the latter stage of complete interpretation may be delayed. As such, the proposal is somewhat similar to another such model which has been influential in research on polysemy, and a few other domains, Frisson and Pickering's Underspecification Model, laid out programmatically in Frisson and Pickering (2001) and Frisson (2009).

The Underspecification Model aims to treat the process of deferred decision-making between senses of a polyseme observed in Frazier and Rayner (1990) and subsequently by e.g. Frisson and Pickering (1999) and Pickering and Frisson (2001). The central claim, following closely with the original Frazier and Rayner proposal, is that initial lexical access of a polysemous word generates a single underspecified meaning, an abstract meaning which is compatible with all possible senses of the word. In a post-lexical process, comprehenders may optionally "home in" on a single sense using information in context; this homing-in process may occur quickly given an informative context, or in the canonical case will be delayed. Such a model predicts that garden-path effects will not be observed until homing in is completed, and crucially, because individual senses are not selected through a process of typical decision among alternatives, purports to predict no subordinate selection costs. Even in cases where comprehenders may be able to rapidly home in on a single sense, asymmetries based on meaning dominance are not expected to affect reading behavior. Indeed, early evidence in Frisson and Pickering (1999) and Pickering and Frisson (2001) found that polysemes lacked both garden paths and subordinate selection costs. Subsequent work has challenged the lack of subordinate selection costs, largely finding that in specifying contexts, specifying a non-dominant sense is associated with slower reading behavior as soon as the polyseme itself (Lowder & Gordon, 2013; Brocher et al., 2016, 2018). This would seem to arbitrate for a model where rapid specification does involve competition among senses, while still allowing that under delay, some non-competitive homing-in may happen gradually.

The proposal has had influence beyond polysemy: underspecification-like proposals for cases of delayed decisions, where a single non-specific representation is formed and later optionally refined, have been explicitly considered and sometimes adopted or rejected in several of the other literatures discussed here, e.g. Radó and Bott (2012) for quantifier scope ambiguity, Pickering et al. (2006) for aspectual ambiguity, Swets et al. (2008) and Logačev and Vasishth (2016) for modifier attachment, etc. Some of these cases (e.g. Swets et al., 2008) exemplify a connection between Underspecification-like models and the hypothesis of Good Enough Processing (Ferreira & Patson, 2007), which, like the present work, tries to reckon with the influence of task demands and comprehender goals on comprehension behavior. Like the Underspecification Model, Good Enough Processing holds that in certain scenarios, comprehenders will adopt non-specific representations, and that these representations allow comprehenders to avoid effortful decision-making. Crucially the latter type of account differs by allowing non-specificity even when the grammar does not afford a single underspecified or otherwise non-specific-but-coherent representation.

The proposal in this dissertation differs from any model that allows for this kind of single incomplete representation, specifically in the claim that alternative analyses are generated and rapidly available even when comprehenders are strategically avoiding a decision. I have not presented evidence for this fact in the case of polysemy,[23] but I have discussed evidence that possible interpretations have been generated and are under active consideration for both the interpretation of *some* in Chapter 3, and the derivation of causal inference in Chapter 4, even as final decisions have been apparently deferred. As reviewed in Table 5.1, there is also some evidence for rapid consideration before selection for quantifier scope and modifier attachment, and I am ultimately not aware of any studies which have noted evidence for the absence of such effects in any of the ten constructions under discussion. Effects of expectation founded on consideration of possible meanings strike me as excellent evidence against an Underspecification-like account, as these accounts predict that when comprehenders delay selection of a specific meaning, they have not engaged in the generation of possible meanings at all.

I note that this allows us to treat selection among meanings for polysemes as the same type of comprehension procedure as selection among meanings for homonyms, i.e. there is no need to posit an alternative mechanism for homing in on a specified meaning without competition. Regardless of whether selection happens immediately or at a delay, the procedure is ultimately the same sort of implicit decision among alternatives.

A strong claim available to me would be that comprehenders never adopt underspecified representations of this kind, and always rapidly generate alternative meanings to decide between. This is not, to my knowledge, untenable: e.g. Frisson (2009) offered that the best argument for underspecification rather than parallel consideration was simply that meaning uncertainty is frequent, and maintenance of all possible analyses in parallel

[23] The right evidence would be, in neutral contexts, evidence of exhaustive access of senses using a priming paradigm, or evidence that reflects rapid expectations before the sentence boundary, perhaps from ERPs or visual world eye-tracking. In this domain, I only know of Chang et al. (2015), who provide cross-modal priming evidence for exhaustive access with at least one kind of polysemy in Mandarin Chinese.

could be incredibly costly. I agree that that leaving a multiplicity of interpretations unresolved may require substantial resources. But I have treated that pressure as a component in a larger calculus regarding cost, where other benefits may conspire to make that behavior more rational. Still, I do not rule out that underspecification may be an option available to the comprehender in certain cases, at least as a temporary stopping point that allows for some partially determinate representation during consideration among possible specified meanings. Future work might make progress here by testing whether there is any evidence for underspecification as a strategy to avoid generating meanings, or whether we can hold onto a model that always generates possible meanings rapidly. The latter is more parsimonious, but the present model could accommodate the former by allowing the timing of generation to vary, as well as the timing of selection.

5.2.4 On the roles of domain-general cognition and grammar

In the argumentation in this dissertation, I have made frequent recourse to the idea that comprehension behaviors may have their source in domain-general cognitive mechanisms. I have suggested that the final stage of ambiguity resolution should be modeled using standard models of explicit forced-choice selection from alternatives; I have suggested that preferences of decision timing should be explained by standard pressures for rational sequential behavior. I adopt these positions as a starting point, to be preferred instead of a language-specific explanation, at least for theoretical parsimony: there is good evidence for these models in other components of human behavior, and it is reasonable to assume that they could apply in language processing in the same ways. But it is an empirical question whether such models suffice to explain the necessary facts.

There are two kinds of evidence that I think are relevant here. First, given an explicit model relying on domain-general components (of the kind sketched but not fully developed here), evidence that the model cannot generate the right predictions litigates for more language-specific modifications and pressures. But we can also marshal evidence in favor of a domain-general model without relying on parsimony, if we can observe that general cognitive variables affect language processing in the same way as they do other tasks. This is a common argument used to pick out shared resources among cognitive faculties, although it relies on experimental designs that elicit valid and reliable measures of individual differences on multiple tasks. Often, familiar cognitive tasks are not well designed to permit such measurements (Hedge et al., 2018), but certain strategic variables, like general response caution, have been shown to have a reliable effect across tasks for the same individuals (Hedge et al., 2019, 2022). If there is individual variance in how comprehenders negotiate the calculus for selection timing, and that calculus is indeed a domain-general phenomenon related to caution, we should expect it to correlate with variables like response caution across other tasks. Evidence for this kind of correlation would be valuable positive evidence to retain a domain-general account of selection timing.

By making an argument that there is a domain-general source for certain aspects of how comprehenders time and execute selection processes during incremental comprehension, I do not mean to pretend that the peculiarities of language as a cognitive object have no importance here.[24] Certainly, the relationship between the input and the possible outputs of language comprehension is determined by the language faculty. Moreover, speaking particularly about the problem of decision timing now, I take as standard the idea that the comprehension mechanism generalizes preferred behavior across categories of stimulus determined by linguistic knowledge. If, for instance, there is a systematic distinction between how comprehenders time selection for nominal homonyms and verbal homonyms (e.g. compare Frazier and Rayner (1990) and Pickering and Frisson (2001)), I do not imagine that this pattern emerges entirely from experience with individual nouns and verbs, but rather that the abstract categories learned as part of the grammar help determine pathways of generalization. In short, I assume that the symbols and abstract representations particular to language are an inextricable component of how we process language. The stance I have taken throughout this dissertation, however, is one of curiosity about how theories of sentence processing might model general pressures on the way we execute that particular task.

As a result of both a commitment to the cognitive reality of linguistic representations and a theoretical stance favoring domain-general explanations, there are a few places here where the border between language-specific and domain-general explanations is of interest. The relevance of a division between strictly grammatical and post-grammatical comprehension decisions is one such place. As I note, the original firm Frazier (1999) split between grammatical decisions made immediately and deferred post-grammatical decisions can cover a lot of ground, and specifically does predict the delay of pragmatic inferences noted in Chapters 3 and 4. But we also have seen evidence that uncertainties related to grammatical representation can be delayed (distributivity, certain homonyms, quanti-

[24] I thank Lyn Frazier for encouraging me to make this point more clearly.

fier scope, modifier attachment...). I do not think that the split, such that decisions which require a distinction of grammatical representation are more likely to occur more rapidly, is an accident. However, given the variation observed among those grammatical decisions, the grammatical status of the decision cannot be treated as an unmitigated determinant of behavior. Here, I have tried to capture how the representational consequences of an ambiguity matter by approaching them from a functional perspective. I think there is a broader lesson here, representative of how one can approach the relevance of grammar while connecting comprehension processes to other cognitive behavior: limits imposed by our linguistic theories matter for psycholinguistic performance, but they may take their effect on behavior indirectly.

5.3 Limitations and directions for future work

I will conclude by highlighting the limitations of this work, and the future directions that I think can help redress those limitations, and make further progress building from the foundation I have suggested here.

