



University of  
Massachusetts  
Amherst

## There and Gone Again: Syntactic Structure In Memory

Item Type	Dissertation (Open Access)
Authors	Andrews, Caroline
DOI	<a href="https://doi.org/10.7275/19826911">10.7275/19826911</a>
Rights	Attribution-ShareAlike 4.0 International
Download date	2026-06-12 23:51:32
Item License	<a href="http://creativecommons.org/licenses/by-sa/4.0/">http://creativecommons.org/licenses/by-sa/4.0/</a>
Link to Item	<a href="https://hdl.handle.net/20.500.14394/18381">https://hdl.handle.net/20.500.14394/18381</a>

University of Massachusetts Amherst

**ScholarWorks@UMass Amherst**

---

Doctoral Dissertations

Dissertations and Theses

---

## There and Gone Again: Syntactic Structure In Memory

Caroline Andrews

Follow this and additional works at: [https://scholarworks.umass.edu/dissertations\\_2](https://scholarworks.umass.edu/dissertations_2)



Part of the [Psycholinguistics and Neurolinguistics Commons](#)

---

**THERE AND GONE AGAIN: SYNTACTIC STRUCTURE IN MEMORY**

A Dissertation Presented

by

CAROLINE ANDREWS

Submitted to the Graduate School of the  
University of Massachusetts Amherst in partial fulfillment  
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

February 2021

Linguistics

© Copyright by Caroline Andrews 2021

All Rights Reserved

# THERE AND GONE AGAIN: SYNTACTIC STRUCTURE IN MEMORY

A Dissertation Presented

by

CAROLINE ANDREWS

Approved as to style and content by:

---

Brian Dillon, Chair

---

John Kingston, Member

---

Adrian Staub, Member

---

Joe Pater, Department Chair  
Linguistics

*To my parents and my brother, who have taught me so many things, not the least of which were courage and curiosity. And to my many wonderful teachers, especially Pat, Paulette, and Laurie, who got me started on writing and on science.*

## ACKNOWLEDGMENTS

First and foremost, many thanks are due to my committee: Brian Dillon, John Kingston, and Adrian Staub.

Brian gracefully accepted the role of my chair well before he had the title officially. He is a natural scientist in the way he thinks, with a deep grasp on how the details matter, and his influence shaped me into a far better scientist. It is thanks to him and his unshakeable commitment to being a good advisor for his students that I was able to go from a first-year with slightly out-there research goals to leaving with a dissertation and a postdoc that meets those goals.

John Kingston has been at once a phenomenal mentor, a climbing buddy, and a friend. I shall miss dearly him for all of these. This document (and much more besides) would certainly not have been possible without him and if I were someday to be half as good a mentor as he is, I should be very pleased.

Adrian Staub's mark on this dissertation comes from his consistently on point scientific scepticism, that at once makes it clear how complex psycholinguistic problems are and makes them more interesting.

In addition to my dissertation committee, there were other faculty members at UMass who deserve serious thanks. Kyle Johnson was an invaluable source of good syntax-related questions, patience with psycholinguists, professional advice, and humor. Ellen Woolford made me a much better writer and was supportive of my interest in fieldwork. Gaja Jarosz has been an enthusiastic adventure buddy, a good friend, and a python mentor. Lyn Frazier was an inspiration for her tremendous insight as both a psycholinguist and teacher. Caren Rotello and Andrew Cohen are responsible for reinventing statistics for me and making it a joy in its own right. Additional thanks goes to Rajesh Bhatt and Peggy Speas. Non-UMass faculty who made my graduate career possible include Claire Halpert, Matthew Wagers, and of course, Pranav Anand.

Beyond the faculty, a number of people deserve credit for enriching my time at UMass. In Linguistics, this included Katia Vostrikova, David Erschler, Jyoti Iyer, Jeremy Pasquereau, and more recently Anissa Neal. Anthony Yacovone and Stephanie Rich belong on this list as well, no less for being undergrads at the time. Equal thanks goes to Will Hopper, Josh Levy, Andrea Cataldo, and Merika Wilson, who made the cognitive division of Psychology a second home.

Much appreciation goes to Coral Hughto, for being witty, kind, and an all-around decent person, Sakshi Bhatia for many hours of syntactic (and non-syntactic) discussion, Rodica Ivan, also for many hours of non-scholastic discussion, Lap-Ching Keung for being my twin in over in Tobin, and Claire Moore-Cantwell, for taking on the role of older psycholinguist, for much needed day-dreaming about adventures, and for freezing in the desert in winter with me. Thuy Bui was my office buddy and all-around best person to kvetch with at a time when kvetching was very much needed. I'm grateful to Ria Mai Geguera for first being the best student in my class, then a truly excellent lab manager, and then a good friend.

Several research assistants contributed greatly to this work, especially Christian Muxica, Bhavya Pant, Amanda Doucette, and Annina van Riper.

There are four additional people who deserve special mention, and who truly shaped my graduate student career. Shayne Sloggett has been an older academic brother in so many ways. Amanda Rysling deserves thanks for inviting me to stay on her (Shayne's) couch, and really just about everything that happened after that, including panda, wine, late nights at the office and their apartment, and the shelter of a like-minded place. Alex Goebel was, well, a friend under what were at times the hardest possible circumstances to be a friend. I'm grateful to him for lending me space in his office to work, for sticky notes, for company on rides home, for playing Ocarina of Time, and being a fellow fan of the Mountain Goats. The last, but certainly not least in this category is Tom Maxfield, who somehow managed to be better at problem solving than I was at creating at problems (no small feat). I owe him for endless hours of chatting, much chocolate, advice, and making so many things possible.

Finally, I'm grateful to my family —Ned, Sharon, and Graham —for being curious alongside me from the beginning.

## ABSTRACT

### THERE AND GONE AGAIN: SYNTACTIC STRUCTURE IN MEMORY

FEBRUARY 2021

CAROLINE ANDREWS

B.A., UNIVERSITY OF CALIFORNIA SANTA CRUZ

Ph.D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor Brian Dillon

This dissertation addresses the relationship between hierarchical syntactic structure and memory in language processing of individual sentences. Hierarchical syntactic structure is a key part of human languages and language processing but its integration with memory has been uneasy ever since Sachs (1967) demonstrated that the syntactic structure of individual sentences is lost in explicit sentence recall tasks much faster than other linguistic information (lexical, semantic, etc.). Nonetheless, psycholinguists have continued to draw on memory in syntactic processing theories, in part due to (i) the explanatory power that memory can give to sentence processing hypotheses, and (ii) the conflicting results that continually replicate the basic findings of Sachs (1967, 1974) on one hand while on the other hand supporting robust, long-term implicit persistence of syntactic structure in the form of abstract syntactic priming.

The dissertation provides three case studies on syntactic structure in memory at three different time points over the course of processing. One case-study revisits syntactic persistence during the timescale which has classically provided the bulk of the evidence against syntactic structure in memory, from late in online processing to early offline processing, using a comparison of ellipsis-antecedent resolution and recognition memory over time. A second case-study looks at the sensitivity of proposed memory-operations to subject-verb agreement versus reflexive anaphora at the earliest timescale, during online sentence processing. Finally, the second half of

the dissertation focuses on the reliability of abstract syntactic priming in comprehension, with an extended test of Syntactic Adaptation theory (Fine, Jaeger, Farmer, & Qian, 2013).

The dissertation argues that while there is still some good evidence in favor of syntactic structure in memory, theories which intend to control most of online sentence processing from memory are probably premature. Even if memory does turn out to play a role in the syntactic processing of individual sentences, domain general, declarative memory is most likely an insufficient architecture to capture even the data which is most supportive of a memory-based account.

# TABLE OF CONTENTS

	Page
ACKNOWLEDGMENTS .....	v
ABSTRACT .....	vii
LIST OF TABLES .....	xii
LIST OF FIGURES .....	xv
 <b>CHAPTER</b>	
<b>1. INTRODUCTION .....</b>	<b>1</b>
1.0.1 Access to Syntactic Structure .....	2
1.0.2 Cue-Based Memory Models .....	6
1.0.3 The Regeneration Hypothesis .....	7
1.0.4 Memory and Syntactic Structure .....	8
<b>2. ACCESS TO SYNTACTIC MEMORY IN VERB PHRASE ELLIPSIS .....</b>	<b>10</b>
2.1 Introduction .....	10
2.2 Experiment 1 .....	15
2.2.1 Procedure .....	15
2.2.2 Materials .....	18
2.2.3 Participants .....	19
2.2.4 Predictions .....	19
2.2.5 RT Data Processing .....	20
2.2.6 Results .....	22
2.2.6.1 Aggregated Over Voice .....	23
2.2.6.2 Voice analysis .....	29
2.3 Discussion .....	34
<b>3. GRAMMATICAL INFLUENCES ON MEMORY OPERATIONS DURING ONLINE     PROCESSING .....</b>	<b>38</b>
3.1 Introduction .....	38
3.2 Experiment 2 .....	46
3.2.1 Norming Study .....	49
3.2.2 Participants .....	50
3.2.3 Procedure .....	50
3.2.4 Results .....	51

3.2.4.1	Agreement Results .....	51
3.2.4.2	Reflexive Anaphor Results .....	53
3.2.4.3	Aggregate Results .....	54
3.2.4.4	Extended Analyses .....	56
3.3	Discussion .....	62
<b>4.</b>	<b>PRIMING IN COMPREHENSION .....</b>	<b>66</b>
4.1	Introduction to Priming .....	66
4.2	Missing Priming in Comprehension .....	70
4.3	Paths Forward .....	75
4.3.1	Special Cases .....	76
4.3.1.1	Thothathiri & Snedecker (2008) .....	76
4.3.1.2	Coordination Priming .....	78
4.3.1.3	Discussion of Special Cases .....	79
4.3.2	Syntactic Adaptation .....	80
4.3.2.1	Cumulative Priming .....	84
4.3.2.2	Mere Exposure and Syntactic Satiation .....	87
4.3.2.3	Item Bias .....	90
4.3.3	Lexically-Filtered Comprehension .....	91
4.3.3.1	Item Bias Reprise .....	93
4.3.3.2	Attachment Priming .....	93
4.3.3.3	Verb Final .....	95
4.4	Discussion .....	97
<b>5.</b>	<b>AN EMPIRICAL TEST OF ADAPTATION AS SYNTACTIC PRIMING .....</b>	<b>102</b>
5.1	Introduction .....	102
5.2	Order Analysis of Staub, Dillon, & Clifton (2017) .....	102
5.2.1	Materials and Procedure .....	102
5.2.2	Overview of Staub et al.'s Original Results .....	103
5.2.3	Results of the Staub et al. Order Analysis .....	103
5.2.4	Discussion of Staub et al. Order Analysis .....	105
5.3	Experiment 3 .....	107
5.3.1	Participants .....	115
5.3.2	Procedure .....	115
5.3.3	Coding Sentence Completions .....	116
5.3.4	Results .....	117
5.3.5	Discussion .....	124
5.4	Experiment 4 .....	126
5.4.1	Participants .....	127
5.4.2	Procedure .....	127
5.4.3	Coding Sentence Completions .....	127

5.4.4	Results .....	129
5.4.5	Discussion.....	135
5.5	Combined Analysis of Experiments 3 & 4 .....	136
5.5.1	Embedded Clause Results .....	136
5.5.2	NP/Z Results.....	138
5.5.3	NP/S Results .....	139
5.5.4	Bayesian Analysis .....	140
5.6	General Discussion .....	145
<b>6.</b>	<b>CONCLUSION .....</b>	<b>149</b>
<b>APPENDIX: LOG-TRANSFORMED READING TIME ANALYSIS FROM CHAPTER</b>		
	<b>3 .....</b>	<b>155</b>
	<b>REFERENCES.....</b>	<b>159</b>

## LIST OF TABLES

Table	Page
2.1	Variable coding for Bayesian regression models. . . . . 23
2.2	For all combinations of task and lag by quantile, the values for $d_a$ , $c_a$ , the Area Under the Curve ( $A_Z$ ). A $p < 0.05$ rejects the hypothesis that the model is a fit for the data (Pazzaglia, Dube, & Rotello, 2013). . . . . 24
2.3	Outcome of the Bayesian logistic model applied to the data when aggregated over voice of the antecedent. The model predicted accurate responses based on Lag (time by quantile), Match/Mismatch condition, and Task. . . . . 27
2.4	For all combinations of task and lag by quantile in the Active voice, the values for $d_a$ , $c_a$ , the Area Under the Curve ( $A_Z$ ). A $p < 0.05$ rejects the hypothesis that the model is a fit for the data. . . . . 30
2.5	For all combinations of task and lag by quantile in the Passive voice, the values for $d_a$ , $c_a$ , the Area Under the Curve ( $A_Z$ ). A $p < 0.05$ rejects the hypothesis that the model is a fit for the data. . . . . 30
2.6	Outcome of the Bayesian logistic model applied to Active voice antecedents only, predicting accurate responses based on Lag (time by quantile), Match/Mismatch, and Task. . . . . 33
2.7	Outcome of the Bayesian logistic model applied to Passive voice antecedents only, predicting accurate responses based on Lag (time by quantile), Match/Mismatch, and Task. . . . . 33
3.1	Means and standard deviations by condition for the pre-Experiment 2 norming study . . . . . 50
3.2	Mean RTs at the critical region for subject-verb agreement (the verb region). Standard errors are given in parentheses. . . . . 52
3.3	Summary of the Linear Models applied to the Agreement results as $\beta$ coefficients. Standard Errors are in parentheses. $t$ -values can be obtained by dividing the $\beta$ value by the associated Standard Error. With the exception of intercepts, dark gray shading indicates significance and light gray shading indicates a marginal effect. . . . . 52
3.4	Mean reading times at the critical region for reflexive anaphors. Standard errors are given in parentheses. . . . . 53

3.5	Summary of the Linear Models applied to the Reflexive results as $\beta$ coefficients. Standard Errors are in parentheses. $t$ -values can be obtained by dividing the $\beta$ value by the associated Standard Error. With the exception of intercepts, dark gray shading indicates significance and light gray shading indicates a marginal effect. ....	54
3.6	Summary of Bayesian estimates for the full 2x2x2 model in Go Past at the critical region. The values for 2.5% and 97.5% in the table refer to the end points of the 95% confidence interval for the parameter. ....	62
5.1	Frequency of the two target structures across Experiments 3 and 4 .....	109
5.2	Relative Frequencies of sentence types the eyetracking portions in Experiment 3 and Experiment 4. Absolute quantities are given in parentheses .....	113
5.3	Coding used in models for Experiment 3. ....	117
5.4	Sentence completion results for dative preambles in Experiment 3 as raw counts and by-subject percentages of tokens per block (each block included of 14 dative preambles). ....	120
5.5	Outcome of the mixed models applied to the dative sentence completion results of Experiment 3 .....	121
5.6	Sentence completion results for relative clause fragments in Experiment 3 as raw counts and by-subject percentages of tokens per block (each block included of 14 RC preambles). ....	121
5.7	Outcome of the mixed effects models applied to the relative clause sentence completion results.....	121
5.8	Outcomes of Bayesian modeling for both embedded clause comparisons in Go Past at the critical NP. ....	124
5.9	Outcomes of Bayesian modeling for the embedded clause comparisons in Total Time at the critical NP. ....	124
5.10	Variable coding used in models for Experiment 4.....	129
5.11	Descriptive results for dative fragments in the sentence completion task of Experiment 4. Mean per Block and Mean % were both calculated by-subjects.....	131
5.12	Outcome of the mixed models applied to the dative sentence completion results of Experiment 4. Pre/Post test was coded as pre-test = -0.5 and post-test = 0.5. ....	132
5.13	Descriptive statistics for Mean per Block and Mean % were both calculated by-subjects.....	132
5.14	Outcome of the mixed effects models applied to the intransitive NP/Z fragment sentence completion results. ....	133

5.15	Descriptive results for the sub-types of the intransitive completions of NP/Z fragments. Mean count per Block and Mean % are both adjusted by-subjects.....	133
5.16	Outcomes of the Bayesian models for both garden path comparisons in Go Past (at the disambiguating verb) .....	134
5.17	Outcomes of the Bayesian models for both garden path comparisons in Total Time (at the disambiguating verb) .....	134
5.18	Between-subjects model coefficients for ORCs. Standard errors are in parentheses. Significant outcomes are highlighted in gray. ....	137
5.19	Between-subjects comparisons for NP/Z→Z garden paths. For the linear models for Go Past and Total Time, the significance statistic is <i>t</i> and it is <i>p</i> for the logistic model applied to probability of regression. ....	139
5.20	Summary of the Bayesian between-subjects models for the Z condition of garden paths.....	141
5.21	Summary of the Bayesian between-subjects models for the Z condition of garden paths.....	142
5.22	Summary of the Bayesian between-subjects models for NP/S garden paths. ....	143
5.23	Summary of the Bayesian between-subjects models for SRCs. ....	144
A.1	Mean RTs at the Critical Region for subject verb agreement (the verb region). Standard errors are given in parentheses.....	155
A.2	Summary of the Linear Models applied to the Agreement results as $\beta$ coefficients. Standard Errors are in parentheses. <i>t</i> -values can be obtained by dividing the $\beta$ value by the associated Standard Error. With the exception of intercepts, dark gray shading indicates significance and light gray shading indicates a marginal effect. ....	156
A.3	Mean reading times at the Critical Region for reflexive anaphors. Standard errors are given in parentheses. ....	156
A.4	Summary of the Linear Models applied to the Reflexive results as $\beta$ coefficients. Standard Errors are in parentheses. <i>t</i> -values can be obtained by dividing the $\beta$ value by the associated Standard Error. With the exception of intercepts, dark gray shading indicates significance and light gray shading indicates a marginal effect. ....	157

## LIST OF FIGURES

Figure	Page
1.1 Results from Sachs (1974) .....	3
1.2 Results of Miller and Isard (1964). .....	4
1.3 Transitional Error Probability for condition 1 of N. F. Johnson (1965). .....	4
1.4 Transitional Error Probability for condition 2 of N. F. Johnson (1965). .....	5
2.1 Distribution of aggregate trial RTs by lag condition after outlier rejection. Bin size is 500ms. Dotted black lines indicate the quantile boundaries used in the analysis. ....	21
2.2 ROCs for Experiment 1 by Lag, aggregated across Voice and participants. Note that the (1,1) point has been left in, but does not contribute meaningfully to the interpretation. ....	23
2.3 ROCs for Experiment 1 split by Voice and Lag. Note that the (1,1) point has been left in, but does not contribute meaningfully to the interpretation. ....	29
2.4 ROCs for Experiment 1 split by Voice and Task. Note that the (1,1) point has been left in, but does not contribute meaningfully to the interpretation. ....	32
3.1 Go Past reading times for both the respective critical and spillover regions. Error bars indicate standard errors. ....	56
3.2 Results of cumulative progression on pairwise comparisons of the attraction effects in the ungrammatical conditions of agreement and reflexives, and then the corresponding cross-dependency interaction. Regions highlighted in light gray are those where the <i>t</i> -tests between the two conditions were all significant. ....	58
3.3 Summary of Bayesian estimates for the full 2x2x2 model in Go Past at the spillover region, including a histogram of posterior of the critical three-way interaction. The values for 2.5% and 97.5% in the table refer to the end points of the 95% credible interval for the parameter. ....	59
3.4 Outcomes of individual pairwise comparisons within the ungrammatical conditions, one for agreement and one for reflexives. <i>Left</i> : Posterior distribution for Lure Number parameter in the agreement pairwise comparison. <i>Right</i> : Posterior distribution for Lure Number parameter in the reflexive pairwise comparison. ....	60
5.1 Reading Times and Probabilities of Regression for each position in the order analysis. Top: Relative NP; Bottom: Relative Verb .....	104

5.2	Reproduced from Wells, Christiansen, Race, Acheson, and MacDonald (2009). Pre- and post-test self-paced RTs for the relative clause conditions. ....	105
5.3	Change in Mean RT of the embedded clause conditions over the course of Experiment 3. Error bars represent standard error. ....	119
5.4	Change in Mean RT of garden path conditions over the course of Experiment 3. Error bars represent standard error. ....	120
5.5	Change in Mean RT of the embedded clause conditions over the course of Experiment 4. Error bars represent standard error. ....	130
5.6	Change in Mean RT of the NP/Z conditions over the course of Experiment 4. Error bars represent standard error. ....	131
5.7	Between-subject plots of change in Mean RT for ORCs sentences. Error bars represent standard errors. ....	137
5.8	Plots of change in Mean RT based on ORDER for NP/Z→Z garden paths ....	138
5.9	Change in Mean RT of the NP/S → S garden paths over the course of both experiments. Error bars represent standard error. ....	139
5.10	Posterior distributions of the EXPERIMENT parameter for the NP/S Bayesian between-subjects models. <i>Left: Go Past; Right: Total Time</i> ....	143

## CHAPTER 1

### INTRODUCTION

The past two decades have shown a marked increase in the detail of psycholinguistic models of memory, and correspondingly, the detail in the research about how online language processing uses memory to interact with syntactic structure (Dillon, 2011; Engelmann, Jäger, & Vasishth, 2015; Kush, 2013; Lewis & Vasishth, 2005; McElree, Foraker, & Dyer, 2003; Van Dyke & McElree, 2006; Wagers, 2008, *inter alia*). The most widely used of these models, “cue-based models”, are attractive not just for their explanatory potential, but also because they are in-line with models from the broader domain of memory research and therefore have the promise of explaining psycholinguistic phenomenon within a largely domain general architecture (Anderson & Neely, 1996; Kahana, 2012). However, cue-based models of psycholinguistics are not without their challenges. One of the foremost of these is that in borrowing domain general architecture intended for both long term and short term memory, linguistic models without modification predict long term retention of syntactic structure (Lewis & Vasishth, 2005; McElree, 2006), when in fact syntactic structure is famously thought to be ephemeral (Jarvella, 1971; Sachs, 1967). The lack of clear evidence for long-term syntactic memory has led to the development of alternative influential theories in which syntactic structure *never* exists in memory (Lombardi & Potter, 1992; Potter & Lombardi, 1990).