First, I want to acknowledge the status of the evidence I build from here. I have attempted to draw on a wide variety of work to develop a general characterization of how comprehenders tend to solve the functional problem of incremental comprehension of uncertain input. Nevertheless, my novel contributions are small, limited in scope, and not always consistent with that previous work. There are several critical effects that the picture I present rests heavily on, and they should be further evaluated carefully across different methods and designs. In particular, this includes the novel effects I report here, the task-specificity of garden paths for polysemy from Chapter 2, and the lack of garden paths for scalar implicature and causal inference in Chapters 3 and 4. This also includes predicted effects that I either did not test or tested but failed to find: facilitation for complement reference due to possible upper-bound meaning for *some*, and offline selection costs for causal inferences. Although I maintain that Rapid Consideration Without Selection is a convincing account of both scalar implicature and causal inference, those facts require further investigation to make a more compelling case. Finally, in the last sections (§5.1-5.2), I have moved towards a broader typology building on constructions which I have not myself investigated here, and where previous studies may not be directly comparable. If the three effects of interest here (garden-path effects, subordinate selection effects, and rapid consideration effects) were pursued more systematically across these constructions, it would provide better empirical support for the conclusions made tentatively in Table 5.1. As it stands, I can argue that it is possible to generalize over these constructions, but in the process I have made many assumptions and predictions that should be further tested—I have tried to flag these where possible in the discussion above.

In addition to the ways the empirical picture merits further development, the model of comprehension I propose has been advanced merely as a series of computational intuitions. Many contemporary models of language comprehension, including those which I purport to argue against here, have found great success in generating explicit, testable predictions through computational cognitive modeling: e.g. ACT-R based models for memory operations during parsing descending from R. L. Lewis and Vasishth (2005) and information-theoretic approaches to incremental expectations descending from Hale (2006) and Levy (2008); see also Bayesian models for the likely outcomes of pragmatic computation descending from Goodman and Frank (2016). Several elements of my proposal here are modelable, at least in some simplified form: the interactive expectation mechanism with variable resources sketched in §3.2, the generalized account of selection costs suggested in §5.2.2, and perhaps most importantly, the value-based approach to selection timing summarized in §5.2.1. Moving towards more explicit models in each of these areas would allow more precise statement of the claims, and mechanisms for extracting and evaluating predictions that could drive future work.

I would also like to point out a few places where my proposal here has consequences or predictions that lie outside of the empirical scope of the dissertation. The way I have presented the proposal, as an account of decision-making in comprehension, should extend not only to the somewhat niche areas of uncertain meaning profiled here, but also the heavily-researched phenomenon of incremental syntactic ambiguity, which has only been touched on here briefly. I cannot defend or even consider the predictions here in nearly as much detail as is necessary, and they almost certainly would be shown to be overly simple if I did. But I will note prospectively that the idea of value-based timing of comprehension decisions may be a useful tool for accounting for cases where syntactic comprehension appears to be less incremental in certain languages. For instance, eye movement data investigated by Duff et al. (in preparation) is compatible with the hypothesis that comprehenders in a language with a fully productive suite of non-intrusive resumptive pronouns for filler-gap dependencies (Santiago Laxopa Zapotec) systematically delay predictions about the base position of a relativized argument, avoiding the rapid predictive associations thought to underlie garden-path-like effects in English relative clauses (Wagers & Pendleton, 2015). This could be interpreted as rational delay given expectation for a later high-value cue (the possible resumptive pronoun).

Another area where the proposal has implications that I have not considered systematically is the development of language comprehension behavior. I have framed the value-based calculation of the timing of comprehension decisions as a way to derive key behavioral generalizations from experience. But this makes testable predictions about how humans actually acquire such behavior, imagining that they move through stages of overly rapid or strikingly delayed decision-making on their way to the rational balance apparently struck by adults. The acquisition of adult-like mechanisms for processing ambiguous input has been an object of much study in the past few decades, although it has been absent from my own discussion so far. The major discoveries, as outlined in helpful reviews from Snedeker (2013) and Omaki et al. (2015), are two-fold (and both anticipated in the seminal study by Trueswell et al., 1999 that raised these questions to prominence): children rapidly consider potential interpretations of ambiguous input using a subset of the cues exploited by adults, and have trouble executing rapid reanalysis. Recent evidence suggests that apparently-ignored cues may be used by children given more time, and ties reanalysis difficulties to developing domain-general cognitive control mechanisms (Qi et al., 2020). These facts are remarkably consistent across the types of ambiguity that have been subjected to the most study so far, including not only syntactic ambiguities (Trueswell et al., 1999, see also e.g. Yacovone et al., 2020), but also pronominal ambiguities (sources reviewed in Snedeker, 2013), quantifier scope ambiguities (sources reviewed in Lidz, 2018), and scalar implicature (e.g. Y. T. Huang and Snedeker, 2009b).[25] In all of these areas, children behave somewhat like adults under high cognitive load, preferring to pick and stick with less context-sensitive interpretations which are often adults' first choice (Omaki et al., 2015). One place where kids have been found to be more surprisingly adult-like is a small literature on homonymy (Rabagliati et al., 2013; Hahn et al., 2015), where interpretations are largely sensitive to preceding context.

Despite the growing amount of work on incremental consideration in this literature, not so much is known about incremental decisions and when they are (or are

[25] The finding that children exhibit fewer upper-bound interpretations of weak scalars is far from unique, offline or online, but is subject to much debate, see Horowitz et al. (2018) for a recent discussion of the lay of the land.

not) made; this is the domain where I make the most predictions. Of course, key evidence for firm decision-making has generally come here from reading time measures, which are ill-suited to the job here: even in investigations with late school-age children, lower-level deficits in the skill of reading are a problematic confound that is hard to avoid (Snedeker, 2013). Perhaps judicious use of paradigms like self-paced listening (Kidd & Bavin, 2007) or pupillometry (Schmidtke, 2018) could offer a better opportunity to measure incremental difficulties due to selection followed by costly reanalysis.

It is my hope that the proposal here, tentative as it must be, will fuel further study of decision-making during incremental comprehension as a generalized phenomenon. There is a laundry list of empirical work that must be done to develop more specific proposals, and in future work any generalized proposal must be further specified in order to make testable predictions. However, it seems to me that the pursuit of a more unified perspective on these decisions would leave our understanding of incremental comprehension richer, specifically by leaving us with a better framework to investigate differences between types of input, diverse languages, and stages of comprehension expertise.

Bibliography

Anderson, C. (2004). *The structure and real-time comprehension of quantifier scope ambiguity* [Dissertation]. Northwestern University.

Arnold, J. E., Eisenband, J. G., Brown-Schmidt, S., & Trueswell, J. C. (2000). The rapid use of gender information: Evidence of the time course of pronoun resolution from eyetracking. *Cognition*, *76*, B13–B26.

Asher, N., & Lascarides, A. (2003). *Logics of conversation*. Cambridge University Press.

Asr, F. T., & Demberg, V. (2012). Implicitness of discourse relations. In *Proceedings of COLING 24* (pp. 2669–2684).

Bader, M. (1998). Prosodic influences on reading syntactically ambiguous sentences. In J. D. Fodor & F. Ferreira (Eds.), *Reanalysis in sentence processing* (pp. 1–46). Kluwer.

Baggio, G., van Lambalgen, M., & Hagoort, P. (2008). Computing and recomputing discourse models: An ERP study. *Journal of Memory and Language*, *59*, 36–53. https://doi.org/10.1016/j.jml.2008.02.005

Barbet, C., & Thierry, G. (2018). When *some* triggers a scalar inference out of the blue: An electrophysiological study of a Stroop-like conflict elicited by single words. *Cognition*, *177*, 58–68. https://doi.org/10.1016/j.cognition.2018.03.013

Baron-Cohen, S., Wheelright, S., Skinner, R., Martin, J., & Clubley, E. (2001). The Autism-Spectrum Quotient (AQ): Evidence from Asperger Syndrome/High-Functioning Autism, males and females, scientists and mathematicians. *Journal of Autism and Developmental Disorders*, *31*, 5–17.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*, 1–48.

Beckers, S. (2016). *Actual causation: Definition and principles* [Dissertation]. KU Leuven.

Bergen, L., & Grodner, D. J. (2012). Speaker knowledge influences the comprehension of pragmatic inferences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(5), 1450–1460.

Bethell-Fox, C. E., & Shepard, R. N. (1988). Mental rotation: Effects of stimulus complexity and familiarity. *Journal of Experimental Psychology: Human Perception and Performance*, *14*(1), 12–23. https://doi.org/10.1037/0096-1523.14.1.12

Binder, K. S., & Rayner, K. (1998). Contextual strength does not modulate the subordinate bias effect: Evidence from eye fixations and self-paced reading. *Psychonomic Bulletin & Review*, *5*(2), 271–276. https://doi.org/10.3758/BF03212950

Black, J. B., & Bern, H. (1981). Causal coherence and memory for events in narratives. *Journal of Verbal Learning and Verbal Behavior*, *20*, 267–275.