The goal of this thesis is to address the role that long-term memory plays in sentence processing, particularly syntactic processing, and consequently what modifications must be made to domain general models of memory if they are to be adapted to the demands of sentence processing. In particular the thesis is based on the following questions:

1. What is the form of syntactic memory?

and

2. How is syntactic memory accessed?

keeping in mind the very real possibility that the answer to (1) may be that there is no such cognitive object.

### 1.0.1 Access to Syntactic Structure

The idea that syntactic structure does not persist in long-term memory (henceforth LTM) dates back quite early in psycholinguistics. In a series of highly influential experiments, Sachs (1967, 1974) presented participants with target sentences embedded in short passages and then asked them to discriminate between the sentence they actually saw and a series of various paraphrases in a sentence recognition task. When participants were given the recognition task immediately after encountering the target sentence, they were quite good at differentiating the exact target sentence from all paraphrases. However, with virtually any delay (starting at 4 seconds), discrimination for the various paraphrases diverged sharply. Participants retained a relatively consistent ability to detect paraphrases which substantially changed the semantics of the event (e.g., changing the polarity of the sentence as in [2a]). However, for paraphrases that differed by a voice alternation (active to passive), lexical substitution, or what Sachs termed “Formal”<sup>1</sup> discriminability quickly dropped to chance-levels (Figure [1.1]).

- (1) Target: The Founding Fathers considered owning slaves to be immoral.
- (2) a. Semantic: The Founding Fathers didn’t consider owning slaves to be immoral.  
b. Active/Passive: Owning slaves was considered to be immoral by the Founding Fathers.  
c. “Formal”: The Founding Fathers considered owning slaves immoral.  
d. Lexical: The Founding Fathers thought owning slaves to be immoral.

Similarly, incorrect rejections of the actual target sentence mirrored incorrect acceptance rates of the non-semantic paraphrases over the same time scale.

Sachs, and others since, took these results to indicate that verbatim memory for syntax is severely limited after the end of a sentence. However, even at the time, Sachs’s findings were in tension with other results which seemed to indicate that syntactic structure served as an organizing factor in memory. The best known of these is the so-called Sentence Superiority Effect (Miller & Isard, 1963, 1964). Unlike Sach’s recognition experiments, Miller and Isard (1964) tested trained recall of sentences versus lexically-matched random strings which were 22 words long, shown in (3) and (4). Participants listened to each string five times and were asked to repeat what they had heard verbatim after each repetition.

---

<sup>1</sup>“Formal” was quite possibly the condition most focused on a purely syntactic change of the three, however the actual change seems to have been inconsistent: another example sentence from the (1974) paper involves the preposition from a particle verb alternating between preceding and following the direct object, while the example in the (1967) paper is a dative alternation. Sachs does not provide any detailed discussion of the construction of the Formal sentences.

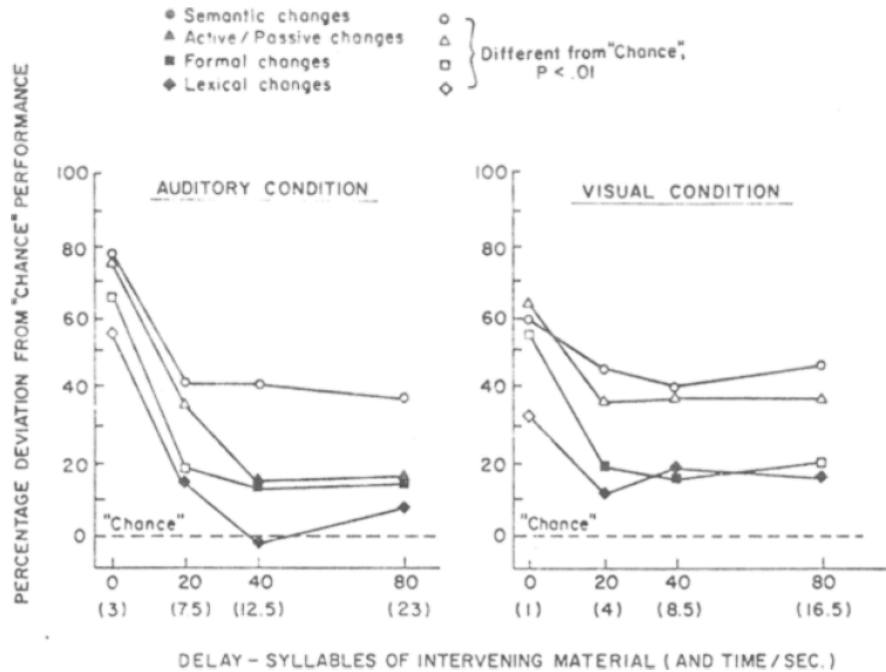


Figure 1.1. Results from Sachs (1974)

- (3) *Base*: She liked the man that visited the jeweler that made the ring that won the prize that was given at the fair.
- (4) *Random*: Won given liked that that the fair man made visited prize the at the the she that jeweler was the ring that
- (5) *Four embeddings*: The prize that the ring that the jeweler that the man that she liked visited made won was given at the fair.

Recall for sentences was dramatically improved over random word sequences, shown in Figure (1.2). However, what made this finding particularly impressive was the manipulation of levels of embedding within the sentence conditions. Using (3) as a base with zero levels of embedding, Miller and Isard tested sentences with 0-4 levels of embedding. At greater than two levels of center embedding, sentences famously become quite difficult to comprehend, as demonstrated by Miller and Isard's most embedded condition in (5). Yet even the deepest embedded conditions continued to substantially outperform random strings.

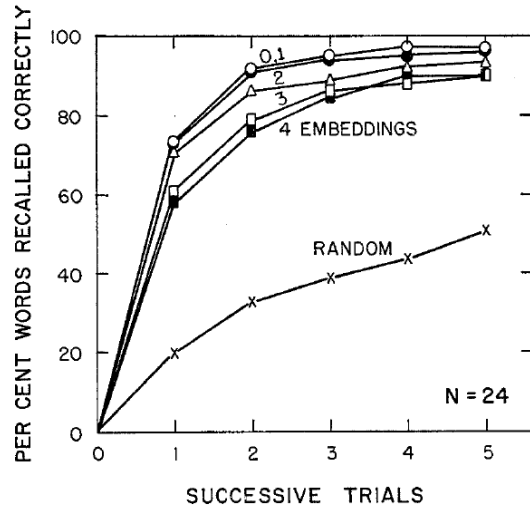


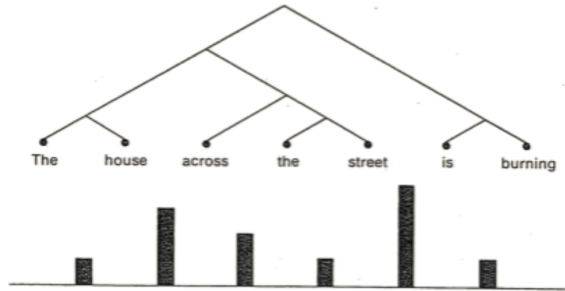
Figure 1.2. Results of Miller and Isard (1964).

The Miller and Isard (1964) findings appeared to demonstrate the ability of syntactic structure to act as an organizing factor in memory, and later studies provided evidence that this extended down to the level of constituent structure. For instance, N. F. Johnson (1965) introduced a measure that he called Transitional Error Probability (TEP) an indicator of recall success at a word-by-word level. For each word  $n$  that was correctly recalled, he tabulated whether  $n+1$  was also correctly recalled. The percentage of errors at any given transition point was taken to be a measure of how closely the two words were related in the memory representation of the sentence.



Figure 1.3. Transitional Error Probability for condition 1 of N. F. Johnson (1965).

Johnson's findings indicated that errors tracked the constituent structure of a sentence, such that transitions spanning larger distances in a syntactic tree corresponded to a greater proportion of errors. This is demonstrated by the height of the bars in Figures (1.3) and (1.4), based on N. F. Johnson (1965) via Fodor, Bever, and Garrett (1974).



**Figure 1.4.** Transitional Error Probability for condition 2 of N. F. Johnson (1965).

Another major development came from a 1986 paper by Bock, which introduced abstract syntactic priming for double object constructions and the active/passive alternation. Starting with this paper, there has been robust literature showing that experience with one structure makes a speaker more likely to use that structure again when given the choice. For instance, Bock and Griffin (2000) initially had speakers listen and repeat back a sentence such as (6). Participants were then showed a picture which depicted an event compatible with a dative verb (e.g., a boy who has a guitar and is offering it to a rock star) and asked to describe the picture. Participants were more likely to use the prepositional object (PO) frame in (8b) when they had already heard and repeated (6) than when they had repeated (7).

- (6) The governess made a pot of tea for the princess.
- (7) The lifeguard tossed the struggling swimmer a rope.
- (8) a. The boy is handing the rock star a guitar.  
b. The boy is handing a guitar to the rock star.

These results seem impossible to explain without some maintenance of abstract syntactic structure in LTM, and indeed ever since Bock (1986), priming has been the strongest argument for long-term maintenance of syntactic structure from individual sentences. This memory would be implicit, in that speakers have no conscious sense of its influence on their syntactic choices, but strong enough to have a noticeable effect from trial-to-trial and persist at minimum over the 10 intervening trials that Bock and Griffin (2000) tested. It also cannot be easily explained away by morphological overlap, as Bock (1986) showed that benefactive double object constructions like the one (6) prime and are primed by dative double object constructions, despite using different critical prepositions (*to* versus *for*). In a similar vein, it has also been shown that passive by-phrases can

prime and be primed-by locative phrases (Bock & Loebell, 1990), indicating that priming works at the level of phrase structure.

Subsequent authors have suggested that the results of Miller and Isard (1964) and N. F. Johnson (1965) and others like them can be explained by appealing to memory for the semantics of a sentence rather than syntactic structure (Dommergues & Grosjean, 1981; Grosjean, Grosjean, & Lane, 1979; Townsend & Saltz, 1972). However, Bock's abstract priming has remained reliable across a host of studies. The contrast between the early findings in favor of syntactic structure's usefulness to long-term memory and Sachs's set the stage for a running tension over the place of memory in sentence processing that never been entirely resolved. This tension has returned in recent years. As the sophistication of psycholinguistic models of sentence processing has increased dramatically, memory has become increasingly prominent in theories of sentence processing *within* a single sentence. For this new wave of memory-based psycholinguistics, researchers have turned to a class of memory models known as cue-based models.

### 1.0.2 Cue-Based Memory Models

Cue-based models have arguably been the predominant models in memory research for some time and come in several different flavors (Clark & Gronlund, 1996). However, the models used by psycholinguists are almost universally descended from Anderson's ACT-R model (Anderson & Neely, 1996; Lewis & Vasishth, 2005; Van Dyke & McElree, 2006).<sup>2</sup> The unifying feature of these models is that items in memory, traces, are composed of features/cues, each of which stand for some property of that item. These features are shared with all other items that also have that property. For instance, if the word *porcupines* appeared in a sentence, the memory representation would include the feature [+PL], which it would share with any other plural words in memory. Moreover, it would come with a set of features related to its role in the specific sentence, such as features for case, thematic role, and grammatical role.<sup>3</sup> At the very least, *porcupines* would also come with many features directly out of the lexicon which form the denotation, such as [+ANIMAL], [+ANIMATE], [-HUMAN], and [-CUDDLY].

What is notable about cue-based models from a sentence processing perspective is that they were all developed for LTM. This was neither a mistake nor a coincidence: they were brought into psycholinguistics in response to a problem memory researchers had already solved. A major

---

<sup>2</sup>Although it is worth noting that in practice, many psycholinguist models frequently abandon much of what made ACT-R distinct from other cue-based models.

<sup>3</sup>Particularly in languages where case and grammatical role diverge.

theoretical hurdle in memory research was how to handle search through LTM. Humans are able to quickly recall information or events associated with the current context even if years have passed since that information was originally encoded or last accessed. It is clearly not feasible to organize a search through all the items in LTM by serially inspecting each memory item for relevance to the current context. Instead, cue-based models implement what is known as content-addressable search: the features in the current context provide direct access to any memory which shares those cues, acting like the address in the name. The more cues a memory shares with the current context, or the probe that launches the search, the more likely it is to be retrieved. This type of mechanism allows items in memory to be retrieved quickly and efficiently whenever they are relevant to the current input, regardless of how old they are.

When content-addressable search was discovered for online sentence processing (McElree, 2000; McElree & Doshier, 1989), cue-based models provided a ready-made content-addressable framework. On the other hand, as we have seen there is evidence that syntactic structure for individual sentences is not ultimately encoded into LTM after a sentence is finished. This creates a precarious theoretical situation: LTM is invoked during the short-term syntactic processing within a sentence<sup>4</sup> but outside sentence boundaries syntactic structure is lost to LTM.

There seem to be two ways to out of this theoretical dilemma. First, it is possible to side wholly with cue-based models and deny that syntactic memory is as short-lived as previously thought. This the option pursued by this thesis and has alternately been represented by Interference models of syntactic forgetting (e.g., Lewis, 1996). A second option is to deny that memory is involved in online syntactic processing at all. This view is most clearly encapsulated by the Regeneration Hypothesis of Potter and Lombardi (1990).

### 1.0.3 The Regeneration Hypothesis

The extreme opposite of the cue-based hypothesis is the denial of any role for memory in syntactic processing at all. This the view taken by the Regeneration hypothesis put forward in Lombardi and Potter (1992); Potter and Lombardi (1990). Lombardi and Potter (1992) used a sentence recall paradigm where the induced lexical intrusion. They began with sentences containing verbs that participate in nearly synonymous syntactic alternations, such as (9).

- (9) a. The rich widow is going to give a million dollars to the university.
- b. The rich widow is going to give the university a million dollars.

---

<sup>4</sup>See McElree (2006) for arguments that sentence processing should be viewed as LTM and not working memory.

(10) The rich widow is going to donate a millions dollars to the university.

Participants were then given a list of words as a distractor task, one of which was a verb that was semantically related to the verb in the target sentence. In the case of (9) the intruder verb was *donate*. Critically, *donate* does not participate in the dative alternation. It could easily be substituted into (9a) without any changes, but a substitution into (9b) would be ungrammatical.

The results showed no detectable difference between the likelihood of *donate* intruding into (9a) and (9b). Moreover, when *donate* intruded into (9b), speakers spontaneously repaired to the grammatical PO form of the sentence. This led Potter and Lombardi to conclude that memory for syntactic structure played no role. Instead, they argued that syntactic structure was ephemeral, lasting only long enough to translate the lexical string into a semantic representation and decaying immediately. In fact, under the Regeneration hypothesis, syntax is correctly viewed as a process rather than a representation. The stable cognitive objects are the lexical and semantic representations, which can be combined to *regenerate* the sentence, including its syntactic structure, if there is need, giving the hypothesis its name.<sup>5</sup>

#### 1.0.4 Memory and Syntactic Structure

Despite decades of research, even some of the most basic questions about the relationship of memory and syntax remain unanswered. With cue-based models putting memory directly into online processing theories, the questions have returned to the forefront with urgency. There are at least two main questions which must be answered.

1. Is there a memory representation for the syntactic structure of a sentence, particularly one that could be applied in online processing?
2. If there is, what is the form of syntactic structure in memory?

Without an affirmative answer to the first question, cue-based parsing models are dead in the water. Without a clear answer to the second question, these models remain critically limited in the predictions they can make and the extend of their usefulness. Furthermore, as we will see in Chapters 4 and 5, it is not necessarily the case that asking (2) is dependent on a (1) being true.

The work in this thesis is split into three distinct parts. Chapter 2 addresses the question (1) by proposing a novel hypothesis, the Privileged Access Hypothesis, and providing a test for it.

---

<sup>5</sup>It is worth noting, however, that Potter and Lombardi were forced to conclude that a mechanism for abstract syntactic priming must still in their system, when the Potter and Lombardi (1998) tested and found priming under similar conditions as the Potter and Lombardi (1990) and Lombardi and Potter (1992) studies.

Chapter 3 looks at syntactic memory from the perspective of online processing, and investigates whether syntactic operations can influence the use of memory (as opposed to the other way around, which has been a more typical question). Finally, Chapters 4 and 5 then consider a hybrid of these questions applied to a specific problem in the domain of syntactic priming.

## CHAPTER 2

### ACCESS TO SYNTACTIC MEMORY IN VERB PHRASE ELLIPSIS

#### 2.1 Introduction

As discussed in Chapter 1, there is a considerable body of evidence that argues against long-term maintenance of syntactic form of individual sentences in memory (Jarvella, 1971; Potter & Lombardi, 1990; Sachs, 1967, *inter alia*). This work suggests that once lexical and semantic memory is accounted for, comprehenders have very little access to the syntactic structure of a sentence they have recently encountered. Chapter 1 also introduced the Regeneration Hypothesis from Potter and Lombardi (1990) as an explicit proposal for how processing might work if there is no memory for syntactic structure.

Although there is considerable support for reduced-to-nonexistent syntactic memory, ultimately the range of grammatical and experimental tests is somewhat narrow. And indeed, when the Regeneration Hypothesis was explicitly tested under conditions known to be conducive to syntactic priming, Potter and Lombardi (1998) were ultimately forced to conclude that there was a separate memory cache for syntactic priming which did exist.

The goal of this chapter is to test the Regeneration Hypothesis in a grammatical context where syntactic memory would provide the parser with an unusually strong advantage: ellipsis. Most theoretical accounts of Verb-Phrase Ellipsis (VPE) argue that a strict syntactic match is required between the elision site and the antecedent (see K. Johnson, 2001, for an overview). Tanenhaus and Carlson (1990) demonstrated this experimentally by showing that the VPE in (11a) has higher “make sense” ratings than the minimally different sentence in (11b). These results were attributed to the fact that the elided constituent in brackets in (11b) cannot both match the exact syntactic form of the antecedent constituent *be taken out* and at the same time form a grammatical second clause.

- (11) a. Someone had to take out the garbage. But Bill refused to [].  
b. The garbage had to be taken out. But Bill refused to [].  
c. The garbage had to be taken out. But Bill refused to do it.

(Tanenhaus & Carlson, 1990)

Notably, the surface anaphor in (11c) is not subject to this syntactic identity constraint, indicating that this is truly a constraint on VPE. Surface anaphors do prefer parallelism with their antecedents but ultimately are claimed to establish discourse-level anaphoric relations rather than strict syntactic matching. It is the fact that (11b) is much worse than (11c) that constitutes Tanenhaus and Carlson's argument for syntactic identity.

A strict matching requirement presents the potential for an uncomfortable state of affairs for the Regeneration Hypothesis. In order to evaluate whether the grammatical requirement for strict matching has been met, the parser needs to have reliable access to the syntax of both the elision and antecedent clauses. From this point of view, anything less than perfect syntactic memory will impede the parser's ability to apply the strict requirement. In order to discriminate between (11a) and (11b), this means that memory for the entire current sentence plus at least the previous independent clause. The implied timescale of perseverance of the elliptical antecedent is at odds with the typical findings of rapid syntactic decays in memory. In other words, VPE presents a case where the needs of the grammar directly conflict with the previous findings of syntactic memory capacity.

There seem to be two ways forward out of this conundrum. The first basic approach takes the previous findings from verbatim memory and the Regeneration Hypothesis as correct. Under this proposal, syntactic structure acts as a critical organizer of information online, but does not exist in memory in its own right.

Interestingly, this is essentially the tack taken by, or at least compatible with, a large segment of the ellipsis processing literature. The studies in this line of thought point to cases where a mismatch between the antecedent and the elided constituent do seem to be tolerated. The single most notable of these is perhaps the voice mismatch effect demonstrated in (12) and (13).

(12) This problem was to have been looked into, but obviously nobody did.

(Kehler, 2002, p. 53)

(13) This information could have been released by Gorbachov, but he chose not to.

(Hardt, 1993, p. 37)

(12) and (13) seem to be largely acceptable even though the antecedent is passive while the elided VP must be active. Other authors have noted that when the elided material and the antecedent don't match there is in fact an acceptability gradient between types of mismatch. Notably, Arregui, Clifton, Frazier, and Moulton (2006) contrasted examples such as the ones in (14).

- (14) a. None of the astronomers saw the comet, /but John did.  
 b. Seeing the comet was nearly impossible, /but John did.  
 c. The comet was nearly unseeable, /but John did.

Of the three sentences in (14), only (14a) is typically judged as fully grammatical (as supported by the results of Arregui et al.'s acceptability judgment study). The critical observation is that although both (14b) and (14c) are degraded, (14c) is noticeably worse. To handle gradience in ellipsis antecedent relations, Arregui et al. propose the VP Recycling Hypothesis, which suggests that in the absence of a syntactically matching antecedent the parser will attempt to "recycle" material from the actual input it received into a licit antecedent, but always with some penalty for not having true syntactic identity. The closer the input material is to a grammatical antecedent, the more successful the parser will be in recycling and the smaller the penalty will be. Thus, the embedded VP of (14b) is easier to accept than the negative adjective in (14c), because although neither structure is truly grammatical, the embedded VP example still contains a VP that can be easily turned into a legal antecedent.

Arregui et al. (2006) consider memory's potential to play a role in recycling, but state the core hypothesis in a way that does not actually require memory to be involved at all. An explicitly memory-based proposal for ellipsis comes from Garnham and Oakhill (1987). They embedded elliptical antecedents in passages that either did or did not support the actual interpretation of the elliptical verb, such as the one in (15)-(16). Importantly, even when the grammatically required interpretation of the ellipsis was technically implausible, the stimuli were constructed in such a way that there was a plausible interpretation in the overall context (e.g., that the nurse had examined the elderly patient).

(15) The elderly patient had been examined by the doctor { \_\_\_\_/during the ward round}.

- (16) a. *Plausible ellipsis*: The child had too.  
 b. *Implausible ellipsis*: The nurse had too.