Boland, J. E., Tanenhaus, M. K., Garnsey, S., & Carlson, G. N. (1995). Verb argument structure in parsing and interpretation: Evidence from wh-questions. *Journal of Memory and Language*, *34*, 774–806. https://doi.org/10.1006/jmla.1995.1034

Bott, L., Bailey, T. M., & Grodner, D. (2012). Distinguishing speed from accuracy in scalar implicatures. *Journal of Memory and Language*, *66*, 123–142. https://doi.org/10.1016/j.jml.2011.09.005

Bott, L., & Noveck, I. A. (2004). Some utterances are underinformative: The onset and time course of scalar inferences. *Journal of Memory and Language*, *51*, 437–457. https://doi.org/10.1016/j.jml.2004.05.006

Bott, L., Rees, A., & Frisson, S. (2016). The time course of familiar metonymy. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*(7), 1160–1170. https://doi.org/10.1037/xlm0000218

Bott, O. (2010). *The processing of events*. John Benjamins. https://doi.org/10.1075/la.162

Bott, O. (2013). The processing domain of aspectual interpretation. In *Studies in the composition and decomposition of event predicates* (pp. 195–229). Springer Netherlands. https://doi.org/10.1007/978-94-007-5983-1_8

Bott, O., & Gattnar, A. (2015). The cross-linguistic processing of aspect: An eyetracking study on the time course of aspectual interpretation in Russian and German. *Language, Cognition and Neuroscience*, *30*(7), 877–898. https://doi.org/10.1080/23273798.2015.1029499

Bott, O., & Hamm, F. (2014). Cross-linguistic variation in the processing of aspect. In *Studies in theoretical psycholinguistics* (pp. 83–109). Springer International Publishing. https://doi.org/10.1007/978-3-319-05675-3_4

Bott, O., & Schlotterbeck, F. (2015). The processing domain of scope interaction. *Journal of Semantics*, *32*(1), 39–92. https://doi.org/10.1093/jos/fft015

Boyce, V., Futrell, R., & Levy, R. P. (2020). Maze Made Easy: Better and easier measurement of incremental processing difficulty. *Journal of Memory and Language*, *111*. https://doi.org/10.1016/j.jml.2019.104082

Boyce, V., & Levy, R. (2023). A-maze of Natural Stories: Comprehension and surprisal in the Maze task. *Glossa Psycholinguistics*, *2*(1). https://doi.org/10.5070/G6011190

Brasoveanu, A. (2013). Modified numerals as post-suppositions. *Journal of Semantics*, *30*, 155–209. https://doi.org/10.1093/jos/ffs003

Brasoveanu, A., & Dotlačil, J. (2015). Sentence-internal *same* and its quantificational licensors: A new window into the processing of inverse scope. *Semantics & Pragmatics*, *8*(1).

Brasoveanu, A., & Dotlačil, J. (2019). Quantification. In C. Cummins & N. Katsos (Eds.), *The Oxford handbook of experimental semantics and pragmatics* (pp. 228–245). Oxford University Press.

Breheny, R. (2019). Scalar implicatures. In C. Cummins & N. Katsos (Eds.), *The Oxford handbook of experimental semantics and pragmatics* (pp. 39–61). Oxford University PRess.

Breheny, R., Ferguson, H. J., & Katsos, N. (2013). Investigating the timecourse of accessing conversational implicatures during incremental sentence interpretation. *Language and Cognitive Processes*, *28*(4), 443–467.

Breheny, R., Katsos, N., & Williams, J. (2006). Are generalised scalar implicatures calculated by default? An on-line investigation into the role of context in generating pragmatic inferences. *Cognition*, *100*, 434–463. https://doi.org/10.1016/j.cognition.2005.07.003

Brennan, J., & Pylkkänen, L. (2008). Processing events: Behavioral and neuromagnetic correlates. *Brain & Language*, *106*, 132–143. https://doi.org/10.1016/j.bandl.2008.04.003

Brocher, A., Foraker, S., & Koenig, J.-P. (2016). Processing of irregular polysemes in sentence reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*(11), 1798–1813. https://doi.org/10.1037/xlm0000271

Brocher, A., Koenig, J.-P., Mauner, G., & Foraker, S. (2018). About sharing and commitment: The retrieval of biased and balanced irregular polysemes. *Language, Cognition and Neuroscience*, *33*(4), 443–466. https://doi.org/10.1080/23273798.2017.1381748

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28. https://doi.org/10.18637/jss.v080.i01

Bürkner, P.-C. (2018). Advanced Bayesian multilevel modeling with the R package brms. *The R Journal*, *10*, 395–411. https://doi.org/10.32614/rj-2018-017

Burton-Roberts, N. (1999). Language, linear precedence, and parentheticals. In P. Collins & D. Lee (Eds.), *The clause in English: In honor of Rodney Huddleston* (pp. 33–52). John Benjamins. https://doi.org/10.1075/slcs.45.05bur

Carreiras, M., & Clifton, C., Jr. (1993). Relative clause interpretation preferences in Spanish and English. *Language and Speech*, *36*(4), 353–372. https://doi.org/10.1177/002383099303600401

Chacón, D. A., Momma, S., & Phillips, C. (2016). Linguistic representations and memory architectures: The devil is in the details. *Behavioral and Brain Sciences*, *39*, e68. https://doi.org/https://doi.org/10.1017/S0140525X15000746

Champollion, L. (2020). Distributivity, collectivity, and cumulativity. In D. Gutzmann, L. Matthewson, C. Meier, H. Rullman, & T. E. Zimmerman (Eds.), *The Wiley Blackwell companion to semantics*. Wiley. https://doi.org/10.1002/9781118788516.sem021

Chang, Y., Lin, C.-J. C., & Ahrens, K. (2015). Conventionalization of lexical meanings and the role of metaphoricity: Processing of metaphorical polysemy using a cross-modal lexical priming task. *Language & Linguistics*, *16*(4), 587–614. https://doi.org/10.1177/1606822X15583240

Chemla, E., & Bott, L. (2014). Processing inferences at the semantics/pragmatics frontier: Disjunctions and *free choice*. *Cognition*, *130*, 380–396. https://doi.org/10.1016/j.cognition.2013.11.013

Chemla, E., & Singh, R. (2014a). Remarks on the experimental turn in the study of scalar implicature, part I. *Language and Linguistics Compass*, *8/9*, 373–386.

Chemla, E., & Singh, R. (2014b). Remarks on the experimental turn in the study of scalar implicature, part II. *Language and Linguistics Compass*, *8/9*, 387–399.

Chierchia, G. (2004). Scalar implicatures, polarity phenomena, and the syntax-pragmatics interface. In A. Belletti (Ed.), *Structures and beyond* (pp. 39–103). Oxford University PRess.

Chierchia, G., Crain, S., Guasti, M. T., Gualmini, A., & Meroni, L. (2001). The acquisition of disjunction: Evidence for a grammatical view of scalar implicatures. In *Proceedings of BUCLD 25* (pp. 157–168).

Christiansen, M. H., & Chater, N. (2015). The Now-or-Never bottleneck: A fundamental constraint on language. *Behavioral and Brain Sciences*, *39*, e62. https://doi.org/10.1017/S0140525X1500031X

Ciardelli, I., Zhang, L., & Champollion, L. (2018). Two switches in the theory of counterfactuals. *Linguistics and Philosophy*, *41*, 577–621. https://doi.org/10.1007/s10988-018-9232-4

Clark, H. H., & Chase, W. G. (1972). On the process of comparing sentences against pictures. *Cognitive Psychology*, *3*, 472–517. https://doi.org/10.1016/0010-0285(72)90019-9

Cohen, J., & Kehler, A. (2021). Conversational eliciture. *Philosophers' Imprint*, *21*(12).

Cozijn, R. (2000). *Integration and inference in understanding causal sentences* [Dissertation]. Tilburg University.

Creemers, A., & Meyer, A. S. (2022). The processing of ambiguous pronominal reference is sensitive to depth of processing. *Glossa Psycholinguistics*, *1*(1), 3. https://doi.org/10.5070/G601166

Cutler, A., Mehler, J., Norris, D., & Segui, J. (1986). The syllable's differing role in the segmentation of French and English. *Journal of Memory and Language*, *25*, 385–400.

Cutler, A., & Norris, D. (1988). The role of strong syllables in segmentation for lexical access. *Journal of Experimental Psychology: Human Perception and Performance*, *14*(1), 113–121.

Davies, M. (2008). The Corpus of Contemporary American English (COCA). https://www.english-corpora.org/coca/

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. https://doi.org/10.1027/1618-3169.54.2.128

Degen, J. (2015). Investigating the distribution of *some* (but not *all*) implicatures using corpora and web-based methods. *Semantics & Pragmatics*, *8*, 11. https://doi.org/10.3765/sp.8.11

Degen, J., & Tanenhaus, M. K. (2016). Availability of alternatives and the processing of scalar implicatures: A visual world eye-tracking study. *Cognitive Science*, *40*, 172–201.

Delogu, F., Crocker, M. W., & Drenhaus, H. (2017). Teasing apart coercion and surprisal: Evidence from eye-movements and ERPs. *Cognition*, *161*, 46–59. https://doi.org/10.1016/j.cognition.2016.12.017

Dickey, M. W. (2001). *The processing of tense*. Kluwer Academic Publishers.

Dieussaert, K., Verkerk, S., Gillard, E., & Schaeken, W. (2011). Some effort for some: Further evidence that scalar implicatures are effortful. *Quarterly Journal of Experimental Psychology*, *64*(12), 2352–2367. https://doi.org/10.1080/17470218.2011.588799

Dillon, B., Andrews, C., Rotello, C. M., & Wagers, M. (2019). A new argument for co-active parses during language comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *45*(7), 1271–1286.

Dopkins, S., Morris, R. K., & Rayner, K. (1992). Lexical ambiguity and eye fixations in reading: A test of competing models of lexical ambiguity resolution. *Journal of Memory and Language*, *31*, 461–476. https://doi.org/10.1016/0749-596X(92)90023-Q

Dotlačil, J., & Brasoveanu, A. (2015). The manner and time course of updating quantifier scope representations in discourse. *Language, Cognition and Neuroscience*, *30*(3), 305–323. https://doi.org/10.1080/23273798.2014.918631

Dotlačil, J., & Brasoveanu, A. (2021). The representation and processing of distributivity and collectivity: Ambiguity vs. underspecification. *Glossa*, *6*(1), 1–22. https://doi.org/10.5334/gjgl.1131

Dowty, D. (1979). *Word meaning and Montague Grammar*. Reidel.

Dowty, D., Wall, R., & Peters, S. (1981). *Introduction to Montague semantics*. Springer.

Drummond, A. (2010). Ibex. https://github.com/addrummond/ibex

Duff, J., Brasoveanu, A., & Rysling, A. (2023). *Task demands supercede minimal effort heuristics in incremental comprehension: Polysemy and distributivity in the Maze* [Manuscript, UC Santa Cruz].

Duff, J., Gomez-Jackson, D., Silva Robles, F., Toosarvandani, M., & Wagers, M. W. (in preparation). *Relative clause parsing in a language with optional subject resumption* [Manuscript, UC Santa Cruz].

Duffy, S. A., Kambe, G., & Rayner, K. (2001). The effect of prior disambiguating context on the comprehension of ambiguous words: Evidence from eye movements. In D. S. Gorfein (Ed.), *On the consequences of meaning selection: Perspectives on resolving lexical ambiguity* (pp. 27–43). APA. https://doi.org/10.1037/10459-002

Duffy, S. A., Morris, R. K., & Rayner, K. (1988). Lexical ambiguity and fixation times in reading. *Journal of Memory and Language*, *27*, 429–446. https://doi.org/10.1016/0749-596X(88)90066-6

Dwivedi, V. D. (2013). Interpreting quantifier scope ambiguity: Evidence of heuristic first, algorithmic second processing. *Plos One*, *8*(11), e81461. https://doi.org/10.1371/journal.pone.0081461

Dwivedi, V. D., Phillips, N. A., Einagel, S., & Baum, S. R. (2010). The neural underpinnings of semantic ambiguity and anaphora. *Brain Research*, *1311*, 93–109. https://doi.org/10.1016/j.brainres.2009.09.102

Ebert, C. (2005). *Formal investigations of underspecified representations* [Dissertation]. King's College London.

Egg, M. (2010). Semantic underspecification. *Language and Linguistics Compass*, *4*(3), 166–181.

Erlich, 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. https://doi.org/10.1016/S0022-5371(81)90220-6

Federmeier, K. D., Wlotko, E. W., De Ochoa-Dewald, E., & Kutas, M. (2007). Multiple effects of sentential constraint on word processing. *Brain Research*, *1146*, 75–84.

Ferreira, F., & Chantavarin, S. (2018). Integration and prediction in language processing: A synthesis of old and new. *Current Directions in Psychological Science*, *27*(6), 443–448.

Ferreira, F., & Henderson, J. M. (1990). Use of verb expectation in syntactic parsing: Evidence from eye movements and word-by-word self-paced reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*(4), 555–568. https://doi.org/10.1037/0278-7393.16.4.555

Ferreira, F., & Patson, N. D. (2007). The 'good enough' approach to language comprehension. *Language and Linguistics Compass*, *1*(1-2), 71–83. https://doi.org/10.1111/j.1749-818X.2007.00007.x

Filik, R., Paterson, K. B., & Liversedge, S. P. (2004). Processing doubly quantified sentences: Evidence from eye movements. *Psychonomic Bulletin & Review*, *11*(5), 953–959.

Fishbein, J., & Harris, J. A. (2014). Making sense of Kafka: Structural biases induce early sense commitment for metonyms. *Journal of Memory and Language*, *76*, 94–112. https://doi.org/10.1016/j.jml.2014.06.005

Folk, J. R., & Morris, R. K. (1995). Multiple lexical codes in reading: Evidence from eye movements, naming time, and oral reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(6), 1412–1429. https://doi.org/10.1037/0278-7393.21.6.1412

Folk, J. R., & Morris, R. K. (2003). Effects of syntactic category assignment on lexical ambiguity resolution in reading: An eye movement analysis. *Memory & Cognition*, *31*(1), 87–99.

Foraker, S., & Murphy, G. L. (2012). Polysemy in sentence comprehension: Effects of meaning dominance. *Journal of Memory and Language*, *67*, 407–425. https://doi.org/10.1016/j.jml.2012.07.010

Forster, K. I., Guerrera, C., & Elliot, L. (2009). The maze task: Measuring forced incremental sentence processing time. *Behavior Research Methods*, *41*(1), 163–171. https://doi.org/10.3758/BRM.41.1.163

Fox, D. (2007). Free choice and the theory of scalar implicatures. In U. Sauerland & P. Stateva (Eds.), *Presupposition and implicature in compositional semantics* (pp. 71–120). Palgrave Macmillan. https://doi.org/10.1057/9780230210752_4

Frazier, L. (1978). *On comprehending sentences: Syntactic parsing strategies* [Dissertation]. University of Connecticut.

Frazier, L. (1999). *On sentence interpretation*. Springer. https://doi.org/10.1007/978-94-011-4599-2

Frazier, L., & Clifton, C., Jr. (1996). *Construal*. MIT Press.

Frazier, L., Pacht, J. M., & Rayner, K. (1999). Taking on semantic commitments, II: Collective versus distributive readings. *Cognition*, *70*, 87–104. https://doi.org/10.1016/s0010-0277(99)00002-5

Frazier, L., & Rayner, K. (1982). Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. *Cognitive Psychology*, *14*(2), 178–210. https://doi.org/10.1016/0010-0285(82)90008-1

Frazier, L., & Rayner, K. (1987). Resolution of syntactic category ambiguities: Eye movements in parsing lexically ambiguous sentences. *Journal of Memory and Language*, *26*, 505–526. https://doi.org/10.1016/0749-596X(87)90137-9

Frazier, L., & Rayner, K. (1990). Taking on semantic commitments: Processing multiple meanings vs. multiple senses. *Journal of Memory and Language*, *29*, 181–200. https://doi.org/10.1016/0749-596X(90)90071-7

Frisson, S. (2009). Semantic underspecification in language processing. *Language and Linguistics Compass*, *3*(1), 111–127. https://doi.org/10.1111/j.1749-818X.2008.00104.x

Frisson, S., & Frazier, L. (2004). *Processing polysemy: Making sense of sense* (Poster at CUNY 17, University of Maryland).

Frisson, S., & Pickering, M. J. (1999). The processing of metonymy: Evidence from eye movements. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*(6), 1366–1383. https://doi.org/10.1037/0278-7393.25.6.1366

Frisson, S., & Pickering, M. J. (2001). Obtaining a figurative interpretation of a word: Support for underspecification. *Metaphor and Symbol*, *16*(3-4), 149–171. https://doi.org/10.1080/10926488.2001.9678893

Garvey, C., & Caramazza, A. (1974). Implicit causality in verbs. *Linguistic Inquiry*, *5*(3), 459–464.

Gernsbacher, M. A. (1989). Mechanisms that improve referential access. *Cognition*, *32*, 99–156. https://doi.org/10.1016/0010-0277(89)90001-2

Gibson, E., Futrell, R., Piantadosi, S. T., Dautriche, I., Mahowald, K., Bergen, L., & Levy, R. (2019). How efficiency shapes human language. *Trends in Cognitive Sciences*, *23*(5), 389–407. https://doi.org/10.1016/j.tics.2019.02.003

Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, *20*(11), 818–829.

Goodman, N. D., & Stuhlmüller, A. (2013). Knowledge and implicature: Modeling language understanding as social cognition. *Topics in Cognitive Science*, *5*, 173–184. https://doi.org/10.1111/tops.12007

Gordon, P. C., Grosz, B. J., & Gilliom, L. A. (1993). Pronouns, names, and the centering of attention in discourse. *Cognitive Science*, *17*, 311–347.

Gordon, P. C., & Scearce, K. A. (1995). Pronominalization and discourse coherence, discourse structure and pronoun interpretation. *Memory & Cognition*, *23*(3), 313–323. https://doi.org/10.3758/BF03197233

Grant, M., Sloggett, S., & Dillon, B. (2020). Processing ambiguities in attachment and pronominal reference. *Glossa*, *5*(1), 77. https://doi.org/10.5334/gjgl.852

Greene, S. B., McKoon, G., & Ratcliff, R. (1992). Pronoun resolution and discourse models. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*(2), 266–283. https://doi.org/10.1037/0278-7393.18.2.266

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and semantics 3: Speech acts* (pp. 41–58). Academic Press.