In a stance that directly anticipates the more general Regeneration hypothesis, Garnham and Oakhill proposed that the parser resolves ellipsis by computing a semantic interpretation that is consistent with all of the comprehender's knowledge (including world knowledge) and use this to "supplement" a strict grammatical evaluation of the ellipsis. They further argue that this effect should be greater with more material intervening between the antecedent and the ellipsis as syntactic information becomes less reliable due to forgetting. In the context of their semantic

manipulation, they predicted that errors should be biased in favor of the more plausible interpretation, and that these errors should increase with more distance between the antecedent and elision site. They found both such main effects of plausibility and distance, as well as interactions of plausibility and distance, indicating that comprehenders increasingly relied on reconstructing an interpretation to fit the semantic context over the strict grammatical reading as distance increased.

While not all of the ellipsis parsing based literature has endorsed a memory-based approach as directly as Garnham and Oakhill (1987), the take away from this section is that much of the literature has been focused on the parser's reaction to imperfect applications of the strict matching requirement between the antecedent and elision site. This makes an account with no proper syntactic memory more attractive, and positions it as one of the two main hypotheses being considered in this chapter. The combined Garnham and Oakhill and Regeneration view of ellipsis may be sketched roughly as: comprehenders quickly forget the true syntax of the antecedent, but will generally attempt to reconstruct it at the ellipsis site. During reconstruction, comprehenders strive to use the semantic and lexical material that persists in memory to formulate a licit antecedent, sometimes at the expense of reformulating the exact grammatical conditions.

However, even if syntactic forgetting has been supported by portions of the ellipsis literature, this does not necessarily mean that an account with long-term syntactic memory is impossible. Where the challenge ellipsis posed for Regeneration was specifying how grammatical operations might operate over distance, the challenge for a long-term syntactic account is making clear how syntactic memory might persist in the face of so much evidence to the contrary from sentence/recall recognition. One approach that could accomplish this would be to assume that *access* to the memory trace is somehow limited. That is, the memory for the syntactic structure of a sentence is maintained in memory without issue, but only a subset of cognitive operations can make use of that type of item in memory. For the purposes of this chapter, this view will fall under the auspices of a novel hypothesis, the Privileged Access Hypothesis. This hypothesis is based on the idea that verbatim memory and ellipsis<sup>1</sup> could yield different conclusions about the time course of syntactic memory if they were different classes of processes. Specifically, ellipsis is a grammatical operation which must access a grammatical representation as part of its sole function. The operations of verbatim memory, on the other hand, —i.e., recall and recognition as applied to whole sentences —are rarely required of speakers outside of labs. Most previous

---

<sup>1</sup>And perhaps relatedly, the abstract syntactic priming discussed in Chapters 1, 4, and 5.

work has viewed this difference as incidental, but an alternative option is to view it as deeply fundamental.

- (17) *Privileged Access Hypothesis*: There is long term memory for syntax, but it is only accessible to linguistic processes. Sentence recall and recognition are not linguistic processes in the relevant sense, but ellipsis is.

While the focus of the chapter is on establishing whether there is psycholinguistic evidence in favor of the PAH, and less on what might lead to such a division between cognitive processes, it is still worthwhile to consider whether it is cognitively reasonable to propose that cognitive processes be separated by the type of linguistic information that they may access in the first place. There are at least two such proposals already in the literature.

First, the separation of memory operations in (17) could be achieved by the notion of an encapsulated language module from Fodor (1983). A language module would certainly contain grammatical knowledge, which would cover both the grammatical rules for ellipsis and syntactic memory (because syntactic memory is a form of grammatical knowledge). A modular view of language ability would also place the operations for parsing linguistic material, (such as the mechanisms behind priming or serial/parallel parsing mechanisms), inside the language module. Furthermore, Fodor's view of modularity strongly restricted information-sharing across modules, which would suggest that the operations recall and recognition borrowed from general memory would have limited success retrieving purely linguistic information. A modularized view of cognitive architecture therefore arguably predicts divisions like the one described in the PAH *a priori*.

However, the PAH itself does not imply a modular view of linguistic capacities. The separation along the lines suggested in the PAH would also be a natural consequence if linguistic memory were procedural memory (Ullman, 2016). The most commonly studied form of memory is declarative memory, which is used for the conscious retrieval of facts and events that a person has experienced. It contrasts with procedural memory, which is a form of implicit memory classically applied to motor skills. The hallmark of procedural memory is the ability to perform tasks (such as skiing or writing by hand) easily and precisely once they have been practiced, but without the ability to consciously describe how the task is being accomplished or even repeat parts of the procedure outside the context of the task.

The characteristics of procedural knowledge bear a striking resemblance to the behavior of syntactic information in memory. Syntactic knowledge is implemented consistently and precisely

once it has been acquired, yet adult speakers must be explicitly taught the rules that they are already following in order to be able to meta-reason about syntax. Thus, while declarative memory is well-suited to some forms of linguistic knowledge, most notably the lexicon, procedural memory may be a better fit for syntactic knowledge.

Without committing ourselves to any particular theoretical view of procedural memory, we may suppose that procedural traces consist of the steps needed to perform an action. Learning consists both of creating the trace in the first place and then practicing it until it becomes an established skill. Under a procedural view of the PAH, accessing syntactic knowledge is a skill in its own right. Like other skills, it can be practiced and easy in the context of some actions but remain difficult to combine with others. Grammatical operations such as ellipsis would have “practiced” using syntactic procedural memory and would have easy access to syntactic memory precisely because of this practice. Recall and recognition, while perhaps quite adept at accessing declarative memory, would have no way to interface with grammatical procedural memory directly. Instead, they would have to rely on the declarative memory for the semantic and lexical representations to guess at the syntactic representation, exactly as suggested in Potter and Lombardi’s Regeneration Hypothesis. If this were correct, the reason that Potter and Lombardi (1990) saw such apparently short syntactic memory would be because they used a specifically non-linguistic operation: recall.

## **2.2 Experiment 1**

Experiment 1 explicitly tests the Privileged Access Hypothesis versus the combined Garnham and Oakhill/Regeneration Hypothesis by comparing a non-linguistic memory-process, recognition, to a linguistic memory processes, ellipsis, over the same sentences. There were three factors of interest: i) the form of the prime sentence, ii) the amount of time intervening between the prime and the target task, and iii) the type of target task.

### **2.2.1 Procedure**

Trials were divided into three parts: presentation of the prime, lag, and target task. Each trial began with a fixation cross in the center of the screen, which ensured that participants were fixated in the correct place for the prime sentence. The prime sentence was then displayed one word at a time using centered presentation, at a rate of 300ms per word with a 100ms inter-word interval (the fixation cross counted as two words for presentation times, including the inter-word delay to ensure that participants had adequate time to focus on it). Prime sentences were in either the active or passive form, such as (18).

- (18) a. Active prime: The politician criticized the journalist over the presentation of the new bill.
- b. Passive prime: The journalist was criticized by the politician over the presentation of the new bill.

Once the prime sentence was finished, the trial entered the lag phase, in which length of time between the prime and the target task was manipulated by varying the number of math problems that a participant needed to complete after seeing the prime (between 0 and 5). In the lag 0 condition, the lag phase was skipped and trial proceeded to the target task without any delay. For lag conditions 2 and 5, simple addition problems were generated randomly in the moment using numbers between 0 and 99 such that the answer was never greater than 99. Math problems only ever consisted of two addends and only used addition. Participants responded by entering their answer in a text box just below the math problem. In order to streamline the process as much as possible, the cursor automatically appeared in the text box and pressing 'Enter' brought up the next math problem/target task (this was pointed out to participants during the instructions so that they did not need to rely on the mouse). There was no restriction on how long participants had to answer to a math problem, but reaction times were recorded from the moment that the math problem loaded. Participants were not given feedback whether they had answered correctly or not. Due to early technical limitations, the randomly generated math problem itself was not recorded for the initial set of participants, but was for later participants.

The final phase of a trial was the target task. Crucially, on any given trial, participants did not know in advance which target task they would be asked to complete, up until the task phase itself.

For critical trials in the recognition task, participants were again given either (18a) or (18b) and asked whether this was the sentence they had just seen. This time the sentences were presented in their entirety with no time limit.

In the ellipsis task, participants were presented with one of the two VP ellipsis continuations of the prime in (19). In this case, they were asked if the ellipsis was a reasonable continuation of the conversation.

- (19) a. *Active Ellipsis Target:* The TV pundit did too.
- b. *Passive Ellipsis Target:* The TV pundit was too.

Participants had two ways of knowing which of the tasks they needed to complete. The first and most prominent of these was one of two images that appeared at the top of screen in the target

task: either a large equals sign with a question mark for the recognition task or a clip-art image of two people talking for the ellipsis conditions. This gave participants an easy, fast, and non-linguistic way to identify the task without reading the full question. In addition, the full question was also presented every time below the image, in case participants needed to be reminded of the exact wording. For recognition, the full question was “Was this the sentence you just saw?”, while the ellipsis question was “Is this a reasonable continuation of the conversation?”.

Note that the ellipsis task actually functions as two conditions in one, because one version of the ellipsis condition asked participants whether (19a) is a valid continuation of the prime and separate trials probed the (19b) condition. The same is true of the recognition task.

Responses encoded old/new judgments and confidence combined on a single, ordinal scale which was the same across tasks. The scale had six-points, split between with *yes* (old/a good continuation) and *no* (new/not a good continuation), with *definitely*, *probably*, and *maybe* as sub-categories of the old/new judgment. Participants could respond using either the mouse or the number keys to indicate their answer on the yes/no/confidence scale (instructions slightly biased toward using the keyboard as a way to indicate an answer faster). Again, no feedback was given on responses.

The experiment was administered through the Ibex Farm online experiment platform. It was preceded by instructions and several examples that demonstrated the speeded presentation of the prime, the image cues for the two tasks, and the use of the response scale. Examples included both ‘*yes*’ and ‘*no*’ responses for the two tasks with explicit instructions pointing out that one answer was correct and another was incorrect. For both tasks, the example sentences used control versions of the tasks as examples.

Participants were also told that the sentences they would see were drawn at random from a larger set of sentences, and therefore they should not expect that the proportion of ‘*yes*’ and ‘*no*’ responses would be balanced across the experiment. This was to address the fact that while the number of *yes/no* responses was balanced across the critical stimuli, the correct response to the foils was always *no*, leading to a greater proportion of *no* responses in the overall experiment. This will likely lead to an overall compensatory *yes*-bias in our data. Our analysis will provide explicit ways to identify bias, therefore the unbalanced proportion is not deadly, but the inclusion of this instruction may help to mitigate some of the compensatory bias as well.

The experiment ended with a short debrief screen that asked participants to identify what type of machine they had used to take the experiment (laptop, desktop, etc), whether they had noticed anything about the sentences and whether they had run into any difficulties.

The experiment took approximately 45 - 55 minutes to complete, if participants worked straight through. However, because this was such a long experiment and distractions were anticipated due to the out-of-lab environment, participants were given breaks every 10 trials to accommodate distractions. These breaks could be as long as the participant chose.

### 2.2.2 Materials

One hundred and forty-four items were constructed according to the format in (18) and (19). All prime sentences used verbs which can take two animate arguments and were reversible such that the patient would be a suitable agent of the verb and vice versa, as in (18a) and (18b). Each prime had only one verb to ensure that voice remains consistent across an item. In addition, all passive stimuli included the agent by-phrase to maintain the greatest degree of lexical and semantic overlap.

Controls were relativized to the target tasks. For recognition, there were two main types of controls. The first of these simply substitute one word somewhere in the sentence for an alternative word with a related meaning, shown in (20). The second control type in (21) switched the agent and patient roles of the prime sentence, which changed the linear position and grammatical roles of the arguments like the active/passive alternation while also changing the semantic content of the sentence.

(20) The journalist was criticized by the politician over the presentation of the new treaty.

(21) The politician was criticized by the journalist over the presentation of the new bill.

Both of these controls serve to draw attention away from the active/passive manipulation so that participants must focus on encoding the whole sentence rather than developing a strategy of remembering whether a prime was passive or not. They also serve as baselines for ability to detect lexical and semantic changes respectively, which can then be compared to the critical manipulation of syntactic change.

The control for the ellipsis condition used a discourse particle which was incongruent with the context, as in (22).

(22) The TV pundit was anyway.

In total, the critical factors (Voice x2 levels; Task x2 levels; Lag [math problems] x3 levels; Outcome [Match/Mismatch/Control] x3) plus the control for ellipsis and two controls for recog-

nition as additional levels of Outcome resulted in 42 possible conditions.<sup>2</sup> In order to decrease the number of conditions and therefore increase the number of observations, each item was only ever presented in one voice across throughout the experiment. In addition, only one of the recognition controls was used for any given item. This reduced the total conditions per item down to 18. However, all items were constructed with the full set of controls. Moreover, the determination of which item would be presented in which voice with which recognition control was made immediately before running the experiment and assigned by alternating through the items, so that no bias influenced the construction of different sets.

### 2.2.3 Participants

Participants were recruited via the Prolific.ac online data collection system. Participants were paid \$6.90-7.00 for their participation. Prolific's own screening system sorts participants by demographic groups based on questions they answer before they take any studies. We used their system as a preliminary screening tool to allow only participants who had previously said that they had no language disorders, were between the ages of 18 and 60, and were native speakers of English. We also included questions within the experiment about language background in the experiment as an additional check to ensure that only native English speakers were included.

As a further measure to ensure quality, after the instructions participants took a short, five question multiple choice test on the most important parts of the instructions for task. Participants had to answer all five questions correctly to be included in data analysis, although they were still compensated for their time. In addition, there was a debriefing at the end of the experiment that asked participants if anything had stood out to them during the experiment. If participants indicated too great an awareness of the task, e.g., indicating that they had realized they needed track whether a sentence was passive or active, they could be excluded from analysis as well.

Data collection continued until there were 54 participants who met all the criteria.

### 2.2.4 Predictions

There are two competing hypotheses being considered in this experiment: the Privileged Access Hypothesis and the Regeneration Hypothesis of Potter and Lombardi (1990). If the Privileged Access Hypothesis in (17) is correct, then performance on the recognition task should decay rapidly as the delay between the prime and test increases, while the performance on the ellipsis

---

<sup>2</sup>Note that this count assumes that the two recognition controls each functioned as a separate condition, while ellipsis had only one control.

task should not exhibit such a steep forgetting curve. The Regeneration Hypothesis predicts no difference in the slope of the forgetting curves for the two tasks.

To see why the two accounts make these predictions, consider the status of syntactic memory between the two tasks. If Regeneration is correct, then it does not matter whether the task is ultimately ellipsis or recognition, there is no such thing as syntactic memory and under no circumstance will the comprehender have access to any long-term record of the structure. Therefore under Regeneration, the ability to discriminate between two response options on the basis of largely syntactic criteria should drop off quickly as the syntactic structure of the prime sentence leaves the Focus of Attention and ceases to exist. On the other hand, the PAH suggests that there is a memory trace for syntactic structure but only procedurally linguistic operations like ellipsis can access it. In this case, recognition will behave exactly as it would under Regeneration: as syntax leaves the Focus Of Attention, the mechanisms driving recognition lose access and experience a tremendous amount of “forgetting” over the course of a very short period. The linguistic mechanisms for ellipsis however, would maintain access to a memory trace of the syntactic structure, leading to a moderate amount of forgetting at a ‘normal’ rate, instead of the massive decrease in accessibility experienced by recognition’s mechanisms.

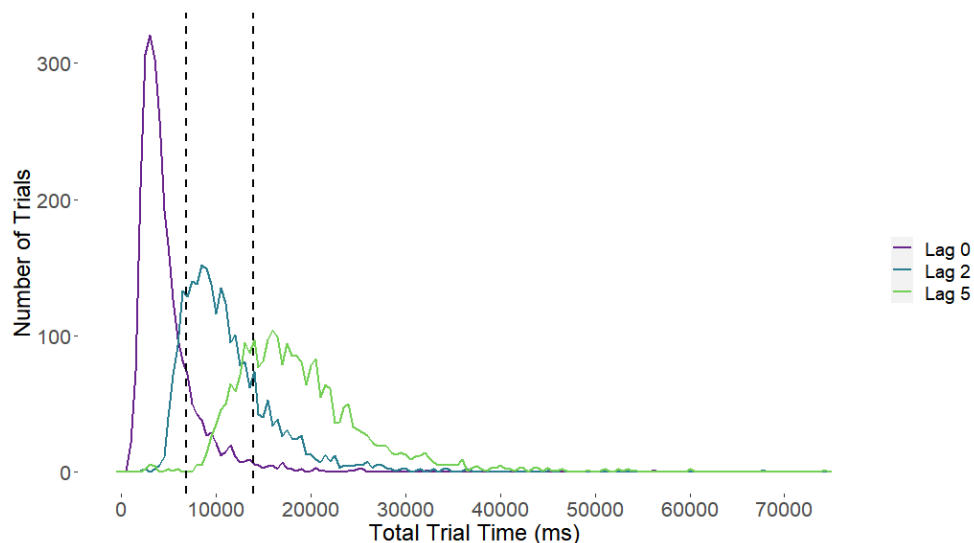
Additionally, the two tasks (though not necessarily the two hypotheses) have different predictions for the voice contrast. In the ellipsis task, the Voice Mismatch Effect is asymmetrical between the two voices and tends to find that active elisions on a passive base sentence are more acceptable than passive elision with an active original. Previous work, including work from recall, has argued that this because passive sentence are more likely to be retrieved as active than the reverse (e.g., Arregui et al., 2006; Mehler, 1963). This would predict higher degrees of sensitivity in the active-prime condition.

The predictions for recognition are somewhat less secure based on the previous literature. There is evidence that passive sentences are more likely to be amended to the active voice during retrieval than vice versa (for the same reasoning that was just applied to the ellipsis task), but passive involves additional, overt morphological cues which should be more memorable. For this reason, we might expect more active responses at lower levels of confidence but greater sensitivity to passive at levels of high confidence.

### **2.2.5 RT Data Processing**

Because this was an online experiment, and indeed a particularly long one, it is expected that participants were sometimes distracted during trials, even with the breaks that were built into the

experiment. This made outlier rejection of RTs particularly important. One trial was excluded for a technical error which caused the RT to be recorded as negative. Remaining trial response times (RTs) were calculated by aggregating the total time from the offset of the prime to the offset of the response screen when participants pressed a response key. Of these, 99.11% of trials took less than 60 seconds. Total trial RTs greater than 75000ms (or 1.25 minutes), were excluded as outliers which corresponded to 0.707% of all trials including foils. As with all outlier rejection, the concern in choosing this cutoff was balancing rejection of true distraction trials with keeping valid trials that simply happen to be long. In order to achieve that balance, it was assumed that true distraction was equally likely to occur on any given screen, which would suggest that true outliers should be roughly distributed among the lag conditions weighted by the total number of screens for those trials. If the outlier rejection were including too many long, valid trials, lag 5 should be over-represented as the cutoff point would be within the valid tail of that distribution. Given that RTs were calculated from the offset of prime screen, the number of screens contributing to the weighting for lag 0 was one (just the target screen), three screens for lag 2 (two math problems and the target) and six for lag 5 (five math problems and the target). Taking this distribution into account, with the 75000ms cutoff at lag 0 there was a 0.270% chance per screen that the trial would be excluded as an outlier, for lag 2 this was a 0.206% chance per screen, and for lag 5 the corresponding exclusion rate was again 0.206% chance per screen. The mean of the excluded trials was 458507ms and the standard deviation was 1504320ms.



**Figure 2.1.** Distribution of aggregate trial RTs by lag condition after outlier rejection. Bin size is 500ms. Dotted black lines indicate the quantile boundaries used in the analysis.

The second part of RT processing concerns how RTs were split into a categorical variable. Plotting the RTs, as in in Figure (2.1), shows considerable for overlap between the distributions of the lags. Therefore, instead of using the lags themselves as the levels of the categorical factor, RTs were split into three quantiles for the analysis. The boundaries for the quantiles are shown as the black dashed lines in the plot. Throughout the remainder of the paper, “*quantiles*” will refer to the levels of the Lag factor in both the ROC and Bayesian models, while lags 0, 2, 5 will be used when it is necessary to refer to the actual conditions that participants saw.

### 2.2.6 Results

The results are broken into two parts. The first part aggregates over voice and focuses on the comparison between the levels of the Task and Lag factors. The second part includes voice as a critical comparison factor. While the voice comparison is exploratory in the sense that each cell of the design necessarily only has half the observations of the aggregate analysis and lacks the power to be a primary analysis in this particular experiment, it is still of critical importance because both the memory and ellipsis literatures predict that voice should have an impact on the encoding or maintenance of sentences in memory. Thus the analysis is exploratory, but the predictions are not. Both parts make use of two different analysis techniques: i) Bayesian ordinal regression analyses and ii) ROC curve analysis from Signal Detection Theory (Macmillan & Creelman, 2004).<sup>3,4</sup>

For the ROC analyses: because of the nature of the experiment, we were unable to collect enough trials to calculate  $d'/d_a$  values for individual participants. We were therefore forced to aggregate data across participants for all of the analyses in this section. For all of the main analyses, unless otherwise noted, a Hit is a *yes*-response when the prime matched the voice of the test and a False Alarm is a *yes*-response when the prime and test mismatched in voice. Distractors are not included, unless specifically noted. Inferential tests for ROCs other than the Bayesian models were run with the pROC R package (Robin et al., 2011).

Bayesian regression models were computed using the brms package in R (Bürkner, 2018), and include random effects for subject and item. Unless otherwise noted, models consisted of four chains at 10,000 iterations each. The variable codings used in the models are given in Table (2.1).

---

<sup>3</sup>For Bayesian models, a random seed was generated through R and then saved so that models are replicable. This information is available on request.

<sup>4</sup>See Dillon and Wagers (2019) for an introduction to Signal Detection Theory intended specifically for psycholinguists and Pazzaglia et al. (2013) for an accessible introduction to modeling in SDT including a worked template.