Grodner, D., Klein, N. M., Carbary, K. M., & Tanenhaus, M. K. (2010). "Some," and possibly all, scalar implicatures are not delayed: Evidence for immediate pragmatic enrichment. *Cognition*, *116*, 42–55. https://doi.org/10.1016/j.cognition.2010.03.014

Grosz, B. J., & Sidner, C. L. (1986). Attention, intentions, and the structure of discourse. *Computational Linguistics*, *12*(3), 175–2014.

Gulordava, K., Bojanowski, P., Grave, E., Linzen, T., & Baroni, M. (2018). Colorless green recurrent networks dream hierarchically. In *Proceedings of NAACL-HLT 16* (pp. 1195–1205). https://doi.org/10.18653/v1/N18-1108

Hahn, N., Snedeker, J., & Rabagliati, H. (2015). Rapid linguistic ambiguity resolution in young children with Autism Spectrum Disorder: Eye tracking evidence for the limits of weak central coherence. *Autism Research*, *8*, 717–726. https://doi.org/10.1002/aur.1487

Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of NAACL 2* (pp. 159–166). https://aclanthology.org/N01-1021

Hale, J. (2006). Uncertainty about the rest of the sentence. *Cognitive Science*, *30*, 643–672.

Hale, J. (2011). What a rational parser would do. *Cognitive Science*, *35*, 399–443.

Halldorson, M., & Singer, M. (2002). Inference processes: Integrating relevant knowledge and text information. *Discourse Processes*, *34*(2), 145–161. https://doi.org/10.1207/S15326950DP3402_2

Hartshorne, J. K., & Snedeker, J. (2014). *The speed of inference: Evidence against rapid use of context in calculation of scalar implicatures* (Manuscript, Harvard University).

Hedge, C., Powell, G., Bompas, A., & Sumner, P. (2022). Strategy and processing speed eclipse individual differences in control ability in conflict tasks. *Journal of Exper-*

imental Psychology: Learning, Memory, and Cognition, *48*(10), 1448–1469. https://doi.org/10.1037/xlm0001028

Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods*, *50*, 1166–1186. https://doi.org/10.3758/s13428-017-0935-1

Hedge, C., Vivian-Griffiths, S., Powell, G., Bompas, A., & Sumner, P. (2019). Slow and steady? Strategic adjustments in response caution are moderately reliable and correlate across tasks. *Consciousness and Cognition*, *75*, 102797. https://doi.org/10.1016/j.concog.2019.102797

Hobbs, J. R. (1979). Coherence and coreference. *Cognitive Science*, *3*, 67–90. https://doi.org/10.1207/s15516709cog0301_4

Hobbs, J. R. (1990). *Literature and cognition.* CSLI Publications.

Hoek, J., Rohde, H., Evers-Vermeul, J., & Sanders, T. J. M. (2021a). Expectations from relative clauses: Real-time coherence updates in discourse processing. *Cognition*, *210*, 104581.

Hoek, J., Rohde, H., Evers-Vermeul, J., & Sanders, T. J. M. (2021b). Scolding the child who threw the scissors: Shaping discourse expectations by restricting referents. *Language, Cognition and Neuroscience*, *36*(3), 382–399.

Horowitz, A. C., Schneider, R. M., & Frank, M. C. (2018). The trouble with quantifiers: Exploring children's deficits in scalar implicature. *Child Development*, *89*(6), e572–e593. https://doi.org/10.1111/cdev.13014

Huang, K.-J., & Staub, A. (2021). Using eye tracking to investigate failure to notice word transpositions in reading. *Cognition*, *216*, 104846. https://doi.org/10.1016/j.cognition.2021.104846

Huang, Y. T., & Snedeker, J. (2009a). Online interpretation of scalar quantifiers: Insight into the semantics–pragmatics interface. *Cognitive Psychology*, *58*, 376–415. https://doi.org/10.1016/j.cogpsych.2008.09.001

Huang, Y. T., & Snedeker, J. (2009b). Semantic meaning and pragmatic interpretation in 5-year-olds: Evidence from real-time spoken language comprehension. *Developmental Psychology*, *45*(6), 1723–1739. https://doi.org/10.1037/a0016704

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. https://doi.org/10.1080/01690965.2010.508641

Huang, Y. T., & Snedeker, J. (2018). *Some* inferences still take time: Prosody, predictability, and the speed of scalar implicatures. *Cognitive Psychology*, *102*, 105–126. https://doi.org/10.1016/j.cogpsych.2018.01.004

Hunt, L., III, Politzer-Ahles, S., Gibson, L., Minai, U., & Fiorentino, R. (2013). Pragmatic inferences modulate N400 during sentence comprehension: Evidence from picture-sentence verification. *Neuroscience Letters*, *534*, 246–251. https://doi.org/10.1016/j.neulet.2012.11.044

Husband, E. M., Stockall, L., & Beretta, A. (2008). *VP-internal event composition* [Manuscript, Michigan State University].

Jasinskaja, K. (2016). *Not at issue any more* [Manuscript, University of Cologne].

Johnson, S. G. B., & Keil, F. C. (2014). Causal inference and the hierarchical structure of experience. *Journal of Experimental Psychology: General*, *143*(6), 2223–2241.

Kambe, G., Rayner, K., & Duffy, S. A. (2001). Global context effects on processing lexically ambiguous words: Evidence from eye fixations. *Memory & Cognition*, *29*(2), 363–372. https://doi.org/10.3758/BF03194931

Keenan, J. M., Baillet, S. D., & Brown, P. (1984). The effects of causal cohesion on comprehension and memory. *Journal of Verbal Learning and Verbal Behavior*, *23*, 115–126.

Kehler, A. (2002). *Coherence, reference, and the theory of grammar*. CSLI Publications.

Kehler, A., Kertz, L., Rohde, H., & Elman, J. L. (2007). Coherence and coreference revisited. *Journal of Semantics*, *25*(1), 1–44. https://doi.org/10.1093/jos/ffm018

Kehler, A., & Rohde, H. (2013). A probablistic reconciliation of coherence-driven and centering-driven theories of pronoun interpretation. *Theoretical Linguistics*, *39*, 1–37. https://doi.org/10.1515/tl-2013-0001

Kehler, A., & Rohde, H. (2017). Evaluating an expectation-driven Question-Under-Discussion model of discourse interpretation. *Discourse Processes*, *54*(3), 219–238. https://doi.org/10.1080/0163853x.2016.1169069

Kehler, A., & Rohde, H. (2019). Prominence and coherence in a Bayesian theory of pronoun interpretation. *Journal of Pragmatics*, *154*, 63–78.

Khorsheed, A., Price, J., & van Tiel, B. (2022). Sources of cognitive cost in scalar implicature processing: A review. *Frontiers in Communication*, *7*, 990044. https://doi.org/10.3389/fcomm.2022.990044

Kidd, E., & Bavin, E. L. (2007). Lexical and referential influences on on-line spoken language comprehension: A comparison of adults and primary-school-age children. *First Language*, *27*(1), 29–52. https://doi.org/10.1177/0142723707067437

Klein, D. E., & Murphy, G. L. (2001). The representation of polysemous words. *Journal of Memory and Language*, *45*, 259–282. https://doi.org/10.1006/jmla.2001.2779

Klepousniotou, E. (2002). The processing of lexical ambiguity: Homonymy and polysemy in the lexicon. *Brain & Language*, *81*, 205–223. https://doi.org/10.1006/brln.2001.2518

Klepousniotou, E., Titone, D., & Romero, C. (2008). Making sense of word senses: The comprehension of polysemy depends on sense overlap. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(6), 1534–1543. https://doi.org/10.1037/a0013012

Kocijan, V., Davis, E., Lukasiewicz, T., Marcus, G., & Morgenstern, L. (2023). The defeat of the Winograd Schema Challenge. *Artificial Intelligence*, *325*, 103971. https://doi.org/10.1016/j.artint.2023.103971

Koev, T. K. (2013). *Apposition and the structure of discourse* [Dissertation]. Rutgers University.

Koornneef, A. W., & Sanders, T. J. M. (2013). Establishing coherence relations in discourse: The influence of implicit causality and connectives on pronoun resolution. *Language and Cognitive Processes*, *28*(8), 1169–1206. https://doi.org/10.1080/01690965.2012.699076

Kratzer, A. (1981). Partition and revision: The semantics of counterfactuals. *Journal of Philosophical Logic*, *10*, 201–216. https://doi.org/10.1007/BF00248849

Kuperberg, G. R., Brothers, T., & Wlotko, E. W. (2020). A tale of two positivities and the N400: Distinct neural signatures are evoked by confirmed and violated predictions at different levels of representation. *Journal of Cognitive Neuroscience*, *32*(1), 12–35. https://doi.org/10.1162/jocn_a_01465

Kuperberg, G. R., Paczynski, M., & Ditman, T. (2011). Establishing causal coherence across sentences: An ERP study. *Journal of Cognitive Neuroscience*, *23*(5), 1230–1246.

Kurtzman, H. S., & MacDonald, M. C. (1993). Resolution of quantifier scope ambiguities. *Cognition*, *48*, 243–279.

Landman, F. (2000). *Events and plurality*. Springer.

Lee, M. D., & Wagenmakers, E.-J. (2013). *Bayesian cognitive modeling: A practical course*. Cambridge University Press.

Levesque, H. J., Davis, E., & Morgenstern, L. (2012). The Winograd Schema Challenge. In *Proceedings of KR 20* (pp. 552–561).

Levinson, L. (2023). Beyond surprising: English event structure in the maze. In *Proceedings of ELM 2* (pp. 176–188). https://doi.org/10.3765/elm.2.5384

Levinson, S. C. (2016). "Process and perish" or multiple buffers with push-down stacks? *Behavioral and Brain Sciences*, *39*, e81. https://doi.org/https://doi.org/10.1017/S0140525X15000862

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177. https://doi.org/10.1016/j.cognition.2007.05.006

Lewis, D. (1973). Causation. *Journal of Philosophy*, *70*(17), 556–567. https://doi.org/10.2307/2025310

Lewis, R. L., & Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science*, *29*, 375–419.

Lewis, S. (2013). *Pragmatic enrichment in language processing and development* [Dissertation]. University of Maryland.

Lidz, J. (2018). The scope of children's scope: Representation, parsing and learning. *Glossa*, *33*(1), 33. https://doi.org/10.5334/gjgl.339

Logačev, P. (2023). The role of underspecification in relative clause attachment: Speed-accuracy tradeoff evidence. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *49*(9), 1471–1493. https://doi.org/10.1037/xlm0001164

Logačev, P., & Vasishth, S. (2016). A multiple-channel model of task-dependent ambiguity resolution in sentence comprehension. *Cognitive Science*, *40*, 266–298. https://doi.org/10.1111/cogs.12228

Lowder, M. W., & Gordon, P. C. (2013). It's hard to offend the college: Effects of sentence structure on figurative-language processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*, 993–1011. https://doi.org/10.1037/a0031671

Luce, R. D. (1986). *Response times: Their role in inferring elementary mental organization*. Oxford.

Luke, S. G., & Christianson, K. (2016). Limits on lexical prediction during reading. *Cognitive Psychology*, *88*, 22–60.

MacDonald, M. C., Pearlmutter, N. J., & Seidenberg, M. S. (1994). Lexical nature of syntactic ambiguity resolution. *Psychological Review*, *101*(4), 676–703. https://doi.org/10.1037/0033-295x.101.4.676

Mackie, J. L. (1974). *The cement of the universe: A study of causation*. Clarendon Press. https://doi.org/10.1093/0198246420.001.0001

Mak, W. M., & Sanders, T. J. M. (2013). The role of causality in discourse processing: Effects of expectation and coherence relations. *Language and Cognitive Processes*, *28*(9), 1414–1437. https://doi.org/10.1080/01690965.2012.708423

Mann, W. C., & Thompson, S. A. (1988). Rhetorical Structure Theory: Toward a functional theory of text organization. *Text*, *8*(3), 243–281.

Marcus, M. P., Santorini, B., Marcinkiewicz, M. A., & Taylor, A. (1999). *Treebank-3*. https://doi.org/10.35111/gq1x-j780

Marty, P., & Chemla, E. (2013). Scalar implicatures: Working memory and a comparison with *only*. *Frontiers in Psychology*, *4*, 403. https://doi.org/0.3389/fpsyg.2013.00403

Marty, P., Chemla, E., & Spector, B. (2013). Interpreting numerals and scalar items under memory load. *Lingua*, *133*, 152–163. https://doi.org/10.1016/j.lingua.2013.03.006

McElree, B. (1993). The locus of lexical preference effects in sentence comprehension: A time-course analysis. *Journal of Memory and Language*, *32*, 536–571. https://doi.org/10.1006/jmla.1993.1028

McElree, B., Frisson, S., & Pickering, M. J. (2006). Deferred interpretations: Why starting Dickens is taxing but reading Dickens isn't. *Cognitive Science*, *30*, 181–192. https://doi.org/10.1207/s15516709cog0000_49

McElree, B., & Griffith, T. (1995). Syntactic and thematic processing in sentence comprehension: Evidence for a temporal dissociation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(1), 134–157.

McHugh, D. (2020). Are causes ever too strong? Downward monotonicity in the causal domain. In *Monotonicity in logic and language* (pp. 125–146). Springer. https://doi.org/10.1007/978-3-662-62843-0_7

McHugh, D. (2023a). *Causation and modality* [Dissertation]. University of Amsterdam.

McHugh, D. (2023b). Exhaustification in the semantics of *cause* and *because*. *Glossa*, *8*(1), 1–38. https://doi.org/10.16995/glossa.7663

Millis, K. K., Golding, J. M., & Barker, G. (1995). Causal connectives increase inference generation. *Discourse Processes*, *20*(1), 29–49. https://doi.org/10.1080/01638539509544930

Millis, K. K., & Just, M. A. (1994). The influence of connectives on sentence comprehension. *Journal of Memory and Language*, *33*, 128–147.

Mirault, J., Snell, J., & Grainger, J. (2018). You that read wrong again! A transposed-word effect in grammaticality judgments. *Psychological Science*, *29*(12), 1922–1929. https://doi.org/10.1177/0956797618806296

Misra, K. (2022). *minicons: Enabling flexible behavioral and representational analyses of transformer language models* (Preprint hosted on arXiv, arXiv:2203.13112).

Miyake, A., Friedman, N. P., Rettinger, D. A., Shah, P., & Hegarty, M. (2001). How are visuospatial working memory, executive functioning, and spatial abilities related? A latent variable analysis. *Journal of Experimental Psychology: General*, *130*, 621–640. https://doi.org/10.1037/0096-3445.130.4.621

Moens, M., & Steedman, M. (1988). Temporal ontology and temporal reference. *Computational Linguistics*, *14*(2), 15–28.

Morris, R. K. (2006). Lexical processing and sentence context effects. In M. J. Traxler & M. A. Gernsbacher (Eds.), *Handbook of psycholinguistics* (2nd ed., pp. 377–401). Academic Press. https://doi.org/10.1016/B978-012369374-7/50011-0

Mulder, G. (2008). *Understanding causal coherence relations* [Dissertation]. Utrecht University.

Murray, J. D. (1997). Connectives and narrative text: The role of continuity. *Memory & Cognition*, *25*(2), 227–236. https://doi.org/10.3758/bf03201114

Myers, J. L., Shinjo, M., & Duffy, S. A. (1987). Degree of causal relatedness and memory. *Journal of Memory and Language*, *26*(4), 453–465. https://doi.org/10.1016/0749-596x(87)90101-x

Nicenboim, B., Schad, D. J., & Vasishth, S. (2022). *An introduction to Bayesian data analysis for cognitive science*. Advance manuscript [version February 21, 2022]. https://vasishth.github.io/bayescogsci/book/

Nicol, J. L., Forster, K. I., & Vereš, C. (1997). Subject-verb agreement processes in comprehension. *Journal of Memory and Language*, *36*(4), 569–587. https://doi.org/10.1006/jmla.1996.2497

Nieuwland, M. S., Ditman, T., & Kuperberg, G. R. (2010). On the incrementality of pragmatic processing: An ERP investigation of informativeness and pragmatic abilities. *Journal of Memory and Language*, *63*, 324–346. https://doi.org/10.1016/j.jml.2010.06.005

Nieuwland, M. S., & Kuperberg, G. R. (2008). When the truth is not too hard to handle: An event-related potential study on the pragmatics of negation. *Psychological Science*, *19*(12), 1213–1218. https://doi.org/10.1111/j.1467-9280.2008.02226.x

Nouwen, R. (2012). Plurality. In M. Aloni & P. J. E. Dekker (Eds.), *Cambridge handbook of formal semantics* (pp. 267–284). Cambridge. https://doi.org/10.1017/CBO9781139236157.010

Omaki, A., Lau, E. F., Davidson White, I., Dakan, M. L., Apple, A., & Phillips, C. (2015). Hyper-active gap filling. *Frontiers in Psychology*, *6*, 384.

Onifer, W., & Swinney, D. A. (1981). Accessing lexical ambiguities during sentence comprehension: Effects of frequency of meaning and contextual bias. *Memory & Cognition*, *9*(3), 225–236. https://doi.org/10.3758/BF03196957

Otake, T., Hatano, G., Cutler, A., & Mehler, J. (1993). Mora or syllable? Speech segmentation in Japanese. *Journal of Memory and Language*, *32*, 258–278.