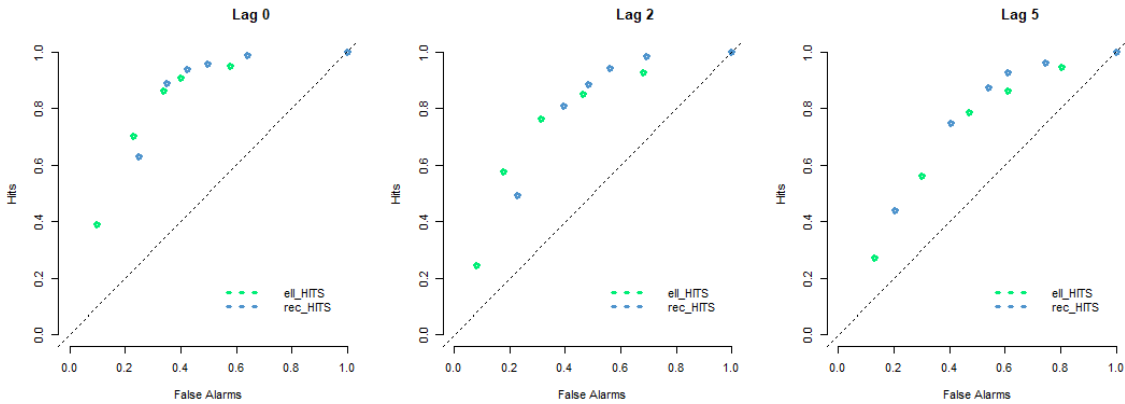
Variable	Level	Value
Lag (quantile)	<i>continuous</i>	quantiles 1,2,3
Match/Mismatch	Match	0.5
	Mismatch	-0.5
Task	Ellipsis	0.5
	Recognition	-0.5
Voice	Active	0.5
	Passive	-0.5

**Table 2.1.** Variable coding for Bayesian regression models.

Note that not all variables are included in every model. Tables from the Bayesian models were compiled with the help of the `shinystan` package from the `STAN` Development Team (2017).

### 2.2.6.1 Aggregated Over Voice

Figure (2.2) presents the ROC curves using just Task and Lag as factors and Table (2.2) presents the associated ROC analysis statistics for the curves.



**Figure 2.2.** ROCs for Experiment 1 by Lag, aggregated across Voice and participants. Note that the (1,1) point has been left in, but does not contribute meaningfully to the interpretation.

Most relevantly to the hypotheses behind the experiment, there is a consistent decrease in sensitivity measured by  $d_a$  as Lag increases, but the magnitude of the decrease is actually greater for ellipsis across the three levels of Lag (the three quantiles), with an overall decrease in sensitivity of 0.61 compared to recognition’s decrease of 0.423 over the same period. Given that the sensitivities for the two tasks appear to be converging by quantile 3, the larger decrease could be an artifact due to ellipsis starting from a higher level of sensitivity. Even if that were so, the fact remains that the direction of the effect runs directly counter to the predictions of the PAH, but can be accommodated by the Regeneration Hypothesis or Garnham and Oakhill (1987). On the other hand, the

higher starting sensitivity for ellipsis is a puzzle in its own right. More than any other piece of the analysis in Table (2.2), the high starting sensitivity for ellipsis could argue that grammatical operations have additional information to draw on that is not available to standard, declarative memory that recognition uses.

The  $d_a$  statistic that is reported in this analysis is already free of the assumptions of equal variance in the signal and noise distributions that was characteristic of  $d'$ . Still,  $d_a$  does retain a number of important assumptions, including that the signal and noise distributions follow a normal distribution (Verde, Macmillan, & Rotello, 2006).  $A_z$  is mathematically and interpretationally closely related to  $d_a$ , but without quite as many model assumptions (Zhang & Mueller, 2005), yet in the Experiment 1 data it tracks quite closely with  $d_a$ . This indicates that whatever is the driving force behind the  $d_a$  patterns, it is not dependent on the differences between parametric statistic  $d_a$  and the non-parametric  $A_z$ .

#### Aggregate Data ROC Analysis

	Lag (quantile)	$d_a$	zROC slope	$c_a$	$A_z$	$p$
Recognition	1	1.228	2.21	-6.129	0.807	0.017
Recognition	2	1.028	1.68	-4.160	0.766	0.067**
Recognition	3	0.805	1.38	-3.066	0.715	0.432**
Ellipsis	1	1.432	1.36	-1.759	0.844	0.002
Ellipsis	2	1.139	1.31	-1.012	0.790	0.006
Ellipsis	3	0.822	1.03	-1.194	0.720	0.042

**Table 2.2.** For all combinations of task and lag by quantile, the values for  $d_a$ ,  $c_a$ , the Area Under the Curve ( $A_z$ ). A  $p < 0.05$  rejects the hypothesis that the model is a fit for the data (Pazzaglia et al., 2013).

From Table (2.2), it is also possible to see that there is a substantial *yes*-bias in the data. Table (2.2) reveals several additional features of the bias. First, the *yes*-bias is heightened for recognition trials across the board far more than it is for ellipsis, based on the overall magnitude of the  $c_a$  values in Table (2.2). This is also visible in Figure (2.2) by the fact that each recognition point is typically up (higher Hit rate) and right (higher False Alarm rate) relative to its specific ellipsis counterpart. The  $c_a$  statistic marks the point on the evaluative dimension (in this case how familiar target task prompt is) where a decision switches from *no* to *yes*. A familiarity-score which is less than  $c_a$  will yield a *no* and a score above  $c_a$  will yield a *yes*-response. The very substantial, negative  $c_a$ -values for recognition will therefore induce an increased rate of *yes*-responses as more and more familiarity scores lie above the threshold, leading to a *yes*-bias. As with  $d_a$ , the higher starting values also pair with a larger decrease over time, although the  $c_a$  statistic does not converge across tasks as  $d_a$  did.

Second, the change over time in recognition's  $c_d$  values is much better behaved than the change over time ellipsis. Where the recognition values decrease consistently, the ellipsis values jump around. This could simply be noise as the  $c_d$ s in ellipsis are quite tightly grouped together relative to recognition. Alternatively, it may be one feature of the overall odd behavior of quantile 2, which is a running theme in these results and will be even more visible in the voice-level analysis.

The  $c_d$  values for recognition are in fact unusually large and it is surprising that they would be largest at quantile 1, when the most evidence is available. There were no predicted changes in response bias over time or intentional manipulation that would have pushed  $c_d$  so low. It might be tempting to attribute some of the *yes*-bias differences between tasks to the differences in the control conditions. Although the responses from the control conditions are not included in the analysis, the recognition control trials tended to be much harder to identify than the ellipsis controls (a changed lexical item or a change in the order of the words of the stimulus versus an ellipsis continuation that was pragmatically infelicitous). While this might have had an effect on the results in some way, the *yes*-bias actually goes in the opposite direction of what would have been expected if ease of identifying controls was the driving force. The easier ellipsis controls could have led participants to view the task as sorting the obviously pragmatically-illicit controls from the pragmatically-licit critical trials, with less attention paid to the voice of the critical trials. Yet this would have led to an extreme *yes*-bias in the ellipsis conditions, not the recognition task.

There are two further features of interest in this data which are worth mentioning. The first is the zROC slopes, which are notable because they exceed 1.00 even though elsewhere in the recognition literature zROCs almost universally remain  $<1.00$  (Ratcliff, Sheu, & Gronlund, 1992). The slope of the zROC is the ratio of the standard deviation of the noise distribution (in this case, the "evidence" suggesting that a target matches the voice of the prime, when it does not) and the signal distribution (the evidence that the target matches the voice of the prime when in fact it does), where the standard deviation of the noise distribution is standardly scaled to be 1. The zROC values above 1.00 can thus be interpreted as the standard deviation of the signal distribution being less than the standard deviation of the noise distribution. We have no firm explanation for this discrepancy from the previous literature, as this result was unexpected and the experiment was not designed to test reasons behind it. This pattern could be consistent with the idea that the cognitive processes of syntactic information are fundamentally different from the declarative information typically studied in memory studies. Nevertheless, there is far too little evidence here to actually make conclusions about why the zROC slopes are so steep.

There is only slightly more to say about the change in zROC slopes across quantiles. The fact that the zROCs decrease over time is an indication that either the standard deviation of the noise distribution is decreasing or the standard deviation of the signal distribution is increasing. Based on the available results, it is impossible to determine which of these is driving the change. Certainly, it is not a stretch to think that the variance of the signal distribution would increase with forgetting, as particularly well-encoded trials remained at the high end of the distribution and forgetting stretched less well-encoded trials into a longer and longer lower tail. It is less immediately clear why the noise distribution would narrow, as forgetting would presumably create more ways for the comprehender to be wrong. For the time being, that logic favors an increase in the width of signal distribution, but more evidence would be needed to confirm that interpretation.

The unusual zROC values (and also somewhat unusual  $c_a$  values) put the focus on a serious caveat in the interpretation of the ROC analysis, namely the model fits. Model fit is indicated by the  $p$ -values in Table (2.2). A  $p$ -value  $< 0.05$  indicates that the model is *not* a fully suitable fit for the data, which means that only two out of the six models in the table have converged successfully. We take lack of success as an indicator that the data does not fit all the process assumptions of the model. ROC models assume that there is a single which represents how familiar the item in memory is, which may be condensed from how familiar each of the item's features/cues are. The familiarity value is compared against one criterion ( $c_a$ ) at a time. If these process assumptions are not correct, then it may be that the ellipsis task in particular maybe more complex.

In addition to the ROC analysis, the data was also subjected to a Bayesian regression analysis in the form of a logistic model applied to the accuracy when the response variable was collapsed to just the *yes/no* binary response scale.<sup>5</sup> The Bayesian model takes on even more importance given the model fits in the ROC analysis, because Bayesian analysis yields the whole posterior distribution, or the estimate of all possible values of the slope and their relative probabilities. Having the entire posterior can sometimes help clarify the interpretation of effects that were marginal or confusing in models that yield single-value outputs (as the ROC analysis does).

---

<sup>5</sup>Confidence information was removed from the dependent variable for this model because it would have resulted in a parabolic relationship between the dependent variable and the predictors, especially Lag. (i.e., more high confidence responses at the extremes of the scale are expected with low lags, while trials with longer RTs should trend toward the low confidence responses in the middle of the scale.)

### Aggregate Data Bayesian Analysis

	$\hat{R}$	$N_{Eff}$	Mean	SD	2.5%	97.5%
Lag (quantile)	1.00	20000	-0.42	0.05	-0.52	-0.32
Match/Mismatch	1.00	15867	2.28	0.22	1.85	2.72
Task	1.00	15326	-0.42	0.22	-0.86	0.01
Lag x Match/Mismatch	1.00	20000	-0.26	0.10	-0.45	-0.08
Lag x Task	1.00	20000	0.09	0.10	-0.10	0.28
Match/Mismatch x Task	1.00	14970	-2.14	0.45	-3.02	-1.27
Lag x Match/Mismatch x Task	1.00	15872	0.33	0.19	-0.05	0.71

**Table 2.3.** Outcome of the Bayesian logistic model applied to the data when aggregated over voice of the antecedent. The model predicted accurate responses based on Lag (time by quantile), Match/Mismatch condition, and Task.

The first result of interest is the reliable negative main effect for Lag, indicating that the manipulation successfully produced forgetting. This is particularly notable given that the vast majority of trials had aggregate RTs under 25 seconds.

The main effect of Match/Mismatch was also reliable, indicating that participants were sensitive to whether the prime and target matched or mismatched in voice, but in a way that privileged matches. That is, participants were more likely to respond accurately when the prime and target did match (a “Hit”) but were less able to correctly reject a trial when the prime and target did not match. This pattern is indicative of the overall *yes*-bias that was also evident in the ROC analysis. The size of the posterior estimation of this effect is the largest in the model and rivaled only the interaction of Match/Mismatch x Task. The interaction is a reflection of the fact that Hits were more common in the recognition condition than ellipsis, while there were substantially more Correct Rejections/fewer False Alarms in ellipsis than recognition (even though Hits were overall still more prevalent than Correct Rejections for ellipsis). Again, this reflects the particularly large *yes*-bias and  $c_a$ -values in the ROC analysis of recognition. Looked at this way for the recognition task, this pattern is perhaps not surprising: the target will look almost exactly the same as the original sentence in either match or mismatch conditions, and will have neither the argument reversals or the lexical substitutions that participants would have come to expect from the foils. What *is* impressive is the ability of participants to reject an ellipsis continuation that mismatched the voice of the antecedent. This might be unexpected because recognition mismatch conditions are unarguably incorrect in the task while the Voice Mismatch Effect in VPE indicates that the mismatch conditions for the ellipsis task should have been at least marginally acceptable. Moreover, the morphological differences between the active and the passive are less dramatic in the ellipsis targets than recognition.

The Lag x Match/Mismatch interaction is much smaller but also technically excludes zero from the credible interval. This interaction corresponds to the decrease in  $c_a$  over time. As noted in the ROC analysis, the change over time in  $c_a$  is much smaller than the difference between the two tasks, thus the much smaller posterior estimate for Lag x Match/Mismatch than for Match/Mismatch x Task.

The last of the main effects was Task, and here the credible interval does include zero, although only at the very edge of the posterior distribution. This is a slightly different picture than the ROC analysis because in this case the ROC analysis provides multiple statistics ( $d_a$ ,  $c_a$ , and  $A_z$ ) where the Bayesian analysis provides only one. No underlying difference between the two tasks is a possibility, but the null effect is much less secure because the bulk of probability mass is associated away from zero. The negative direction of the coefficient argues that recognition had a higher accuracy overall, if the two tasks do differ underlyingly.

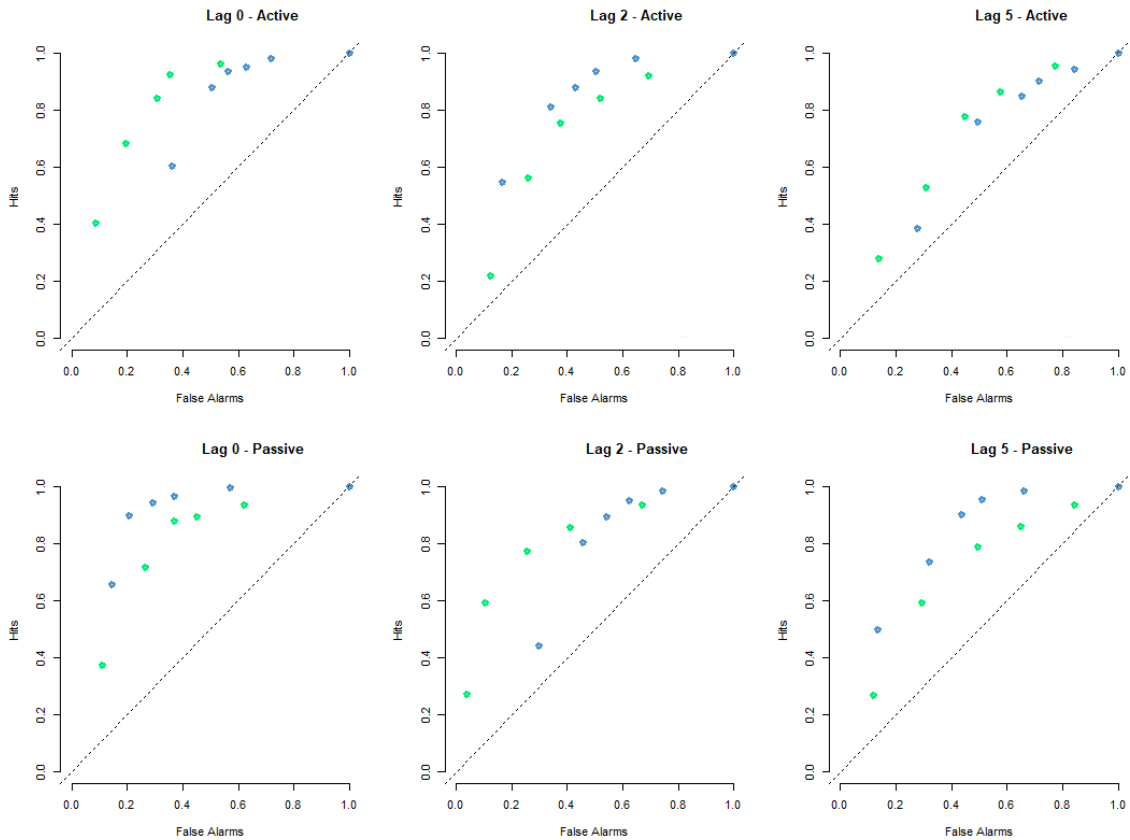
Most importantly, the interaction of Lag x Task is not reliable in the aggregate data, nor is zero particularly close to the tails of the probability density function. This was the critical interaction that was predicted by the PAH, but not by the Regeneration Hypothesis. It indicates that there is no reliable difference in the forgetting rates associated with recognition versus ellipsis. The way that the posterior distribution is situated about zero gives little indication about how to interpret it. In Bayesian posteriors, while a null effect is still ambiguous between the same two interpretations as in traditional NHST statistics, a posterior distribution which only included zero at the very edge (like the posterior for Task) can be seen as an indication that a false rejection<sup>6</sup> might be more likely, while a distribution centered on zero would bias interpretation toward a true null effect. This posterior distribution has neither of those characteristics, giving it the flavor of a fully ambiguous null result.

Given how much variance in this model is attributed to Match/Mismatch, another reasonable possibility could have been for the critical predicted effect to appear in the higher-order three-way interaction rather than the lower Lag x Task interaction. The three-way interaction is also not reliable, although here zero is much closer to the edge of the CI, raising the very serious possibility that the effect exists but is simply very small relative to the sensitivity of this experiment. With this in mind, we turn to the results split by voice of the antecedent, to see whether there are sub-patterns to the data that could meaningfully inform the interpretation of the critical interactions.

---

<sup>6</sup>Or its equivalent, as “rejection” really implies the binary decision-making of of NHST and not the gradient probability allocation over hypotheses that is Bayesian decision making.

### 2.2.6.2 Voice analysis



**Figure 2.3.** ROCs for Experiment 1 split by Voice and Lag. Note that the (1,1) point has been left in, but does not contribute meaningfully to the interpretation.

Figure (2.3) shows the same data as Figure (2.2) but with the addition of the voice contrast. Tables (2.4) and (2.5) provide the new ROC analysis fits split on the voice contrast. This view demonstrates that the aggregate picture in Figure (2.2) potentially glosses over important internal structure in the data.<sup>7</sup>

For instance, at quantile 1 the ellipsis task had high sensitivity to active antecedents, but in the recognition task passive antecedents boosted sensitivity. Notably, this switch is not driven by a change in just one of the tasks, because they appear to differ in initial sensitivity between the two voice conditions. That is, the sensitivity for ellipsis is higher in the active than in the passive *and* recognition has higher sensitivity in the active than the passive (where either one of these shifts

---

<sup>7</sup>A reminder: all of the results in this section are reported with the understanding that they are preliminary results only due to lack of power.

would have been enough to reverse the relative order). This change roughly comports with the predictions for voice.

#### Active ROC Analysis

	Lag (quantile)	$d_a$	zROC slope	$c_a$	$A_z$	$p$
Recognition	1	0.880	2.12	-6.826	0.733	0.124**
Recognition	2	1.038	1.40	-3.360	0.769	0.211**
Recognition	3	0.606	1.31	-3.252	0.666	0.146**
Ellipsis	1	1.412	1.60	-1.646	0.841	0.006
Ellipsis	2	0.863	1.52	-1.137	0.729	0.025
Ellipsis	3	0.837	1.08	-1.339	0.723	0.570**

**Table 2.4.** For all combinations of task and lag by quantile in the Active voice, the values for  $d_a$ ,  $c_a$ , the Area Under the Curve ( $A_z$ ). A  $p < 0.05$  rejects the hypothesis that the model is a fit for the data.

#### Passive ROC Analysis

	Lag (quantile)	$d_a$	zROC slope	$c_a$	$A_z$	$p$
Recognition	1	1.661	2.18	-5.030	0.880	0.520**
Recognition	2	0.997	2.08	-5.383	0.760	0.063**
Recognition	3	0.997	1.47	-2.889	0.760	0.268**
Ellipsis	1	1.463	1.18	-1.844	0.850	$9e^{-4}$
Ellipsis	2	1.425	1.11	-0.816	0.843	0.158**
Ellipsis	3	0.800	1.01	-1.069	0.714	0.010

**Table 2.5.** For all combinations of task and lag by quantile in the Passive voice, the values for  $d_a$ ,  $c_a$ , the Area Under the Curve ( $A_z$ ). A  $p < 0.05$  rejects the hypothesis that the model is a fit for the data.

By quantile 3 these patterns clearly begin to degrade across both voices, as the sensitivity decreases with forgetting. Interestingly, recognition in the passive still retains an advantage over recognition in the active and ellipsis in the passive, suggesting that the advantage from the markedness of the passive may persist better in the recognition task. On the other hand, it could also be a reflection of the fact that sensitivity for the passive in the recognition task starts so high, and the absolute magnitude of forgetting is approximately the same (though see the Bayesian analysis for an argument against this interpretation).

Two of the Task and Voice combinations, ellipsis in the active and recognition in the passive, follow the pattern for sensitivity metrics that might have been most expected. For these two combinations,  $d_a$  starts off high, has a large drop between quantile 1 and quantile 2 and then a smaller drop from quantile 2 to quantile 3. This is the pattern that would be expected if quantile 1 were benefiting from lingering information in the Focus of Attention and therefore using qualitatively different processes. (Alternatively, Regeneration’s system of syntax only existing during online

processing and disappearing immediately after would produce the same pattern.) Quantiles 2 and 3 would rely on a very different process to evaluate the target task and would pattern together. This would be a fairly sensible underlying process set, but it is unclear why it would not apply to all four combinations of Task and Voice, and why these two would be the ones that did use it.