Paape, D., & Vasishth, S. (2021). Is reanalysis selective when regressions are consciously controlled? *Glossa Psycholinguistics*, *1*(1), 2. https://doi.org/10.5070/G601139

Pacht, J. M., & Rayner, K. (1993). The processing of homophonic homographs during reading: Evidence from eye movement studies. *Journal of Psycholinguistic Research*, *22*(2), 251–271. https://doi.org/10.1007/BF01067833

Paczynski, M., Jackendoff, R., & Kuperberg, G. (2014). When events change their nature: The neurocognitive mechanisms underlying aspectual coercion. *Journal of Cognitive Neuroscience*, *26*(9), 1905–1917. https://doi.org/10.1162/jocn_a_00638

Partee, B. H. (1984). Nominal and temporal anaphora. *Linguistics and Philosophy*, *7*(3), 243–286. https://doi.org/10.1007/bf00627707

Paterson, K. B., Filik, R., & Liversedge, S. P. (2008). Competition during the processing of quantifier scope ambiguities: Evidence from eye movements during reading. *Quarterly Journal of Experimental Psychology*, *61*(3), 459–473. https://doi.org/10.1080/17470210701255317

Pickering, M. J., & Frisson, S. (2001). Processing ambiguous verbs: Evidence from eye movements. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*(2), 556–573. https://doi.org/10.1037/0278-7393.27.2.556

Pickering, M. J., McElree, B., Frisson, S., Chen, L., & Traxler, M. J. (2006). Underspecification and aspectual coercion. *Discourse Processes*, *42*(2), 131–155. https://doi.org/10.1207/s15326950dp4202_3

Piñango, M. M., Winnick, A., Ullah, R., & Zurif, E. (2006). Time-course of semantic composition: The case of aspectual coercion. *Journal of Psycholinguistic Research*, *35*(3), 233–244. https://doi.org/10.1007/s10936-006-9013-z

Piñango, M. M., Zurif, E., & Jackendoff, R. (1999). Real-time processing implications of enriched composition at the syntax-semantics interface. *Journal of Psycholinguistic Research*, *28*(4), 395–414. https://doi.org/10.1023/A:1023241115818

Polanyi, L. (1985). A theory of discourse structure and discourse coherence. In *Proceedings of CLS 21*.

Polanyi, L. (1988). A formal model of the structure of discourse. *Journal of Pragmatics*, *12*, 601–638.

Politzer-Ahles, S., & Fiorentino, R. (2013). The realization of scalar inferences: Context sensitivity without processing cost. *PLOS ONE*, *8*(5), e63943. https://doi.org/10.1371/journal.pone.0063943

Poppels, T., & Levy, R. (2016). Structure-sensitive noise inference: Comprehenders expect exchange errors. In *Proceedings of CogSci 38* (pp. 378–383).

Potter, M. C. (2016). Conceptual short-term memory (CSTM) supports core claims of Christiansen and Chater. *Behavioral and Brain Sciences*, *39*, e88. https://doi.org/https://doi.org/10.1017/S0140525X15000928

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A., & Webber, B. (2008). The Penn Discourse TreeBank 2.0. In *Proceedings of LREC 6* (pp. 2961–2968).

Pylkkänen, L., Llinás, R., & Murphy, G. L. (2006). The representation of polysemy: MEG evidence. *Journal of Cognitive Neuroscience*, *18*(1), 97–109. https://doi.org/10.1162/089892906775250003

Pylkkänen, L., & McElree, B. (2006). The syntax-semantics interface: On-line composition of sentence meaning. In M. J. Traxler & M. A. Gernsbacher (Eds.), *Handbook of psycholinguistics* (2nd ed., pp. 539–579). Academic Press. https://doi.org/10.1016/B978-012369374-7/50015-8

Qi, Z., Love, J., Fisher, C., & Brown-Schmidt, S. (2020). Referential context and executive functioning influence children's resolution of syntactic ambiguity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *46*(10), 1922–1947. https://doi.org/10.1037/xlm0000886

R Core Team. (2016). R: A language and environment for statistical computing. https://www.R-project.org/

Rabagliati, H., Pylkkänen, L., & Marcus, G. F. (2013). Top-down influence in young children's linguistic ambiguity resolution. *Developmental Psychology*, *49*(6), 1076–1089. https://doi.org/10.1037/a0026918

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). *Language models are unsupervised multitask learners* (Report). OpenAI.

Radó, J., & Bott, O. (2012). Underspecified representations of scope ambiguity? In *Logic, language and meaning: 18th Amsterdam Colloquium* (pp. 180–189). Springer. https://doi.org/10.1007/978-3-642-31482-7_19

Ratcliff, R. (1985). Theoretical interpretations of the speed and accuracy of positive and negative responses. *Psychological Review*, *92*(2), 212–225. https://doi.org/10.1037/0033-295X.92.2.212

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation*, *20*, 873–922. https://doi.org/10.1162/neco.2008.12-06-420

Ratcliff, R., Smith, P. L., Brown, S. D., & McKoon, G. (2016). Diffusion decision model: Current issues and history. *Trends in Cognitive Sciences*, *20*(4), 260–281.

Rayner, K., & Frazier, L. (1989). Selection mechanisms in reading lexically ambiguous words. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*(5), 779–790. https://doi.org/10.1037/0278-7393.15.5.779

Rayner, K., Pacht, J. M., & Duffy, S. A. (1994). Effects of prior encounter and global discourse bias on the processing of lexically ambiguous words: Evidence from eye fixations. *Journal of Memory and Language*, *33*, 527–544. https://doi.org/10.1006/jmla.1994.1025

Rayner, K., Reichle, E. D., Stroud, M. J., Williams, C. C., & Pollatsek, A. (2006). The effect of word frequency, word predictability, and font difficulty on the eye movements of young and older readers. *Psychology and Aging*, *21*(3), 448–465. https://doi.org/10.1037/0882-7974.21.3.448

Reed, A. V. (1973). Speed-accuracy trade-off in recognition memory. *Science*, *181*, 574–576. https://doi.org/10.1126/science.181.4099.574

Reichman, R. (1978). Conversational coherency. *Cognitive Science*, *2*, 288–327.

Rigalleau, F., Caplan, D., & Baudiffier, V. (2004). New arguments in favour of an automatic gender pronominal process. *Quarterly Journal of Experimental Psychology*, *57A*(5), 893–933. https://doi.org/10.1080/02724980343000549

Roberts, C. (1987). *Modal subordination, anaphora, and distributivity* [Dissertation]. UMass Amherst.

Rohde, H. (2019). Pronoun interpretation and production. In C. Cummins & N. Katsos (Eds.), *The Oxford handbook of experimental semantics and pragmatics* (pp. 452–473). Oxford University PRess.

Rohde, H., & Horton, W. S. (2014). Anticipatory looks reveal expectations about discourse relations. *Cognition*, *133*, 667–691.

Rohde, H., Levy, R., & Kehler, A. (2011). Anticipating explanations in relative clause processing. *Cognition*, *118*(3), 339–358. https://doi.org/10.1016/j.cognition.2010.10.016

Rohde, H., Tyler, J., & Carlson, K. (2017). Form and function: Optional complementizers reduce causal inferences. *Glossa*, *2*(1), 53.

Rooth, M. (1985). *Association with focus* [Dissertation]. UMass Amherst.

Rouder, J. N., Province, J. M., Morey, R. D., Gomez, P., & Heathcote, A. (2015). The lognormal race: A cognitive-process model of choice and latency with desirable psychometric properties. *Psychometrika*, *80*(2), 491–513. https://doi.org/10.1007/S11336-013-9396-3

Samson, D., Apperly, I. A., Braithwaite, J. J., Andrews, B. J., & Bodley Scott, S. E. (2010). Seeing it their way: Evidence for rapid and involuntary computation of what other people see. *Journal of Experimental Psychology: Human Perception and Performance*, *36*(5), 1255–1266. https://doi.org/10.1037/a0018729

Sanders, T. J. M. (2005). Coherence, causality, and cognitive complexity in discourse. In *Proceedings of the symposium on the exploration and modelling of meaning*.

Sanders, T. J. M., & Noordman, L. G. M. (2000). The role of coherence relations and their linguistic markers in text processing. *Discourse Processes*, *29*(1), 37–60.

Sasaki, K. (2021). *Components of coherence* [Dissertation]. UC Santa Cruz.

Schad, D. J., Nicenboim, B., Bürkner, P.-C., Betancourt, M., & Vasishth, S. (2022). Workflow techniques for the robust use of Bayes factors [Advance online publication]. *Psychological Methods*. https://doi.org/10.1037/met0000472

Schmidtke, J. (2018). Pupillometry in linguistic research: An introduction and review for second language researchers. *Studies in Second Language Acquisition*, *40*, 529–549. https://doi.org/10.1017/S0272263117000195

Schwarzschild, R. (1994). Plurals, presuppositions and the sources of distributivity. *Natural Language Semantics*, *2*, 201–248. https://doi.org/10.1007/BF01256743

Sedivy, J. C., Tanenhaus, M. K., Chambers, C. G., & Carlson, G. N. (1999). Achieving incremental semantic interpretation through contextual representation. *Cognition*, *71*, 109–147. https://doi.org/10.1016/S0010-0277(99)00025-6

Sherman, M. A. (1976). Adjectival negation and the comprehension of multiply negated sentences. *Journal of Verbal Learning and Verbal Behavior*, *15*, 143–157. https://doi.org/10.1016/0022-5371(76)90015-3

Singer, M., & Halldorson, M. (1996). Constructing and validating motive bridging inferences. *Cognitive Psychology*, *30*, 1–38. https://doi.org/10.1006/cogp.1996.0001

Sloggett, S., Duff, J., Van Handel, N., Sasaki, K., Rich, S., Orth, W., Anand, P., & Rysling, A. (2023). *Ambiguous isn't necessarily underspecified: Evidence from three tasks* [Manuscript, Northwestern University].

Snedeker, J. (2013). Children's sentence processing. In R. P. G. van Gompel (Ed.), *Sentence processing* (pp. 189–220). Psychology Press.

Soederberg Miller, L. M., & Stine-Morrow, E. A. L. (1998). Aging and the effects of knowledge on on-line reading strategies. *Journal of Gerontology: Psychological Sciences*, *53B*(4), P223–P233.

Stalnaker, R. C. (1968). A theory of conditionals. In N. Rescher (Ed.), *Studies in logical theory* (pp. 98–113). Blackwell. https://doi.org/10.1007/978-94-009-9117-0_2

Stan Development Team. (2019). Stan modeling language users guide and reference manual, version 2.26. https://mc-stan.org

Staub, A. (2007). The parser doesn't ignore intransitivity, after all. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(3), 550–569.

Stewart, A. J., Holler, J., & Kidd, E. (2007). Shallow processing of ambiguous pronouns: Evidence for delay. *Quarterly Journal of Experimental Psychology*, *60*(12), 1680–1696. https://doi.org/10.1080/17470210601160807

Stine-Morrow, E. A. L., Soederberg Miller, L. M., & Hertzog, C. (2006). Aging and self-regulated language processing. *Psychological Bulletin*, *132*(4), 582–606.

Stockall, L., Husband, E. M., & Beretta, A. (2010). *The online composition of events* [Queen Mary's Occasional Papers Advancing Linguistics, 19].

Stone, K., Nicenboim, B., Vasishth, S., & Rösker, F. (2023). Understanding the effects of constraint and predictability in ERP. *Neurobiology of Language*, 221–256. https://doi.org/10.1162/nol_a_00094

Stowe, L. A. (1986). Parsing WH-constructions: Evidence for on-line gap location. *Language and Cognitive Processes*, *1*(3), 227–245. https://doi.org/10.1080/01690968608407062

Sun, C., & Breheny, R. (2020). Another look at the online processing of scalar inferences: An investigation of conflicting findings from visual-world eye-tracking studies. *Language, Cognition and Neuroscience*, *35*(8), 949–979. https://doi.org/10.1080/23273798.2019.1678759

Sun, C., Pankratz, E., & van Tiel, B. (2023). *The effect of polarity on scalar implicature processing* [Poster at AMLaP Asia 2].

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction* (2nd ed.). MIT Press.

Swets, B., Desmet, T., Clifton, C., Jr., & Ferreira, F. (2008). Underspecification of syntactic ambiguity: Evidence from self-paced reading. *Memory & Cognition*, *36*(1), 201–216. https://doi.org/10.3758/mc.36.1.201

Swinney, D. A. (1979). Lexical access during sentence comprehension: (re)consideration of context effects. *Journal of Verbal Learning and Verbal Behavior*, *18*, 645–659.

Terkel, S. (1974). *Working*. Pantheon Books.

Todorova, M., Straub, K., Badecker, W., & Frank, R. (2000). Aspectual coercion and the online computation of sentential aspect. In *Proceedings of CogSci 22* (pp. 523–528).

Tomlinson, J. M., Jr, 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*, 18–35. https://doi.org/10.1016/j.jml.2013.02.003

Townsend, D. J. (2013). Aspectual coercion in eye movements. *Journal of Psycholinguistic Research*, *42*, 281–306. https://doi.org/10.1007/s10936-012-9216-4

Trabasso, T., & van den Broek, P. (1985). Causal thinking and the representation of narrative events. *Journal of Memory and Language*, *24*(5), 612–630. https://doi.org/10.1016/0749-596x(85)90049-x

Traxler, M. J., Pickering, M. J., & Clifton, C., Jr. (1998). Adjunct attachment is not a form of lexical ambiguity resolution. *Journal of Memory and Language*, *39*, 558–592. https://doi.org/10.1006/jmla.1998.2600

Trueswell, J. C., Sekerina, I., Hill, N. M., & Logrip, M. L. (1999). The kindergarten-path effect: Studying on-line sentence processing in young children. *Cognition*, *73*, 89–134. https://doi.org/10.1016/S0010-0277(99)00032-3

Tsvilodub, P., van Tiel, B., & Franke, M. (2023). The role of relevance, competence, and priors for scalar inferences. In *Proceedings of ELM 2* (pp. 288–298).

Tunstall, S. (1998). *The interpretation of quantifiers: Semantics & processing* [Dissertation]. UMass Amherst.

Van Dyke, J. A., & Lewis, R. L. (2003). Distinguishing effects of structure and decay on attachment and repair: A cue-based parsing account of recovery from misanalyzed ambiguities. *Journal of Memory and Language*, *49*, 285–316. https://doi.org/10.1016/s0749-596x(03)00081-0

van Gompel, R. P. G., Pickering, M. J., Pearson, J., & Liversedge, S. P. (2005). Evidence against competition during syntactic ambiguity resolution. *Journal of Memory and Language*, *52*, 284–307. https://doi.org/10.1016/j.jml.2004.11.003

van Gompel, R. P. G., Pickering, M. J., & Traxler, M. J. (2000). Unrestricted race: A new model of syntactic ambiguity resolution. In A. Kennedy, R. Radach, D. Heller, & J. Pynte (Eds.), *Reading as a perceptual process* (pp. 621–648). https://doi.org/10.1016/B978-008043642-5/50029-2

van Tiel, B., & Pankratz, E. (2021). Adjectival polarity and the processing of scalar inferences. *Glossa*, *6*(1), 32. https://doi.org/10.5334/gjgl.1457

van Tiel, B., Pankratz, E., & Sun, C. (2019). Scales and scalarity: Processing scalar inferences. *Journal of Memory and Language*, *105*, 93–107. https://doi.org/10.1016/j.jml.2018.12.002

van Tiel, B., & Schaeken, W. (2017). Processing conversational implicatures: Alternatives and counterfactual reasoning. *Cognitive Science*, *41*, 1119–1154. https://doi.org/10.1111/cogs.12362 van Tiel, B., van Miltenburg, E., Zevakhina, N., & Geurts, B. (2016). Scalar diversity. *Journal of Semantics*, *33*, 137–175. https://doi.org/10.1093/jos/ffu017

Vasishth, S., Yadav, H., Schad, D., & Nicenboim, B. (2022). Sample size determination for Bayesian hierarchical models commonly used in psycholinguistics. *Computational Brain and Behavior*, *6*, 102–126. https://doi.org/10.1007/s42113-021-00125-y

Vendler, Z. (1957). Verbs and times. *The Philosophical Review*, *66*(2), 143–160.

Vogel, R. M. (2020). The geometric mean? *Communications in Statistics: Theory and Methods*, *51*(1), 82–94. https://doi.org/10.1080/03610926.2020.1743313

Wagers, M., & Pendleton, E. (2015). Structuring expectation: Licensing animacy in relative clause comprehension. In *Proceedings of WCCFL 33* (pp. 29–46).

Warren, T., White, S. J., & Reichle, E. D. (2009). Investigating the causes of wrap-up effects: Evidence from eye movements and E-Z Reader. *Cognition*, *111*(1), 132–137. https://doi.org/10.1016/j.cognition.2008.12.011

Webber, B. (1991). Discourse modelling: Life at the bottom. In *Proceedings of the AAAI Fall Symposium Series on Discourse Structure in Natural Language Understanding and Generation* (pp. 146–151).

Weiss, A. F., Kretzschmar, F., Schlesewsky, M., Bornkessel-Schlesewsky, I., & Staub, A. (2018). Comprehension demands modulate re-reading, but not first-pass reading behavior. *Quarterly Journal of Experimental Psychology*, *71*(1), 198–210. https://doi.org/10.1080/17470218.2017.1307862

Winograd, T. (1972). Understanding natural language. *Cognitive Psychology*, *3*, 1–191. https://doi.org/10.1016/0010-0285(72)90002-3

Witzel, N., Witzel, J., & Forster, K. I. (2012). Comparisons of online reading paradigms: Eye tracking, moving-window, and maze. *Journal of Psycholinguistic Research*, *41*, 105–128. https://doi.org/10.1007/s10936-011-9179-x

Wlotko, E. W., & Federmeier, K. D. (2015). Time for prediction? The effect of presentation rate on predictive sentence comprehension during word-by-word reading. *Cortex*, *68*, 20–32. https://doi.org/10.1016/j.cortex.2015.03.014

Wolfe, M. B. W., Magliano, J. P., & Larsen, B. (2005). Causal and semantic relatedness in discourse understanding and representation. *Discourse Processes*, *39*(2-3), 165–187.

Wright, R. W. (1985). Causation in tort law. *California Law Review*, *73*, 1735–1828. https://doi.org/10.2307/3480373

Yacovone, A., Rigby, I., & Omaki, A. (2020). Children's comprehension and repair of garden-path *wh*-questions. *Language Acquisition*, *27*(4), 363–396. https://doi.org/10.1080/10489223.2020.1769623

Yano, M. (2018). Predictive processing of aspectual information: Evidence from event-related brain potentials. *Language, Cognition and Neuroscience*, *33*(6), 718–733. https://doi.org/10.1080/23273798.2017.1416150

Zehr, J., & Schwarz, F. (2018). *PennController for internet based experiments (IBEX)*. https://doi.org/10.17605/OSF.IO/MD832

Zondervan, A., Meroni, L., & Gualmini, A. (2008). Experiments on the role of the Question Under Discussion for ambiguity resolution and implicature computation in adults. In *Proceedings of SALT 18* (pp. 765–777). https://doi.org/10.3765/salt.v18i0.2486