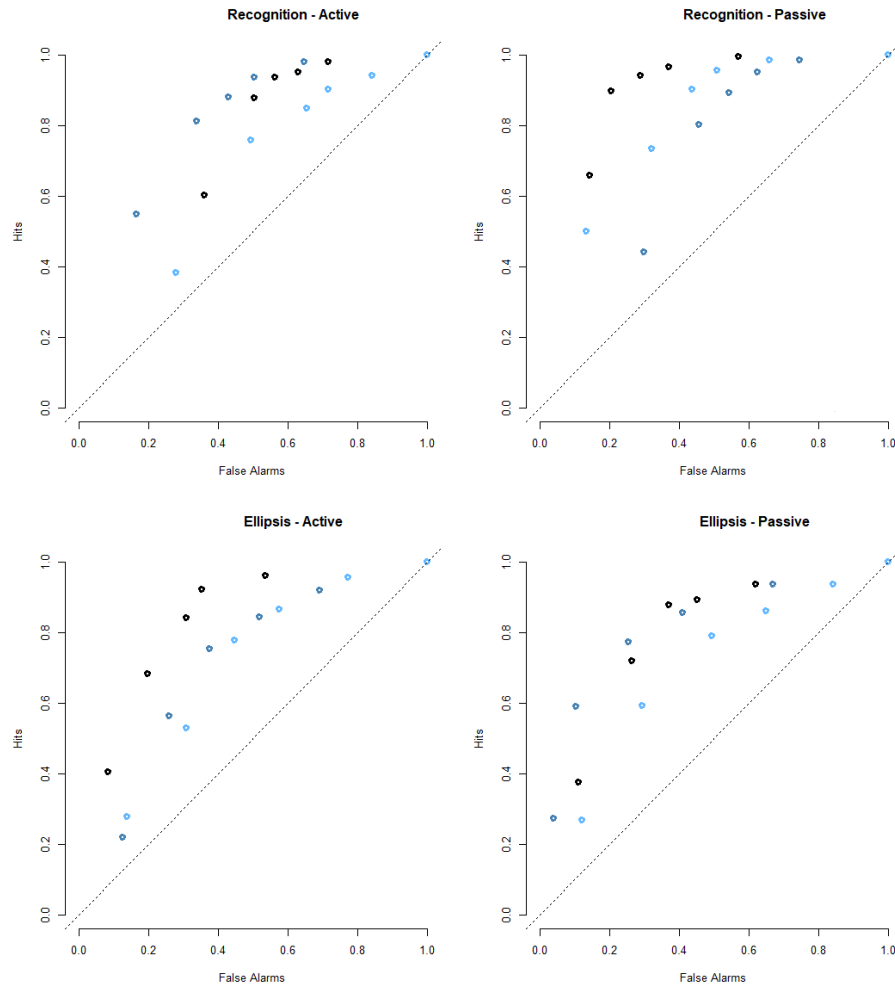
The other two Task and Voice combinations, recognition in the active and ellipsis in the passive, are part of the highly unexpected behavior of quantile 2 in this experiment. Figure (2.4) provides some insight into the contribution of Lag 2 by showing the same data as in Figure (2.3), but reconfigured so that the lags can be directly compared. Figure (2.4) confirms that while the behavior of Lag 0 and Lag 5 is always consistent with endpoints on the forgetting curve, Lag 2 is generally more erratic.

In both the active recognition conditions and passive ellipsis conditions, and quantile 2 disrupts what would otherwise be a reasonable forgetting curve. The most dramatic case is active-voice recognition, where  $d_a$  for quantile 2 jumps much higher than either quantile 1 or 3. Without further information it is difficult to know how to interpret such a change over time, but noise or a violation of model assumptions seems likely. The jump in  $A_z$  at quantile 2 for active recognition is relatively small compared to the corresponding  $d_a$ . That could indicate that a violation of the normality assumption of  $d_a$  is partially to blame for the odd behavior, although  $p$ -values indicate that all the recognition models are at least plausible fits for the data.

A secondary instance of the odd behavior in quantile 2 is the response criterion  $c_a$  in the passive ellipsis conditions. The quantile 2  $c_a$  value of -0.816 is well out of step with quantiles 1 or 3 and implies that the *yes*-bias weakens considerably in the middle of the scale before dropping back to a stronger bias value. The aggregate data analysis argued that  $c_a$  was a more reasonable statistic to see readjustment if there was a qualitative process change between quantile 1 and quantile 2. If the quantile 2 value is set aside, the remaining  $c_a$  values appear roughly comparable to the values in the aggregate data and should have the same interpretation as they did there.

Interestingly, the zROC slopes at quantile 2 never show the odd behavior of the other statistics, arguing that whatever processes disrupt the transition from quantile 1 to quantile 3, they do not implicate the standard deviations of the signal or noise distribution.

One thing that is more clear in the voice analysis is that across the board, the ROC models fit the recognition task better than ellipsis, as demonstrated by the  $p$ -values. Considering that Signal Detection Theory was developed to analyze recognition data, that is a perfectly coherent outcome.



**Figure 2.4.** ROCs for Experiment 1 split by Voice and Task. Note that the (1,1) point has been left in, but does not contribute meaningfully to the interpretation.

It may indicate that the process model assumptions of the ROC analysis are not as good a fit for the ellipsis task as they are for recognition.

Because the ROC and Bayesian models focus on different aspects of data structure, even with the new view of the internal data structure from the voice contrast in the ROC analysis, many of the conclusions from the aggregate Bayesian models carry into the voice-specific data. In both of the Bayesian voice models, the main effect of Lag is fully reliable and negative, indicating that the forgetting manipulation worked.

### Active Bayesian Analysis

	$\hat{R}$	$N_{Eff}$	Mean	SD	2.5%	97.5%
Lag (quantile)	1.00	34000	-0.40	0.07	-0.54	-0.27
Match/Mismatch	1.00	19563	2.31	0.31	1.71	2.93
Task	1.00	17970	-0.20	0.31	-0.82	0.41
Lag x Match/Mismatch	1.00	20400	-0.24	0.13	-0.51	0.02
Lag x Task	1.00	18847	0.06	0.14	-0.21	0.32
Match/Mismatch x Task	1.00	17815	-2.79	0.62	-4.03	-1.59
Lag x Match/Mismatch x Task	1.00	18607	0.58	0.27	0.06	1.11

**Table 2.6.** Outcome of the Bayesian logistic model applied to Active voice antecedents only, predicting accurate responses based on Lag (time by quantile), Match/Mismatch, and Task.

### Passive Bayesian Analysis

	$\hat{R}$	$N_{Eff}$	Mean	SD	2.5%	97.5%
Lag (quantile)	1.00	34000	-0.46	0.07	-0.61	-0.32
Match/Mismatch	1.00	34000	2.35	0.34	1.70	3.02
Task	1.00	34000	-0.64	0.34	-1.31	0.01
Lag x Match/Mismatch	1.00	34000	-0.30	0.14	-0.58	-0.02
Lag x Task	1.00	34000	0.12	0.14	-0.16	0.40
Match/Mismatch x Task	1.00	34000	-1.71	0.68	-3.04	-0.41
Lag x Match/Mismatch x Task	1.00	34000	0.04	0.29	-0.51	0.61

**Table 2.7.** Outcome of the Bayesian logistic model applied to Passive voice antecedents only, predicting accurate responses based on Lag (time by quantile), Match/Mismatch, and Task.

The overall largest effects continue to be the main effect of Match/Mismatch and the corresponding interaction of Match/Mismatch x Task, reflecting the *yes*-bias and the fact that  $c_a$  values are generally high and even higher for recognition than ellipsis across both voices.

Technically, the Lag x Match/Mismatch posterior straddles zero in the active but not the passive. Nonetheless, the posterior estimates across the active, passive, and aggregate data are all so similar that the active and passive most likely still represent the same underlying process with only some noise to separate them.

As in the aggregate data, the posterior of the critical interaction of Lag x Task interaction is fairly tightly clustered around zero, contrary to the PAH but consistent with the predictions of Regeneration.

The two largest differences in the models between the aggregate data and the data split by voice are the main effect of Task in the active and the three-way interaction of Lag x Match/Mismatch x Task. The posterior of the main effect of Task is centered around zero in the active where in the passive and aggregate data the bulk of the probability mass was associated away from zero. This may reflect the same noisiness that was noted in the active recognition  $d_a$ 's in the ROC analysis. The behavior of quantile 2 and the relatively low  $d_a$  for quantile 1 suggest that something more

complicated is behind the active recognition levels than would easily be captured as a single-level of a main effect.

Lastly, the reliable three-way interaction of Lag x Match/Mismatch x Task with active antecedents but not passive ones clarifies the interpretation of the three-way interaction in aggregate data, where it was not reliable in an NHST sense, but zero was quite close to the edge of the posterior distribution. In the split data, it is clear that this interaction presents like a true null in the passive, but a fully reliable interaction the active voice. However, the descriptive data in Figure (2.3) and the  $d_a$  values in Table (2.4) indicate that the difference is more likely attributed to the unexpected behavior of quantile 2, rather than the PAH's predicted difference in forgetting rates between the two tasks.

### 2.3 Discussion

The results of Experiment 1 indicate that the Privileged Access Hypothesis is an unlikely fit for the data. Mostly importantly, ellipsis starts off with a higher overall accuracy than the recognition task, but the two tasks decay to essentially the level of sensitivity with a very short period of time ( $\sim 30$  seconds). To get this pattern, accuracy rates actually decay faster in ellipsis than in the recognition task, indicating that ellipsis has a higher rate of syntactic forgetting. This is the opposite of the direction predicted by the PAH, but it could be predicted under certain assumptions of the Regeneration Hypothesis or Garnham and Oakhill (1987)'s hypothesis.

Intriguingly, there are still features of the data that point to a qualitatively different underlying decision-making process for ellipsis, just not the one described in the PAH. The first of these is the changes needed in the Regeneration Hypothesis to accommodate the higher starting sensitivity for ellipsis. That problem is that even immediately after the sentence, when the Regeneration Hypothesis says that syntactic information *should be* generally available, recognition has more difficulty making use of this information than specifically grammatical tasks. This could be most easily accommodated with the very PAH-like caveat that ellipsis has privileged access to syntactic information at the transition from online to offline processing. Specifically, syntactic information is prioritized by the parser initially, but then quickly disappears such that by quantile 3 retrieval can only rely lexical and semantic information that recognition has access to. This is in some sense a minor change in the priorities of the parser from the point of view of the Regeneration Hypothesis, but implies a special status for syntactic information when it is available. That said, this difference may not be entirely reliable, as the Task x Lag interactions in the Bayesian analysis

overall indicate small to non-existent effects and the behavior of the Task x Match/Mismatch x Lag interaction is only slightly better.

The second potential indicator that ellipsis was qualitatively different from recognition was the poor ROC model fits. The voice analysis showed that the poorer fits were concentrated on the ellipsis data, while the model fits in recognition were acceptable. The ROC analysis makes strong assumptions about the processes that underlie the data, and this process-model has been tailored to recognition memory. The poor fits across the board for ellipsis suggest that the ROC process model is not right for the ellipsis task. Interestingly, the Regeneration Hypothesis probably predicts that the ellipsis decision process should be identical to recognition. In Regeneration, standard recognition (or recall) memory is the only mechanism that the cognitive system can call on, and it does so to retrieve lexical and semantic information. As both the recognition and ellipsis tasks would have the accessible information and the same cognitive mechanism, there would be no reason for the decision process for the two task to be qualitatively different. The Garnham and Oakhill (1987) hypothesis may provide a way forward: their hypothesis was that syntactic information was checked as a supplement to more reliable lexico-semantic information. Though this hypothesis does not require two separate checking mechanisms for syntactic versus lexico-semantic information, it is compatible with them. Two separate familiarity scores would certainly lead to a distinct process model that the ROC analysis could struggle to fit.

There are a number of open mysteries in this data. One mystery is the odd values for the zROC slopes, and to a lesser extent the  $c_a$  values. The results section suggested that this could be another byproduct of the sometimes poor model fits in the ROC analysis, which is in turn likely indicative that the data somehow violates the process assumptions of the model. However, in this case the model fits can only be a partial answer, because the higher  $d_a$  and most negative  $c_a$  values were in recognition, where, again, the model fits were overall satisfactory.

Another mystery is the behavior of quantile 2. In several combinations of Task, Match/Mismatch, and Voice, the Hit and False Alarm values for quantile 2 do not lie in between the values for quantile 1 and quantile 3, as would have been expected of a forgetting process. The direction of the deviation is not always the same: in Figure (2.3) for instance, the passive recognition conditions, much of the quantile 2 ROC curve for the passive recognition conditions appears to be below both the quantile 1 and quantile 3 curves, while corresponding ROC of the ellipsis task is largely above both of its quantile 1 and quantile 3 counterparts. First, there is the very real possibility that this is noise. It is extremely odd that a participant would have better performance in a task after they had completed two intervening math problems in lag 2 compared to when they completed the

task immediately after the initial presentation as in lag 0. It may simply be that two math problems were not sufficient to produce reliable decay and therefore a true mid-point on the forgetting curve. If, for instance easy math problems did not produce real decay and hard math problems were sufficient to result in actual forgetting, a participant would be left in a lag 0-like state on some lag 2 trials, and a more lag 5-like state on others. If this *is* a real effect, then it is a true puzzle and we have very little explanatory to say about it at this point, except to point out that the statistic that tracks this behavior most closely, though by no means perfectly is  $c_a$ , the response bias statistic. All theories currently on the table suggest that the transition from lag/quantile 1 to lag/quantile 2 could involve at least a partial qualitative change as the input from the antecedent sentence passes from the Focus of Attention to a later state of offline processing. Of the parameters in the model as it stands, the response bias might be a reasonable candidate to be influenced by such a transition.

Additionally, the predictions for the voice contrast were largely not borne out. Discrimination for the ellipsis task was no better when the original prime was an active than when it was passive, even though this would have been the reflex of the well-attested Voice Mismatch Effect. As noted before, the voice contrast in this experiment was exploratory only due to the power for either individual voice condition and therefore any interpretation should be postponed for a full power replication of these contrasts.

In addition to discriminating between the PAH and Regeneration/Garnham and Oakhill hypotheses, this experiment provides tangential support for theories of ellipsis acceptability based on memory decay and regeneration (e.g., Arregui et al., 2006), insofar as the outcome predicted by the PAH would not have been as compatible with the forgetting as the driving force between acceptability variation within ellipsis. Critically, Arregui et al. (2006) argue that even though parts of the syntactic structure may have been forgotten, syntactic structure is still necessary. That is, some formal theories of ellipsis antecedent resolution rely more on the semantics (e.g., Kehler, 2002), but Arregui et al. conclude that those theories would not be consistent with their findings. It could be tempting to combine a semantic theory of ellipsis with Regeneration-style syntactic forgetting and assume that syntax is simply not used in elliptical processing. Arregui et al. argue that syntax is necessary to process ellipsis, whether it is a slightly degraded form due to forgetting or regenerated from the semantics and lexical input.

In order to resolve the interpretation of this experiment, including lag 2, the next step needs to be clarification of how much these results are dependent on the specifics of the task. A follow up experiment would ideally make two changes. First, the time between the lags should be extended.

Second, the intervening task should be made to be linguistic, as there is evidence dating back Sachs (1974) that memory for linguistic material may be more susceptible to linguistic intervening tasks. This might mean that the intervening task could be a short list task, or alternatively the experiment could be re-configured into a version of the N-back task (Kirchner, 1958). Only the results which replicated in such a follow-up should be taken as secure.

As discussed in Chapter 1, this experiment is not new for finding that syntactic structure disappears quickly from explicit memory. What *is* newer is that this experiment was also designed to test for a type of non-explicit/declarative memory that could have facilitated syntactic processing even after traditional declarative memory is lost. If the Privileged Access Hypothesis had been supported, it would have provided a way to theoretically unify the memory-based accounts of within-sentence processing that are the subject of Chapter 3 with the findings of disappearing syntax from sentence recall and recognition studies, as well as the long-term memory implied by abstract syntactic priming that will be the focus of Chapters 4 and 5. It also would have green-lighted projects to adjust the cue-based structure of modern memory theories to accommodate hierarchical structure, which these theories do not easily do in their present incarnations (Kush, 2013; Lewis & Vasishth, 2005). But the complicated process of adapting memory theories to include hierarchical syntactic structure is only worthwhile if there is evidence that syntax exists in memory, and at present that evidence is limited.

## CHAPTER 3

### GRAMMATICAL INFLUENCES ON MEMORY OPERATIONS DURING ONLINE PROCESSING

#### 3.1 Introduction

Chapter 2 presented new evidence that syntactic structure in memory behaves like declarative memory regardless of the reason it was accessed, based on rates of decay. However, different mechanisms or pathways for memory access might differ on several dimensions, of which rate of decay is only one.

Dillon (2011) and Dillon, Mishler, Sloggett, and Phillips (2013) proposed another possible division between memory access profiles in language, this time in online comprehension. Dillon and colleagues exploited the prediction of memory intrusion made by the cue-based memory models which were introduced in Chapter 1. To review, these models view each item in memory as a collection of features, for example [+noun, +plural, +human] etc. Retrieval of a particular item from memory begins with a probe consisting of several cues as required by the situation which prompted retrieval in the first place. For instance, on encountering the pronoun *her* in (23) the processor would likely try to recall the individuals in the discourse which might be reasonable antecedents. The pronoun provides a *probe*, or a set of cues which the retrieval will need to satisfy, in this case: [*animate, female, in – common – ground, given – in – discourse*].

- (23) a. Odessa said that Cody had hugged her.  
b. Sean said that Cody had hugged her.

The retrieval process then finds the items within memory that are the best fits for the probe. While items which match all the probe cues will have a distinct advantage (*Odessa* in [23a]), the cue-matching retrieval system also has the possibility of returning partial matches. If, as in (23b), there is no individual in the discourse who fits all the required cues, the comprehender might consider a female who was in the common ground for both interlocutors but not in the current discourse, or even the possibility that the speaker made a speech error and intended a male antecedent that was both in the common ground and salient in the discourse. This system has the

advantage of flexibility, which allows it accommodate errors or other surprises without a complete crash.

The fact that this retrieval mechanism could in principle return several different items, so long as their features largely match the retrieval probe's cues, introduces the potential for interference between items with similar features. And it suggests that the presence versus absence of interference is a test for cue-based retrieval as a specific kind of retrieval mechanism being used in comprehension. Prior to Dillon et al. there was already considerable evidence for interference which fits this description in subject-verb agreement dependencies. Wagers (2008) and Wagers, Lau, and Phillips (2009) proposed that cue-based retrieval interference could account for the comprehension version of a phenomenon known as agreement attraction (Bock & Miller, 1991). The agreement attraction pattern in comprehension is demonstrated by the pair of sentences in (24).<sup>1</sup>

- (24) a. The key to the cabinet are on the table.  
b. The key to the cabinets are on the table.

Both (24a) and (24b) are ungrammatical, yet Bock and Miller found that speakers are more likely to produce the ungrammatical form of the verb (*are*) in (24b) than in (24a). Sometimes known as a mismatch paradigm, the key features of this manipulation are an ungrammatical morphological dependency (here between the number features of subject and the verb) and a lure noun which either matches or mismatches the morphological features of verb. The presence of the plural lure noun, *cabinets*, in (24b) appears to lead the producer to accidentally plan a plural-marked verb. In this case, and perhaps in most cases, the lure noun is linearly between the subject and the verb, but this is not a requirement for attraction to occur (Wagers et al., 2009).

Agreement attraction in production has a number of correlates in comprehension, including increased acceptability and shorter reading times for (24b) (Wagers et al., 2009). A curious feature of agreement attraction which is apparent in comprehension, though difficult to observe in production, is that there is a grammatical asymmetry. That is, although the ungrammatical sentences in (24) are influenced by intrusion from the number feature on *cabinet(s)*, the grammatical counterparts in (25) do not typically show any sensitivity to intrusion. In comprehension, this would have been expected to manifest as lower acceptability or longer reading times for the sentence in (25b) with a plural lure noun than for (25a) where the singular lure noun matches the singular

---

<sup>1</sup>For the sake of clarity, this chapter will use *intrusion* to mean the hypothesized memory retrieval phenomenon, *attraction* to refer to the data pattern which is invoked as evidence for intrusion, and *lure* to refer to the noun that causes attraction in the mismatch design paradigm. While all of these terms are used the literature, many other papers use different mappings.

verb. However, the majority of studies don't find any such difference between the grammatical sentences, and certainly not with the reliability of the contrast in the ungrammatical sentences.

- (25) a. The key to the cabinet is on the table.  
b. The key to the cabinets is on the table.

Recent work by Hammerly, Staub, and Dillon (2019) suggests that the grammatical asymmetry may be at least partially methodological and that the examples in (25) may show sensitivity to intrusion when comprehenders have a reduced *a priori* bias to expect grammatical sentences. Nevertheless, in most studies the traditional hallmark of agreement attraction is an interaction of Grammaticality x Intrusion such that the ungrammatical, matching lure condition in (24a), is substantially more difficult to comprehend than the other three conditions.

Wagers (2008) and Wagers et al. (2009) suggested two separate ways that memory retrieval could produce the agreement attraction pattern, both of which rely on a retrieval operation which is triggered at the verb in order to verify the form of the agreement morpheme. In one of the two proposals, retrieval is triggered only when the initial comprehension processes detect a mismatch between the previously encoded subject noun and the incoming agreement morpheme at that verb, i.e., (24). In the other proposal, retrieval is always triggered at the agreement morpheme. Only the first of these strongly predicts the grammatical asymmetry, but both predict the original contrast in the ungrammatical sentences.

Regardless of the specific proposal, the probe for this retrieval operation must have a set of cues which is approximately [+NOM, +NOUN, +PL]. In the sentences in (24), agreement is ungrammatical because the subject head noun, *key*, does not actually have the [+PL] feature to satisfy the probe. But *cabinets* does have a [+PL] feature, and by virtue of being another noun inside the subject phrase will have many overlapping features in common with the subject head noun and therefore the probe. This causes retrieval interference such that the presence of a plural feature makes the ungrammaticality seem more acceptable, even though the number feature isn't hosted on the right noun.

There is nothing about the logic of either of the retrieval-interference proposals in Wagers et al.'s account that is restricted to agreement in particular. Based on their account for (24), the same pattern of results would be expected to hold for any feature-matching between two lexical items separated by some amount of intervening material. To the point of this chapter, another dependency which fits this description is the relationship between a reflexive anaphor and a preceding antecedent. Yet despite the fact that agreement and reflexives appear to present

similar feature-matching problems for the parser, attraction has been far more elusive for reflexives than for agreement. This was what led Dillon et al. (2013) to suggest that different grammatical constructions recruit memory in fundamentally different ways.

First, consider why agreement and reflexives might be predicted to behave similarly as memory retrieval operations. Both agreement and reflexive anaphors involve two elements which *must* have matching  $\phi$ -features, realized morphologically and this means that they have many similarities whether they are compared in either formal grammatical or memory theories. In a Minimalist framework, both subject-verb agreement and reflexive anaphora are underlyingly implemented by *AGREE*, one of the two atomic operations in Minimalism. *AGREE* requires that there is a c-command relationship between the two elements of the dependency, as the subject c-commanding the verb or correspondingly the antecedent c-commanding the reflexive. (Frequently, of course, the antecedent in the anaphor dependency is also a subject.)<sup>2</sup> At a grammatical level, these dependencies have a considerable amount in common.

Furthermore, what matters from a memory perspective is that in both dependencies the two elements the two elements are part of different constituents and can be linearly and syntactically distant from each other. Since decay of syntactic memory is so rapid, this distance is enough that the first element will (at least sometimes) no longer be in the focus of attention when the second element is, and will therefore need to be retrieved from memory (McElree et al., 2003). It may also be important that the contentful versions of the features —that is, the lexical item that is actually plural or has gender derived from event/world knowledge —is first, although psycholinguistics has lacked access to a sufficient range of long-distance dependencies of this sort to thoroughly test for the effect of order. From the perspective of memory retrieval over linear order, then, agreement and reflexives present the parser with nearly identical parsing problems: obligatory  $\phi$ -feature-matching over distance.

Differences arise because agreement and reflexive anaphors are governed by potentially very different grammatical functions. *AGREE* is standardly motivated for subject-verb agreement by the need for the agentive NP to receive case and to facilitate raising of the subject to SpecTP if raising will happen in that sentence. In English and many other languages, subject-verb agreement holds between two very specific positions in the syntactic tree: the head of the NP complement

---

<sup>2</sup>*AGREE* and memory theories share the notions of probing and features, but implement them in different ways. Most notably, for *AGREE* the probe is in the canonical subject position, SpecTP, and probes down. Memory theories on the other hand locate the probe at the verb and suggests that the direction of probing is backwards. What matters for the current purposes is not the absolute difference between the two theoretical implementations, but rather that under either framework subject-verb agreement and reflexive anaphora receive roughly the same treatment.

in SpecTP and the highest verbal head in the VP projection (auxiliary or main verb). Reflexive anaphors, however, are a binding relationship under Principle A, which roughly states that a reflexive anaphor must be bound by an antecedent within its binding domain. This means that unlike agreement, even though antecedents frequently turn out to be subjects, anaphora cannot reliably look to a single position in the tree and must instead consider what positions are available in a particular domain. The precise conditions on size of a binding domain is still a matter of some debate, but in the sentences that are at issue in this chapter it will be sufficient to treat the binding domain as a clause.

With so many properties in common, grammatically and in the linear form given to incremental processing, there are good *a priori* reasons to believe that the parser might apply the same or substantially similar algorithms to resolving agreement and reflexive morphological dependencies. If there is a difference in how sensitive the two dependencies are to intrusion, then it is most likely attributable to the grammatical differences between agreement and Principle A. The intrusion profile for agreement is well-attested in the form of agreement attraction, indicating that a cue-based retrieval account might be fruitfully applied to reflexives. On the other hand, while there is a considerable literature on reflexive anaphors in the configuration as agreement attraction, a truly robust and reliable pattern of intrusion to rival the one from agreement has yet to emerge.

The lack of interference for reflexives might be predicted if reflexive processing did not rely on memory retrieval, but this does not seem to be the right explanation. Evidence that memory does play a role in reflexive processing comes from work done by Nicol (1988) and which is summarized in Nicol and Swinney (1989). This study used a cross-modal priming task where (26) was presented auditorily and then interrupted at the \* for a visual lexical decision on a word that was semantically associated with *boxer*, *skier*, or *doctor*.

- (26) The boxer<sub>i</sub> told the skier<sub>j</sub> that the doctor<sub>k</sub> for the team would blame {*himself<sub>k</sub> / him<sub>i/j</sub>*} \*  
for the recent injury.

When the test point occurred after a reflexive, the response time for lexical decision with the associate of *doctor* had a 104ms advantage over the average response times for the lexical decision task. This would be expected if *doctor* were active at the test point, and therefore able to semantically prime its associate. However, when the sentence included a pronoun instead, the advantage for the associate of *doctor* disappeared and the associates of *boxer* and *skier* had a semantic priming advantage instead. This is precisely what would be predicted by the parser

trading off applying Principle A and its sister constraint, Principle B, which requires that all pronouns are free in their binding domains. The fact that the semantic associate of *doctor* was not facilitated in the pronoun condition demonstrates that *doctor* was not simply maintained over the course of the sentence, but rather that it must have been retrieved at the reflexive. In other words, on reaching the reflexive, the parser retrieved the only possible antecedent, *doctor*, allowing it to be sufficiently active to semantically prime in the lexical decision task. Thus, while Nicol's results do not address the details of the retrieval mechanism itself, they do demonstrate that memory retrieval is a necessary component of parsing reflexive anaphora over sufficient distances.

One of the first studies to specifically address interference for reflexives was Sturt (2003), which compared pairs of sentences based on those in (27) and (28). Each pair includes of a context sentence that introduced one of the referents for the (a) and (b) test sentences. In the test sentences themselves, Sturt used a paradigm based on the mismatch paradigm for agreement attraction in (24) and (25)

(27) Jonathan was pretty worried at the City Hospital.

- a. The surgeon who treated Jennifer had pricked himself with a needle.
- b. The surgeon who treated Jonathan had pricked herself with a needle.

(28) Jennifer was pretty worried at the City Hospital.

- a. The surgeon who treated Jennifer had pricked himself with a needle.
- b. The surgeon who treated Jennifer had pricked herself with a needle.

Sturt found no facilitation for (28b) —the intrusion condition that corresponds to (24b)—in early eyetracking measures, but did find an interference effect in very late measures (second pass reading). In Sturt's own assessment, this is indicative that “[binding] constraints were applied extremely early” in the course of processing (Sturt, 2003, pg. 549), much like Nicol and Swinney (1989)'s conclusion. This is striking in comparison to agreement attraction, which effectively constitutes the grammatical constraint being partially overridden by the presence of the plural feature on the intruding noun.

Dillon et al. (2013) remarked on the apparent difference between intrusion profiles of agreement v.s. reflexive anaphor, and designed a direct comparison with the items in (29). Like Sturt (2003) and many others working on reflexives, they used a version of the mismatch paradigm, manipulating the number of the lure noun and the critical verb/anaphor.

- (29) a. The new executive who oversaw the middle manager( $\emptyset$ /s) apparently was dishonest about the company's profits.
- b. The new executive who oversaw the middle manager( $\emptyset$ /s) apparently were dishonest about the company's profits.
- c. The new executive who oversaw the middle manager( $\emptyset$ /s) apparently doubted himself on most major decisions.
- d. The new executive who oversaw the middle manager( $\emptyset$ /s) apparently doubted themselves on most major decisions.

In both Total Time and Probability of Regression, the agreement conditions showed an interaction of grammaticality  $\times$  intrusion, i.e., the traditional agreement attraction effect, even though the pairwise comparisons between intrusion and no intrusion versions of the ungrammatical conditions only attained significance in Total Time. For reflexive conditions, there was a main effect of grammaticality in First Pass and Total Times. But, in accord with many previous findings from reflexive intrusion studies, none of the three reported measures showed the critical grammaticality  $\times$  intrusion interaction, nor was the pairwise comparison of intrusion versus non-intrusion of the ungrammatical conditions significant in any measure. Finally, the interaction in the full  $2 \times 2 \times 2$  ANOVA, which included dependency as a factor, was significant by participants in Total Times, indicating a reliable difference in the sensitivity of reflexives and agreement to intrusion.

Based on their results, Dillon et al. argued that agreement and reflexives have fundamentally distinct memory intrusion profiles. Following the argument that predicts intrusion from cue-based memory, they proposed that only agreement employed true cue-based retrieval. As an alternative to cue-based retrieval, they suggest that reflexives access memory via a Search operation that only considers candidate antecedents which already comply with the constraints of the grammar. Search is a fundamentally different mode of retrieval which checks each element in memory sequentially for whether it fully matches the probe or not. Search is an intuitive way to apply hard constraints like grammatical rules, but it is less efficient when searching a large candidate space.

If the parser can apply different retrieval mechanisms to fulfill different grammatical requirements, this would be profound. A possible implication might be that in some cases, grammatical rules like Principle A can be enforced as such hard constraints that the parser employs a less efficient retrieval in order to ensure that syntactic demands are respected. Moreover, if specific grammatical constructions can actually determine how memory is accessed, this would suggest

a much tighter and more integrated relationship between sentence processing and memory than would be required under other circumstances (for instance, if memory were applied purely on the basis of linear distance).

But Dillon et al. (2013)'s claim is not without challengers. In point of fact, the strict version of the finding —that reflexives *never* show interference —is already known to be false. For instance, King, Andrews, and Wagers (2012) employed the mismatch paradigm from Sturt (2003) with a manipulation of the distance of the reflexive from the verb, shown in (30).

- (30) a. The bricklayer who employed Gregory/Helen shipped himself/herself sacks of mortar...
- b. The bricklayer who employed Gregory/Helen shipped sacks of mortar to himself/herself ...

As in previous studies, the reflexive in direct object position showed no sensitivity to the lure noun. However, when the reflexive was distanced from the verb by being inside a PP, the anaphor was susceptible to interference. In this sense, they showed that, at minimum, Principle A is not solely responsible for determining presence/absence of interference.

Sloggett (2017) identified a separate case of reflexive interference based on his own work and the work of Parker and Phillips (2017). The critical case is when the reflexive anaphor was embedded under a speech predicate, shown in (31). In a series of experiments, Sloggett showed that interference obtains under the speech predicate *said*, but not under the perception predicate *heard*.

- (31) The [*librarian/janitor*] [*said/heard*] that the schoolgirl misrepresented herself at the meeting.

Interestingly, the anaphoric configuration in (31) is exactly where some languages have grammaticalized a special pronominal form known as a logophor.<sup>3</sup> While the best known languages with true logophoric pronouns are Niger-Congo languages in West Africa, Sloggett notes that even some varieties of English show logophoric-like properties, notably Iron Range English in Northern Minnesota. The existence of specialized pronominal forms suggests that speech predicates create an environment that the forces which shape grammars can be quite sensitive to, indicating that this has the potential to be processed differently even when the surface morphological form is the same.

---

<sup>3</sup>Notably different from other types of Principle A exempt anaphors which are sometimes also called logophoric.

While King et al. (2012); Parker and Phillips (2017); Sloggett (2017) all demonstrate reflexive interference, they are less direct challenges to the original claim of Dillon et al. (2013) and more of a clarification of the scope. Between them, they identify two separate cases in which reflexives are susceptible to interference, but ultimately either confirm or at least do not contradict the lack of reflexive interference in Dillon et al.'s configuration of a reflexive anaphor as the direct object in the environment of a non-speech predicate. Thus while these studies are technically cases of reflexive interference, the original question about whether there is a hard contrast in intrusion profiles between agreement and Dillon et al.'s reflexives is still valid.

On this front too, there have been some challenges to Dillon et al. (2013)'s pattern, although with less clear results. For instance, recent work by Jäger, Mertzen, Van Dyke, and Vasishth (2018) claimed to have found nearly the exact opposite pattern of results. Based on their own large scale study with items similar to Sturt (2003) and Dillon et al.'s, Jäger et al. ran a Bayesian analysis which not only found evidence of reflexive attraction but also estimated overlapping 95% credible intervals for the agreement and reflexive attraction effects in their data, implying no reliable difference between the susceptibility of the two dependencies to attraction. They found this pattern precisely in the Total Time measure where Dillon et al. (2013) had found their critical interaction. Jäger et al. attribute the difference between their findings and those in Dillon et al. to higher power in the 2018 study. This point is well taken, and suggests that future studies should aim for increased power if they mean to address this issue. On the other hand, Jäger et al. do show a contrasting pattern between agreement and reflexives in regression-based eye movement measures. This indicates that power alone may not be the whole story.

### **3.2 Experiment 2**

If Dillon et al. (2013)'s hypothesis that reflexive and agreement dependencies use fundamentally different retrieval algorithms is correct, this would be an important finding on the role that grammar plays in interfacing language processing with mechanisms from general cognition and memory in particular.

The current study has the same motivating idea as Dillon et al. (2013), and the same design at a macro-level. But, while the stimuli in this study are modeled on the sentences in Experiment 1 of Dillon et al., there are several notable differences. First are the changes to how the interference manipulation was implemented in the current experiment. In the present study, the verb and reflexive were always plural and grammaticality was manipulated by modulating the number

of the subject, allowing the critical region to be identical across grammatical and ungrammatical conditions. In particular, the form of the reflexive anaphor was always *themselves*, which eliminated the role of gender features in the design, making the reflexive conditions a closer minimal comparison with the agreement conditions. Because keeping identical critical regions across conditions meant that within the experimental items themselves, the form of the verb or reflexive and the grammaticality of the sentence overall was predictable from the number value of the subject, the fillers balanced the predictability of the morphological forms. Across the experiment, there were an equal number of grammatical and ungrammatical sentences with a plural subject and either an agreeing auxiliary or reflexive. The same was true of grammatical versus ungrammatical sentences with a singular subject. In addition, some of the remaining fillers included other forms of the reflexive (*herself, himself*).

Another substantial difference from the Dillon et al. (2013) stimuli is the spillover region following the reflexive. The change to only using *themselves* as the critical form of the reflexive makes the spillover particularly important when comparing attraction for inflected auxiliaries to reflexives. Regioning of eyetracking sentences tends to put the auxiliary at the beginning of a region containing it and the main verb, while it is natural for the reflexive to be its own critical region. This is potentially a problem since it means that the primary reliable cue to plurality came at the beginning of the agreement critical region and the end of the region. If, for instance, evidence of the intrusion effect began to appear reliably on the fixation immediately following the cue to plurality, then the effect would appear on the critical region for agreement dependencies, but not in the critical region for reflexives.

Dillon et al. approached this problem by always including a function word following the anaphor in the critical region for the reflexive. Including the function word prevented the plural morphology of the reflexive *themselves* from being the last thing in the critical region. However, skipping rates on short function words like the prepositions used in Dillon et al. can be quite high. An alternative is to control the spillover region for reflexives so that it can be analyzed in its own right. Dillon et al. did not intend to analyze the region after the critical reflexive region, and it was therefore not controlled for length, sometimes varied across conditions, and was often the last region of the sentence, leaving it open to wrap-up effects (Just & Carpenter, 1980; Mitchell & Green, 1978). If Dillon et al.'s inclusion of a function word following the reflexive was not sufficient for any reason, it would not have been possible to check the spillover region for any intrusion effect that could have appeared there. The current approach puts *themselves* in a region

of its own, but controls the spillover such that it was always three words (either a prepositional phrase or a D-Adj-N sequence) and remained the same across all conditions within an item.

- (32) a. **Agreement-Gram-NoLure:** The aunts/ of the actress/ definitely/ have been embarrassed/ at the gala/ before.
- b. **Agreement-Gram-Lure:** The aunts/ of the actresses/ definitely/ have been embarrassed/ at the gala/ before.
- c. **Agreement-Ungram-NoLure:** The aunt/ of the actresses/ definitely/ have been embarrassed/ at the gala/ before.
- d. **Agreement-Ungram-Lure:** The aunt/ of the actress/ definitely/ have been embarrassed/ at the gala/ before.
- (33) a. **Reflexive-Gram-NoLure:** The aunts/ of the actress/ definitely/ embarrassed/ themselves/ at the gala/ before.
- b. **Reflexive-Gram-Lure:** The aunts/ of the actresses/ definitely/ embarrassed/ themselves/ at the gala/ before.
- c. **Reflexive-Ungram-NoLure:** The aunt/ of the actresses/ definitely/ embarrassed/ themselves/ at the gala/ before.
- d. **Reflexive-Ungram-Lure:** The aunt/ of the actress/ definitely/ embarrassed/ themselves/ at the gala/ before.

A final crucial contrast with the materials in Dillon et al. is placement of the lure inside a PP modifier to the subject rather than a relative clause. In all of the studies on reflexive intrusion that we are aware of prior to the current experiment, the intruder noun was placed in a separate clause, attached to the true antecedent. This was seen as way to guarantee that the lure was not a grammatical antecedent of the anaphor, since it was then clearly outside of the reflexive's binding domain. However, it has been known since Bock and Miller (1991) that embedding a lure noun phrase in a relative clause yields a weaker agreement attraction effect than if the lure is inside a PP. If reflexive attraction were extant but inherently weaker than attraction in agreement, then it is possible that previous designs might have simply made the effect too small to detect. Moreover, switching to PPs as the container phrase still does not allow the lure noun to be a grammatical candidate antecedent of the reflexive, as it will still fail to c-command the anaphor.

### 3.2.1 Norming Study

Although switching to *themselves* as the critical word for all reflexive conditions has several advantages, it also brings with it a concern. Language change in American English increasingly allows speakers to use *they*, and correspondingly *themselves*, to refer to single individuals. If this is true of our participants, it could make the intended ungrammatical sentences in the reflexive conditions, (33c) and (33d), grammatical.

However, the acceptability of singular *they/themselves* appears to be tied to the prominence of the gender of the antecedent. If the gender of the antecedent is clearly defined in the discourse, then singular *they* is no longer an acceptable pronominal form. Therefore, to counteract any concerns about singular *they* increasing the acceptability of the ungrammatical conditions, the stimuli were all constructed with strongly gendered nouns in both the subject and lure noun positions, informed in part by the stereotypical gender biases compiled by Kennison and Trofe (2003). The fully constructed reflexive conditions were then normed to ensure that singular *themselves* was unacceptable. This was done by replacing the complex subject phrase with either just the subject head noun or the lure noun, crossed with whether the noun was plural (acceptable) or singular (unacceptable). For example, the item in (33) yielded the norming item set in (34).

- (34) a. **Subject.Pl:** The aunts definitely embarrassed themselves at the gala before.  
b. **Subject.Sg:** The aunt definitely embarrassed themselves at the gala before.  
c. **Lure.Pl:** The actresses definitely embarrassed themselves at the gala before.  
d. **Lure.Sg:** The actress definitely embarrassed themselves at the gala before.

These items were given to 40 subjects on Amazon Mechanical Turk to rate on a 7-point likert scale and the resulting data served two purposes. First, the norming data was used to identify items for which the singular was not enough to cause the sentence to be read as unacceptable, usually because the noun was not sufficiently gender biased enough to make *themselves* an unlicensed antecedent of the singular. Items which had less than 2.5-point spread on the likert scale between the singular and plural versions for either noun phrase were either repaired or replaced. The mean difference between all singular and plural pairs was 3.56 and the standard deviation was 0.76.

The second purpose of the norming study was to ensure that there were no systematic differences between subject and lure nouns with regards to the effectiveness of the number manipulation. That is, in the Reflexive-Subj<sub>PL</sub>-NoLure condition, where both the subject and lure noun were plural, we wanted to ensure that neither noun was a systematically more acceptable antecedent

of *themselves*. The data from norming suggest our items conformed to this criterion. Means and standard deviations for each condition are given in Table (3.1). Within the Subject pairs (plural acceptability - singular acceptability), the mean difference score was 3.54 and the standard deviation was 0.76, while within the intruder pairs the mean difference was 3.58 and the standard deviation was 0.76.

	Intruder		Subject	
	Plural	In-sg	Out-pl	Out-sg
Mean	6.017	2.440	5.902	2.367
SD	1.348	1.653	1.426	1.619

**Table 3.1.** Means and standard deviations by condition for the pre-Experiment 2 norming study

Finally, a paired sample *t*-test confirmed that there was no significant difference between the within-subject difference scores and the within-intruder difference scores ( $t(47)=0.258, p=0.797$ ).

### 3.2.2 Participants

Sixty-four participants were recruited from the UMass community in exchange for course credit. Eight participants were removed from analysis because they had lower than 80% accuracy on comprehension questions. Participants would also have been rejected if more than 25% of the critical trials had blinks or other track loss on the regions of interest, however no participants lost that many trials.

### 3.2.3 Procedure

Eye movements were recorded using a Eyelink 1000. Participants were regularly calibrated at the beginning of the experiment and at the mid-point following a break. To check that calibration remained consistent throughout the experiment, before every trial participants saw a small, black box on the side of the screen where the sentence would begin (a gaze-contingency box). The system brought up the sentence automatically once it registered a fixation to the gaze contingency box. This practice also ensured that participants were already fixated at the beginning of the sentence when recording for a trial began. In addition, participants were re-calibrated whenever they or the experimenter felt it was routinely taking too long for the system to register a fixation on the gaze contingency box.

### 3.2.4 Results

For both agreement and reflexives, we report the results for the respective critical regions and the spillover. Reading times were analyzed with linear mixed effects models and probability of regression was analysed with logistic mixed effects models. The intrusion factor was coded as the factor LURE NUMBER, which meant that the critical interaction indicating interference from the lure noun for models applied within a single dependency would be GRAMMATICALITY X LURE NUMBER.

Except where otherwise noted, all analyses in this section take raw reading times as their input. This deserves a special note, because many of the previous studies in this area, (although not the Dillon et al. (2013) precursor experiment to this one) analyzed reading times which had been log-transformed. Log-transformation is a standard practice with reading time data, because raw reading times are generally right-skewed and log-transforming forces the data into a closer approximation of the normal distribution. The decision to use raw reading times in this analysis was made for two reasons. First, log transformation is known to have the ability to destroy interactions based on additivity (i.e., non-crossover interactions). Given that the key prediction of this experiment is the *lack* of an attraction interaction for reflexives, and that attraction interactions are super-additive, it was deemed more conservative for our hypothesis to use raw RTs. Second, Box-Cox tests were performed on both the untransformed and log transformed reading times, and it was found that the transformed data had a  $\lambda$  value which was even farther removed from 0 than untransformed data (0 indicates that there is no need for a transformation to satisfy the normality assumption of the model)(Osborne, 2010). For the sake of comparability with previous research, we also performed a log-transformed analysis. Interested readers can find this analysis in the Appendix.

All models include the maximal random effects structure by default, any exceptions will be noted. Following Gelman and Hill (2007),  $t=2$  is the assumed level of significance.

#### 3.2.4.1 Agreement Results

Mean RTs for agreement are given in Table (3.2); models are given in Table (3.3).

In First Pass, there was a main effect of grammaticality at the verb region ( $\beta=-52.10$ ,  $SE=13.66$ ,  $t=-3.81$ ), but no other effects were significant in either the verb region or the spillover.

In Go Past at the verb region, there were significant main effects of grammaticality ( $\beta=-180.02$ ,  $SE=31.39$ ,  $t=-5.74$ ) and lure number ( $\beta=-92.49$ ,  $SE=29.64$ ,  $t=-3.12$ ). The interaction did not reach significance, but was marginal ( $\beta=131.76$ ,  $SE=66.50$ ,  $t=1.98$ ). The same pattern was repeated in

### Mean RTs for the Agreement Conditions

	First Pass	Go Past	P(Regression)	Total Time
<i>Verb Region</i>				
Agreement-Gram-Lure	446 (19)	606 (28)	0.205	692 (28)
Agreement-Gram-NoLure	447 (20)	578 (32)	0.142	698 (36)
Agreement-Ung-Lure	480 (21)	694 (42)	0.216	765 (33)
Agreement-Ung-NoLure	517 (24)	850 (50)	0.289	890 (45)
<i>Spillover Region</i>				
Agreement-Gram-Lure	469 (17)	624 (27)	0.166	673 (26)
Agreement-Gram-NoLure	467 (18)	613 (37)	0.159	698 (32)
Agreement-Ung-Lure	448 (16)	616 (29)	0.188	670 (28)
Agreement-Ung-NoLure	461 (15)	738 (43)	0.228	734 (31)

**Table 3.2.** Mean RTs at the critical region for subject-verb agreement (the verb region). Standard errors are given in parentheses.

### Agreement Model Summaries

	Intercept	Grammaticality	Lure Number	Grammaticality x Lure Number
<i>Verb Region</i>				
First Pass $\beta$	472.08(20.45)	-52.10(13.66)	-17.47(13.32)	40.13(27.00)
Go Past $\beta$	682.24(32.47)	-180.02(31.39)	-92.49(29.64)	131.76(66.50)
P(Regression)	-1.45(0.16)	-0.52(0.16)	-0.44(0.14)	-0.05(0.31)
Total Time $\beta$	761.03(33.80)	-131.68(22.29)	-60.60(22.42)	131.17(59.06)
<i>Spillover Region</i>				
First Pass $\beta$	462.49(16.29)	13.66(12.25)	-6.72(13.62)	12.45(23.79)
Go Past $\beta$	651.29(27.53)	-59.65(27.84)	-67.02(30.51)	108.53(64.27)
P(Regression)	1.66(0.14)	-0.32(0.16)	-0.15(0.15)	0.23(0.32)
Total Time $\beta$	694.20(28.41)	-16.08(21.19)	-20.24(20.53)	93.37(42.90)

**Table 3.3.** Summary of the Linear Models applied to the Agreement results as  $\beta$  coefficients. Standard Errors are in parentheses.  $t$ -values can be obtained by dividing the  $\beta$  value by the associated Standard Error. With the exception of intercepts, dark gray shading indicates significance and light gray shading indicates a marginal effect.

the spillover region, grammaticality ( $\beta=-59.65$ ,  $SE=27.84$ ,  $t=-2.14$ ) and lure number ( $\beta=-67.02$ ,  $SE=30.51$ ,  $t=-2.19$ ) were again significant, but the interaction was not.

Probability of regression followed the same pattern as Go Past at the critical region with both main effects significant (grammaticality:  $\beta=-0.52$ ,  $SE=0.16$ ,  $p < 0.001$ ; lure number:  $\beta=-0.44$ ,  $SE=0.14$ ,  $p < 0.005$ ), but no significant interaction ( $\beta=0.05$ ,  $SE=0.31$ ,  $p < 0.87$ ). Only grammaticality ( $\beta=-0.32$ ,  $SE=0.16$ ,  $p < 0.05$ ) was significant in the spillover region.

Finally, in Total Times at the verb region, both main effects of grammaticality ( $\beta=-131.68$ ,  $SE=22.29$ ,  $t=-5.91$ ) and lure number ( $\beta=-60.60$ ,  $SE=22.42$ ,  $t=-2.70$ ) were significant, and further-

more the interaction was also significant ( $\beta=131.17$ ,  $SE=59.06$ ,  $t=2.22$ ). In the spillover region, neither main effect reached significance but the interaction did ( $\beta=93.37$ ,  $SE=42.90$ ,  $t=2.18$ ).

### 3.2.4.2 Reflexive Anaphor Results

Raw RTs for the reflexive conditions are given in Table (3.4) and reflexive model summaries are given in Table (3.5). Results are reported for the reflexive region and its spillover, as given in (33).

In First Pass there were no significant effects in either region. In Go Past at the critical reflexive region, GRAMMATICALITY was significant ( $\beta=-42.94$ ,  $SE=19.99$ ,  $t=-2.15$ ). No other effects were significant in the reflexive region and none were significant in the spillover.

Finally, at the critical region in Total Times, grammaticality was again significant ( $\beta=-54.32$ ,  $SE=14.32$ ,  $t=-3.79$ ) as it was in the spillover region ( $\beta=-59.18$ ,  $SE=18.73$ ,  $t=-3.16$ ). These were the only significant effects in Total Time.

#### Mean RTs for the Reflexive Conditions

	<i>First Pass</i>	<i>Go Past</i>	<i>P(Regression)</i>	<i>Total Time</i>
<i>Reflexive Region</i>				
Reflexive-Gram-Lure	264 (9)	340 (16)	0.124	381 (18)
Reflexive-Gram-NoLure	259 (8)	332 (18)	0.132	365 (16)
Reflexive-Ung-Lure	273 (11)	385 (32)	0.134	427 (27)
Reflexive-Ung-NoLure	272 (8)	364 (21)	0.135	422 (22)
<i>Spillover Region</i>				
Reflexive-Gram-Lure	469 (20)	562 (27)	0.114	640 (27)
Reflexive-Gram-NoLure	454 (19)	551 (35)	0.089	639 (30)
Reflexive-Ung-Lure	459 (19)	568 (30)	0.109	687 (32)
Reflexive-Ung-NoLure	479 (18)	645 (46)	0.132	705 (35)

**Table 3.4.** Mean reading times at the critical region for reflexive anaphors. Standard errors are given in parentheses.

Notably, the critical interaction of grammaticality x lure number was not significant or marginal for any comparison in the reflexive analyses. Even though none of the signature tests for intrusion reach significance, there is still some evidence in this data that reflexive intrusion might still exist. Every one of the analyzed measures at the spillover region has a clear trend for the *ungrammatical-Lure* condition to be noticeably facilitated relative the *ungrammatical-NoLure* condition. Despite this, planned comparisons between the ungrammatical conditions never found a significant interaction. In the critical region, this was true in both Go Past ( $\beta=26.38$ ,  $SE=35.75$ ,  $t=0.74$ ) and Total Time ( $\beta=6.19$ ,  $SE=21.40$ ,  $t=0.29$ ). This was also true in the spillover region (Go Past:  $\beta=-76.32$ ,  $SE=46.50$ ,  $t=-1.64$ ; Total Time:  $\beta=-18.26$ ,  $SE=31.39$ ,  $t=-0.58$ ).

### Reflexive Model Summaries

	Intercept	Grammaticality	Lure Number	Grammaticality x Lure Number
<i>Reflexive Region</i>				
First Pass $\beta$	267.72(7.45)	-12.19(6.51)	-1.77(7.38)	-10.16(12.69)
Go Past $\beta$	356.88(18.32)	-42.94(19.99)	9.16(22.49)	-33.62(38.45)
P(Regression)	-2.18(0.17)	-0.06(0.20)	0.02(0.18)	0.09(0.36)
Total Time $\beta$	397.38(17.55)	-54.32(14.32)	-4.16(13.54)	-19.52(28.23)
<i>Spillover Region</i>				
First Pass $\beta$	464.98(19.05)	-9.58(11.54)	-17.12(11.05)	3.98(27.01)
Go Past $\beta$	584.39(28.19)	-52.17(34.30)	-44.19(28.11)	62.16(58.26)
P(Regression)	-2.49(0.19)	-0.21(0.22)	-0.26(0.19)	-0.02(0.47)
Total Time $\beta$	665.91(29.48)	-59.18(18.73)	-9.33(20.19)	17.36(37.48)

**Table 3.5.** Summary of the Linear Models applied to the Reflexive results as  $\beta$  coefficients. Standard Errors are in parentheses.  $t$ -values can be obtained by dividing the  $\beta$  value by the associated Standard Error. With the exception of intercepts, dark gray shading indicates significance and light gray shading indicates a marginal effect.

#### 3.2.4.3 Aggregate Results

A key prediction of Dillon et al. (2013) was not just that reflexive intrusion effects shouldn't hold for the within-dependency comparisons, but that there should be a three-way interaction confirming the lack of attraction for anaphors relative to agreement. This allowed them to move beyond merely taking null effects as evidence for a lack of reflexive attraction. The models in this section are all 2x2x2 models which address this prediction.

In First Pass times at the critical region, there was a significant main effect of dependency ( $\beta=204.89$ ,  $SE=18.28$ ,  $t=11.21$ ). This effect is unsurprising, since the critical region for the two dependencies was different, and on average the critical verb region for agreement was longer than the critical reflexives region. Because the interpretation of the dependency main effect is not central to the interpretation of the results, we will leave off reporting it for the remaining measures, although it will appear in the model summaries given in the tables. There was also a significant main effect of grammaticality ( $\beta=-31.84$ ,  $SE=8.32$ ,  $t=-3.83$ ) and an interaction of dependency x grammaticality ( $\beta=-40.34$ ,  $SE=14.45$ ,  $t=-2.79$ ). The interaction may be an indication of an overall greater penalty for unacceptability for agreement in early measures, but again, interpretation of this particular effect is confounded by the unmatched length at the critical regions. No other effects were significant in the critical region in First Pass.

In Go Past at the critical region,<sup>4</sup> all three main effects were significant (grammaticality  $\beta=-110.95$ ,  $SE=18.99$ ,  $t=-5.84$ ); lure number  $\beta=-42.29$ ,  $SE=17.49$ ,  $t=-2.42$ ). There was an interaction of dependency  $\times$  grammaticality ( $\beta=-137.89$ ,  $SE=41.08$ ,  $t=-3.56$ ). Moreover, there was an interaction of dependency  $\times$  lure number ( $\beta=-100.64$ ,  $SE=39.32$ ,  $t=-2.56$ ), following the fact that the intrusion effect also appears larger for agreement than reflexives. Lastly, the key interaction of dependency  $\times$  grammaticality  $\times$  lure number was significant ( $\beta=163.97$ ,  $SE=75.32$ ,  $t=2.18$ ), suggesting that attraction in ungrammatical sentences was reduced or non-existent in reflexives compared to agreement.

The three-way interaction was only marginal in Total Times at the critical region ( $\beta=147.01$ ,  $SE=73.90$ ,  $t=1.99$ ), but all other main effects and interactions were significant. The interpretation of the three main effects remains the same as in other models, grammaticality ( $\beta=-93.00$ ,  $SE=14.28$ ,  $t=-6.51$ ) indicates longer reading times for ungrammatical sentences, dependency is consistent with longer RTs for agreement ( $\beta=364.82$ ,  $SE=25.05$ ,  $t=14.57$ ), and lure number is associated with overall longer reading times when the lure noun's number was singular ( $\beta=-32.64$ ,  $SE=13.13$ ,  $t=-2.49$ ). As usual, the interpretation of main effects is tempered by the presence of the interactions. The dependency  $\times$  grammaticality interaction was significant ( $\beta=-77.54$ ,  $SE=26.24$ ,  $t=-2.96$ ) as was dependency  $\times$  lure number ( $\beta=-56.01$ ,  $SE=27.88$ ,  $t=-2.01$ ), though the previously discussed length confound applies. Lastly, grammaticality  $\times$  lure number was also significant ( $\beta=56.95$ ,  $SE=27.90$ ,  $t=2.04$ ), indicating the grammatical asymmetry for sensitivity to attraction.

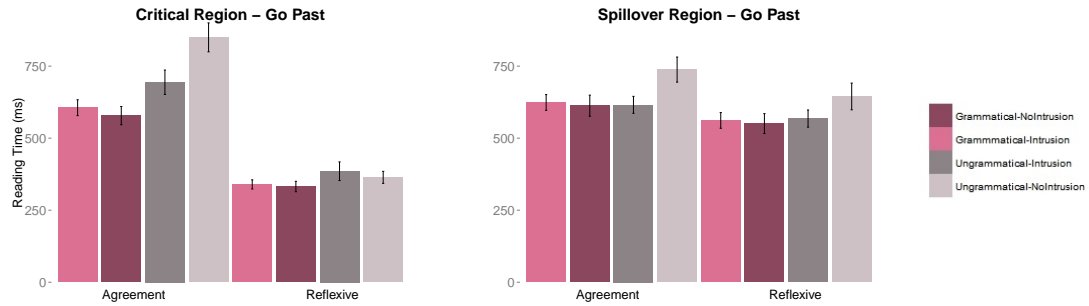
The interactions in Go Past and potentially in Total Time replicate the key three-way interaction which Dillon et al. (2013) used to argue for different retrieval mechanisms for the two dependencies. However, because the current study has a controlled spillover region, it is clear from the raw reading times that the trends in favor of reflexive attraction don't begin until the spillover, as demonstrated in Figure (3.1).

We now turn to the three-way models applied to the spillover region. This region is exactly the same across the two dependencies, and therefore avoids the length confound that hampered interpretation of the critical region comparisons by dependency.

There were no significant effects in First Pass at the spillover, but in Go Past all three of the main effects were significant (Dependency: ( $\beta=67.01$ ,  $SE=22.88$ ,  $t=2.93$ ; Grammaticality:  $\beta=-55.05$ ,

---

<sup>4</sup>The aggregate results on raw RTs, including the Total Time model at the critical region, has the greatest deviance in findings from the log-transformed analysis. However, because a number of the effects which are different between the two models are subject to confounds due to the difference in length of the critical regions, the additional terms which are significant here have a reduced impact on our final interpretation. The interested reader is directed to the Appendix for more details.



**Figure 3.1.** Go Past reading times for both the respective critical and spillover regions. Error bars indicate standard errors.

SE=22.01,  $t=-2.50$ ; Lure Number:  $\beta=-54.82$ , SE=18.56,  $t=-2.95$ ). The three-way interaction, however, did not reach significance ( $\beta=47.63$ , SE=75.88,  $t=0.63$ ).

Finally, in Total Times at the spillover, there was a significant main effect of grammaticality ( $\beta=-38.53$ , SE=14.96,  $t=-2.58$ ), and marginal interaction of grammaticality x lure number ( $\beta=54.39$ , SE=27.80,  $t=1.96$ ), but again the three-way interaction was not significant ( $\beta=75.44$ , SE=58.95,  $t=1.28$ ).

### 3.2.4.4 Extended Analyses

Because the focus of this experiment is really on the presence or absence of reflexive intrusion, this effect warrants some additional attention. All of the analyses in this section are exploratory rather than planned and should be treated as such during interpretation. However, we believe that they may be valuable, particularly in informing the hypotheses and design of future studies. Moreover, for two of these analyses we applied Bayesian techniques, using the *brms* R package (Bürkner, 2018), because Bayesian somewhat mitigates concerns about multiple comparisons and the difference between exploratory and planned analyses.

Unless otherwise stated, all Bayesian models use the same formula as the standard *lme4* model, but with the shifted lognormal distribution as the base distribution for fitting RTs. The shifted lognormal is exactly like the lognormal distribution except that it includes an additional parameter which shifts the position of the distribution along the x-axis, given constant values of the standard normal parameters  $\mu$  and  $\sigma$ .

#### *Cumulative Progression*

We begin, however, with a non-Bayesian analysis. The fact that agreement attraction in this data appears to be centered on the critical region while any potential reflexive intrusion effect

is centered on the spillover region makes the issue of timecourse key. Typical eyetracking while reading analyses of the sort discussed so far are bound to analysis by sentence regions. Even when regions are well-chosen and well-founded, they may be an awkward way to carve up the continuous eye-movement record.

There are two questions about the impact of the particular region schema in this experiment that could be helped by a more continuous view of eye movements. First, is there any evidence that the region schema unfairly weakened the ability of inferential statistics to detect reflexive attraction? For instance, if the agreement attraction effect were entirely contained within the critical region but the reflexive attraction effect began in the critical region and finished in the middle of the spillover, this could lead to a situation in which reflexive attraction never had sufficient statistical power to reach significance in either region even though in theory it might have the same effect size as agreement attraction. Second, to what extent does the fact that the evidence in favor of any reflexive intrusion appears in the spillover region actually indicate that reflexive attraction is later than agreement attraction? Like the first question, this second one arises from a question about why, given that there are clear trends in favor of reflexive attraction in the spillover, these trends fail to attain significance and how this should impact conclusions about the difference between agreement and reflexive processing based on the present data. If some version of reflexive attraction did exist, but it only appeared later and weaker than agreement attraction this would be theoretically significant (and notably in accord with the line of conclusions in the reflexive literature beginning with Sturt, 2003).

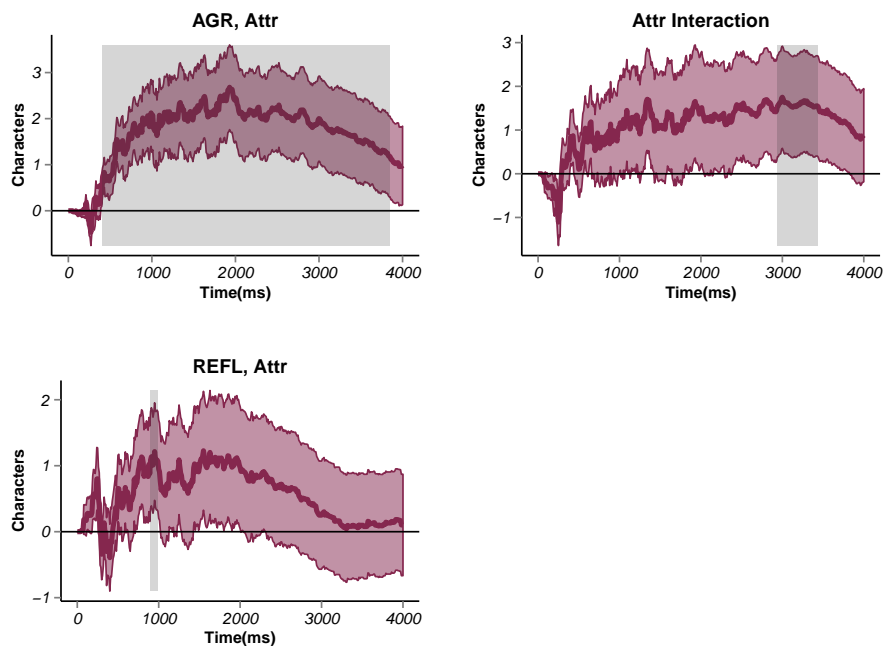
We address these two questions using cumulative progression. Given two conditions, cumulative progression tracks the number of characters that the average reader has progressed from a particular starting point in one condition versus the other (Kreiner, Sturt, & Garrod, 2008).

(35) ... actress(es) definitely/ had been embarrassed...

For instance, if the starting point in the agreement conditions were the start of the critical region at the slash in (35), and 500ms after passing the slash readers in the lure condition were on average fixated on the *m* of *embarrassed* while readers of the non-lure condition were on average still fixated on the *n* of *been*, this would be evidence that the lure condition was being read faster than its non-lure counterpart. The cumulative progression algorithm checks the rightmost progression of eye movements every 10ms and then runs a *t*-test to check for significance between the conditions<sup>5</sup>.

---

<sup>5</sup>An up-to-date version of the script used in this section can be found at [github.com/ssloggett/cumulativeprogression](https://github.com/ssloggett/cumulativeprogression)



**Figure 3.2.** Results of cumulative progression on pairwise comparisons of the attraction effects in the ungrammatical conditions of agreement and reflexives, and then the corresponding cross-dependency interaction. Regions highlighted in light gray are those where the  $t$ -tests between the two conditions were all significant.

Cumulative progression was applied to the effect of attraction on agreement and reflexives separately and then as an interaction between them, starting at the onset of the respective critical regions. For the sake of simplicity, only ungrammatical conditions were included in this analysis, so that the separate agreement and reflexive cumulative progression analyses are effectively pairwise comparisons. The outcome of cumulative progression is shown in Figure (3.2). Stretches of time where all the  $t$ -tests in a row are significant are highlighted in gray.

Figure (3.2) shows that any concerns that the regioning schema might be solely responsible for the lack of reflexive attraction in this experiment are unfounded. Agreement attraction appears as a sustained difference between the lure and non-lure conditions beginning  $\sim 500$ ms and remains for more than 3 seconds. In contrast, reflexive attraction manifests as a significant effect only for a very short span around 1000ms after onset of the critical region. While it is possible that this region of significance for reflexive might be disrupted by the region boundary, cumulative progression makes it clear that reflexive attraction is disadvantaged relative to agreement attraction because it is a shorter and weaker effect, not because of an accident of region placement.

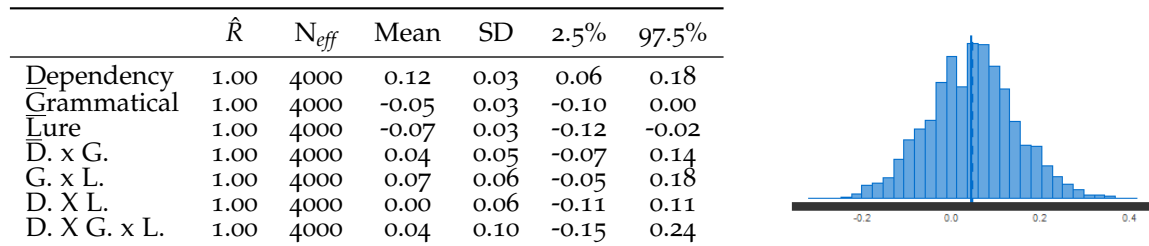
*Bayesian Analysis of Standard Regioning*

The first Bayesian analysis in this section is a direct counterpart of a subset of the traditional NHST analysis in the preceding parts of the Results section. It uses the same data in all respects, including the same regioning schema.

The goal of this analysis was to supplement the NHST statistical findings with estimations of the range of credible values for the key parameters in these findings. The values of interest are those that estimate the size of any reflexive attraction and compare it to the corresponding size of the agreement attraction effect. This analysis takes on all the more importance given the pattern of null NHST results for reflexive attraction combined with a repeating trend in the raw means that appears to be reflexive attraction. For this reasons it concentrates on the spillover region, where the reflexive attraction pattern is evident in the raw data. Finally, the present analysis facilitates comparison of the current results with those of Jäger et al. (2018), whose key claim was that they found no difference in the size of the effect for agreement and reflexive attraction.

Figure (3.3) gives the parameter estimates for the full 2x2x2 model in Go Past at the spillover, along with a histogram of the posterior distribution of the three-way interaction term itself. The estimates and the posterior distribution demonstrate that not only is zero within the 95% credible interval (as expected for a term that was non-significant in NHST), it is actually quite near the center of the probability mass. A *de minimus* value for this interaction indicates no difference of consequence between agreement and reflexive attraction.

There is a question about how this interaction term should be understood. On the one hand, the key prediction of Dillon et al. (2013) was that there should be an three-way interaction. The lack of it in the region which is completely lexically matched for both dependencies seems troubling for that argument. On the other hand, the NHST results merely say that there is no reliable difference, with relatively little insight into the underlying data patterns. On the surface, a null interaction is

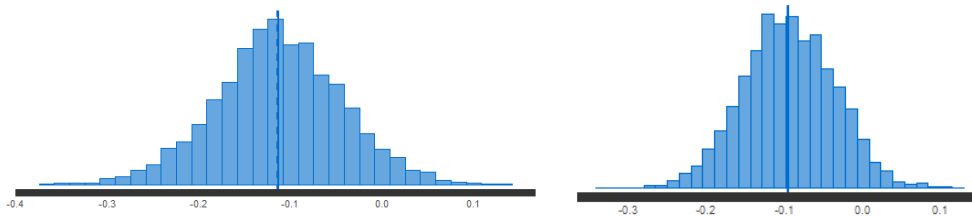


**Figure 3.3.** Summary of Bayesian estimates for the full 2x2x2 model in Go Past at the spillover region, including a histogram of posterior of the critical three-way interaction. The values for 2.5% and 97.5% in the table refer to the end points of the 95% credible interval for the parameter.

possible because both dependencies have attraction or because neither do. The NHST statistics at the spillover indicate no reliable attraction effect for reflexives, but the descriptive data pattern in the raw means in Tables (3.2) and (3.4) leaves open the possibility that attraction exists and only reaches sufficient power with both dependencies in the model.

To address these concerns, the Bayesian analysis continues with Figure (3.4), which gives the parameter estimations for several pairwise comparisons testing the lure number effect within the ungrammatical sentences, including histograms of the posterior for both dependencies in Go Past at the spillover. The pairwise comparison of agreement in Total Time at the critical region was chosen as a banner case of the agreement attraction, when the sensitivity to lure number is at its apogee. While the absolute values are small —owing to lognormal space —the credible interval for this parameter clearly does not include zero as a potential value, indicating that lure number does lead to reliable attraction for agreement at this point.

	Region	Measure	$\hat{R}$	$N_{eff}$	Mean	SD	2.5%	97.5%
Agreement	Critical	Total Time	1.000	4000	-0.146	0.052	-0.247	-0.044
Agreement	Spillover	Go Past	1.000	4000	-0.113	0.069	-0.253	0.021
Reflexive	Spillover	Go Past	1.000	4000	-0.095	0.059	-0.211	0.018



**Figure 3.4.** Outcomes of individual pairwise comparisons within the ungrammatical conditions, one for agreement and one for reflexives. *Left:* Posterior distribution for Lure Number parameter in the agreement pairwise comparison. *Right:* Posterior distribution for Lure Number parameter in the reflexive pairwise comparison.

In contrast, the posterior distribution for the two comparisons at the spillover demonstrate that zero is not only within the credible interval for both dependencies at this point, but occupies a position associated with roughly equivalent amounts of probability mass in both distributions. While not conclusive then, the evidence from Figure (3.4) suggests that the reason the three-way interaction fails to reach significance in the spillover is likely because neither dependency has a reliable attraction effect in this region, *not* because they *both* have an attraction effect. Or at least, if such an attraction effect exists it reduced compared to agreement attraction at its height.

The outcome of this analysis appears to be that the lack of interaction in the spillover region is not a clear challenge to Dillon et al.’s conclusion. Like the cumulative progression analysis, it

also highlights how time-dependent attraction is, since the signature of agreement attraction has already dissipated by the spillover.

*Bayesian Analysis with Critical Regions of Equal Length*

A running concern in the interpretation of the comparison between agreement and reflexives in the NHST analysis was the difference in length of the two critical regions. The Bayesian analysis in this section is meant to address those concerns. The critical region for reflexives was always *themselves*. The standard critical region for agreement on the other hand, always included the number-marked auxiliary and the main verb. This meant that the agreement stimuli had a great deal more variability in the length of the critical region, and as demonstrated by the sample stimuli in (32) the overall critical region could sometimes be quite long.

The difference in the length of the two critical regions is capable of producing a dependency x grammaticality or a dependency x lure number interaction in its own right. In the longer agreement conditions, the increased length and numbers of fixations times will fall more on the critical region, while fixations at the same distance from the verb (the cue to ungrammaticality) will fall on the spillover. More fixations which are also longer will lead to a super-additive effect of grammaticality or lure number. While this type of length effect is technically possible in any eyetracking experiment, it is normally assumed to be mitigated by regions that have equivalent variability in length. In this experiment, where one condition's critical region is completely constant and another isn't, the concern is more acute.

To address this concern, the agreement conditions were re-regioned so that the critical region always contained twelve characters, just like the reflexive critical region. The start point of the critical region remained the same, and as a result the end of the critical region in this schema frequently came in the middle of a word. The spillover then became the next 17 characters starting from wherever the critical region ended (consequently the spillover no longer matched across dependencies).

- (36) a. The aunts of the actress definitely/ have been e/mbarrassed at the/ gala before.  
b. The aunts of the actress definitely/ embarrassed/ themselves/ at the gala befo/re.

The fact that the agreement region interrupts words means that the comparison between the critical regions across dependencies is still not exact, but the adjusted region schema removes mere length in characters as a concern.

Table (3.6) gives the new estimates for the model parameters at the critical region in Go Past with the adjusted region schema. The findings from this model largely concur with the NHST model at the level of the 2x2 interactions, in that the 95% credible intervals for dependency x grammaticality and dependency x lure number excluded 0. This mitigates concerns that these were driven purely by length differences in the original analysis. On the other hand, the 95% credible interval for the 2x2x2 interaction does include 0. This could be an indication that the three-way interaction term in the NHST analysis was impacted by the fact that the agreement critical region was longer. Alternatively, it could reflect the timecourse of the attraction effect, if the new region schema cuts the critical region too early in the development of the attraction effect for the interaction to be secure.

	$\hat{R}$	$N_{eff}$	Mean	SD	2.5%	97.5%
Depend.	1.000	2957	0.606	0.044	0.521	0.692
Gramm.	0.999	4000	-0.128	0.032	-0.191	-0.065
Lure	0.999	4000	-0.071	0.032	-0.133	-0.006
DependXGramm.	0.999	4000	-0.157	0.065	-0.285	-0.029
GrammXLure	0.999	4000	0.023	0.061	-0.097	0.138
DependXLure	1.000	4000	-0.137	0.063	-0.261	-0.014
DependXGrammxLure	0.999	4000	0.083	0.122	-0.156	0.322

**Table 3.6.** Summary of Bayesian estimates for the full 2x2x2 model in Go Past at the critical region. The values for 2.5% and 97.5% in the table refer to the end points of the 95% confidence interval for the parameter.

### 3.3 Discussion

Ostensibly, the current study replicated the results of Dillon et al. (2013). Slightly paradoxically, however, the current findings make the possibility of reflexive intrusion seem much more likely than it had been previously. They support a somewhat weaker version of Dillon et al.'s claim of fundamentally different memory retrieval mechanisms for subject-verb agreement and Principle A dependencies.

The significant results of this study across-dependencies are largely in accord with the findings of Dillon et al. (2013). There was an interaction of agreement and reflexive anaphora as predicted, consistent with an attraction effect for agreement but not for reflexives. This was particularly true at the critical region, which was the region that Dillon et al. investigated. In fact, if anything, the significant results strengthen the findings from Dillon et al., who only found the interaction in Total Times by participants. We find the omnibus interaction in Go Past, a much earlier measure

than Total Times with a full linear regression model that includes random effects for participants and items.

Moreover, within-reflexive comparisons never found a significant interaction of grammaticality x intrusion, or for that matter even a main effect of attraction. This is notably in contrast to agreement, where the key interaction approaches significance in Go Past, does reach significance in Total Times and is accompanied by a significant main effect of intrusion in both measures in the critical region.

All of these findings do suggest a difference in the intrusion profiles of agreement and reflexive anaphors, and consequently a difference in the way that number features are retrieved from memory in agreement and reflexive dependencies. However, the inclusion of a controlled spillover region in this version of the comparison relative to Dillon et al. is enlightening. The trend for reflexive attraction is present across every measure in the spillover. Moreover, as suggested in the description of the stimuli for this experiment, reflexive attraction patterns in the spillover are not inherently later just because they are in the spillover and deserve to be considered on par with the critical region. This would suggest that interpreting the null findings of the regression models as a true lack of intrusion would be an overreach in the context of the current results. This is arguably supported by the lack of interaction in the omnibus 2x2x2 regression at the spillover, where the reflexive attraction effect begins to appear. Following the logic that predicted the three-way interaction in the first place, the absence of the interaction would suggest that there is no reliable difference in the intrusion profiles of the two dependencies at that point. It is entirely possible that the lack of any significant reflexive attraction in this experiment is due to a lack of power. This could be supported by the suspiciously weak agreement attraction pattern in the data even with the switch to lures embedded in PPs, which should strengthen the attraction. A parallel concern is the late and weak development of the grammaticality effect, when grammaticality status should not be in question for these sentences.

Even so, the data from this experiment, particularly the extended analyses, does indicate that the attraction which appears in the spillover region is weaker. Even with clear trends consistent with reflexive attraction, it is important that the intrusion effect never reaches significance despite a myriad of chances. Furthermore, the extended analyses confirmed that the lack of significant effects is not a fault of semi-arbitrary decisions like region schema. Cumulative progression shows that reflexive attraction patterns are shorter-lived independent of region boundary. And the Bayesian analysis on the standard region schema demonstrates that the lack of 2x2x2 interaction

at the spillover is just as easily understood as *neither* agreement or anaphora having attraction at that point in the eye record.

Based on the current evidence, it would be too much to reject the existence of reflexive attraction outright. But we also don't find evidence to match Jäger et al. (2018)'s claim of similar effect sizes for attraction across dependencies. The point at which the 95% credible intervals overlap also includes zero, and claiming the same effect size when 'no effect' is under serious consideration is not a productive way forward.

Given this state of affairs, we propose a modification of Dillon et al. (2013)'s, not entirely unlike the proposal from Sturt (2003). Under the current hypothesis, primary reflexive retrieval is impervious to intrusion. However, the targeted retrievals which are meant to resolve specific morphosyntactic functions run parallel to constant, general retrievals aimed at interpretation and integrating current input into previous semantic material for a full, coherent event representation, based on the work handed to the interpretation unit by lexical morphological and syntactic parsing. While morphosyntactic retrievals for anaphors are not susceptible to interference, interpretation retrievals do use standard cue-based retrieval in the proposed system, and therefore are susceptible to intrusion. This hypothesis accounts not just for our data but is consistent with Dillon et al. (2013) and Sturt (2003).

Regardless of whether the current null results for reflexive intrusion are true null effects or merely the outcome of low power, we believe it is important to consider what it would mean for theories of the language-memory interface that different grammatical dependencies would use separate retrieval strategies. We see two options, depending on why the split occurs.

One possibility is that it is the role of memory is functionally adaptive. Resolving the antecedent for reflexive anaphors has fundamental interpretational consequences, while the best available evidence indicates the English agreement has very little interpretational value (Lau, Wagers, Stroud, & Phillips, 2008; Schlueter, 2017; Schlueter, Parker, & Lau, 2019). This would leave open the possibility that languages with more robust agreement systems and free (or freer) word order might rely more on agreement for interpretation and therefore impose the stricter memory retrieval requirements that English reflexive anaphors apparently exhibit. This is a view of memory which is dynamic and responsive, adapting to find the most efficient processing algorithms for each individual language.

An alternate possibility is that memory simply interfaces with AGREE and Principle A differently, and does so across languages. If so, this would speak to fundamental representational disconnect, perhaps indicating that what theoreticians view as syntax is not necessarily stored

as a single unit from the point of view of the stable, long-term representation of grammatical knowledge nor handled the same in processing.

If there were fundamental differences in the representation of the grammatical knowledge that leads to subject-verb agreement and reflexive anaphora, then it would not be surprising that they employ distinct processing algorithms online. Unfortunately, without well-articulated theories of how hierarchical structure, particularly c-command, is implemented in memory, it is difficult to articulate what the difference in representation might be. What would be clear is that the linear form of a morphological dependency is not a reliable indicator of how memory access proceeds online.

## CHAPTER 4

### PRIMING IN COMPREHENSION

Syntactic priming was introduced in Chapter 1 as the finding that a speaker who has recently been exposed to a particular syntactic construction is then more likely to employ that syntactic construction in the near future. This tendency can hold when even the meaning of the sentence, lexical items, and task have all changed (Mahowald, James, Futrell, & Gibson, 2016; Pickering & Ferreira, 2008). In production, priming is one of the best established and most robust findings in psycholinguistics, covering a full gamut of syntactic constructions and experimental modalities. Its ubiquity has encouraged researchers to posit theories of priming based on equally ubiquitous cognitive mechanisms, such as activation and decay patterns or generalized learning mechanisms.

However, in comprehension, priming has not been so robust, to the point that at times researchers had suggested that comprehension priming might not exist at all. In point of fact, there are now quite a number of contexts under which comprehension priming has been argued to exist, although each of these comes with its own set of uncertainties and caveats.

This chapter considers priming, why comprehension priming (or lack thereof) presents a problem, and then outlines two current theoretical approaches that address both the limited distribution of comprehension priming and make predictions about where priming should be found. As part of evaluating the two theories, the chapter also reviews a range of proposed instances of comprehension priming based on which theory predicts them and evaluates the overall strength of evidence for each case.

#### **4.1 Introduction to Priming**

Priming has been introduced earlier in the dissertation as one of the crucial pieces of evidence for the existence of abstract linguistic structure in memory. The role of priming as evidence for abstract structure has played a key part in determining the direction of priming research. In particular, demonstrating syntactic structure priming requires removing all other possible similarities between the prime and the target that could be confounds for a syntactic interpretation of the effect, such as lexical, semantic, pragmatic and prosodic overlap. For this reason, initial research

beginning with Bock (1986) focused almost entirely on syntactic alternations, most classically the dative alternation, with relatively small to no meaning difference between the alternate forms.

In Bock's original study, participants were prompted to produce either the Prepositional Object frame (PO, [37]) or the corresponding Double Object (DO) frame of a prime sentence by repeating the sentence after the experimenter had read it out loud. They were then given a picture, such as a man sitting with a boy and a book, and were asked to make up a one sentence description for the picture.

(37) PO Prime: The rock star sold some cocaine to an undercover agent.

(38) a. PO Target: The man is reading a story to the boy.

b. DO Target: The man is reading the boy a story.

(Bock, 1986)

Participants who had recently produced a PO prime became more likely to produce a PO picture description instead of the equivalent DO description (38b) than if they had not already encountered (37). The reverse held true when participants had just produced a DO prime: they then became more likely to produce a DO description.<sup>1</sup>

Subsequent studies found that a second, distinct type of priming occurs when the prime and target do share a lexical item. An early example from Pickering and Branigan (1998) used a sentence completion task with primes that included the first post-verbal argument, such as (39a) and (39b) (emphasis added), as way to force/strongly bias either a PO or DO completion. The target fragments, including (40), ended at the verb to give participants the freedom to provide a DO or PO completion as they saw fit.

(39) a. The racing driver showed the torn overall...

b. The racing driver gave the torn overall...

(40) The patient showed...

Pickering and Branigan demonstrated that when participants were primed with (39a) their priming-compatible PO completions of (40) rose to ~17 percentage points more likely than the unprimed DO. This was compared to only a ~4 percentage point spread between the primed and

---

<sup>1</sup>In Bock (1986)'s original dative findings, the most common response was always the primed one, i.e., PO responses outnumbered DO responses when PO was primed and vice versa. However, this need not be the case. For alternations with a greater skew between the frequency of the two alternatives, such as active and passive, priming the less frequent alternative may result in a relative increase in the number of prime-compatible tokens produced, while the more frequent alternative remains numerically more common even when it wasn't primed.

unprimed structures when (39b) primed (40), i.e., abstract priming. Priming the DO structure instead yielded comparable differences between the lexical boost and abstract priming.

Bock's original type of priming came to be known as abstract priming, because only abstract structure was held in common between the target and the prime, while the version with lexical overlap has been called the "lexical boost". Critically, these two types of priming behave quite differently. In general, the lexical boost is much stronger than abstract priming, continuing the trend set in Pickering and Branigan (1998). The lexical boost is also much shorter lived: most studies agree that in the production paradigms discussed thus far, the effect of the lexical boost disappears within a single trial (Branigan & McLean, 2016; Hartsuiker, Bernolet, Schoonbaert, Speybroeck, & Vanderelst, 2008). Abstract priming on the other hand, can persist over at least 10 intervening trials (Bock, Dell, Chang, & Onishi, 2007; Bock & Griffin, 2000) and possibly considerably more (Gurevich, Johnson, & Goldberg, 2010).

Because abstract priming is one of the foundations of the argument for abstract syntax in memory, the remainder of the thesis will not really be concerned with the lexical boost, to the point that we will frequently simply refer to "priming" when abstract priming is intended.

In the decades that have followed Bock (1986), priming has proved to be impressively robust. In addition to Bock's original demonstration of priming for the dative alternation and active/passive alternations, priming has been shown to extend to a number of constructions including general prepositional phrase structure (Bock & Loebell, 1990), participle-auxiliary order in Dutch subordinate clauses (Hartsuiker & Westenberg, 2000), relative clause attachment (Scheepers, 2003), prepositional attachment (Branigan, Pickering, & McLean, 2005), optional *that* complementizers (V. Ferreira, 2003), placement of the particle in particle verb constructions (Gries, 2003), and the form of the genitive construction (Bernolet, Hartsuiker, & Pickering, 2012). It has even been found that non-linguistic operations can participate in syntactic priming: the order of operations in a short arithmetic equation, as determined by parentheses, can prime relative clause attachment (Scheepers et al., 2011).

Priming has equally been shown across a range of methods from sentence repetition and picture description in Bock (1986)'s original study, to both auditory and written sentence completion (Branigan, Pickering, & Cleland, 1999; Hartsuiker & Westenberg, 2000; Pickering & Branigan, 1998), production latencies (M. Smith & Wheeldon, 2001), corpus studies (Gries, 2005; Szmrecsanyi, 2005, 2006), and visual world eyetracking (Arai, van Gompel, & Scheepers, 2007). It appears in children (Branigan & McLean, 2016; Rowland, Chang, Ambridge, Pine, & Lieven, 2012) and across languages (Bernolet et al., 2012; Hartsuiker, Pickering, & Veltkamp, 2004).

The fact that priming applies across such a broad range of grammatical and experimental contexts would seem to require that any explanation must rely on cognitive machinery which held in common by all the relevant tasks. And indeed, the theories that developed to explain the original priming findings fall into two broad families —Activation-based theories and Implicit Learning theories —based on which mechanism from cognitive research they exploit.

Activation-based theories are the oldest, dating back to Bock (1986)'s initial discovery of priming. These theories make use of the general cognitive mechanism of activation and subsequent decay, applied to a speaker's store of syntactic knowledge. This knowledge is divided up into units or "rules": for illustration purposes it is sufficient to think of these rules as treelets, consisting of all the non-lexical nodes that are necessary to distinguish one construction from another.<sup>2</sup> For instance, in the dative alternation this would likely consist of the verb node and the nodes of its two arguments in the configuration for either DO or PO frames. All syntactic rules are assumed to have a baseline resting activation, which varies based on their use. More frequent syntactic rules would have higher resting activations than rarer grammatical structures. In addition, each use of a syntactic structure creates a spike of activation in the moment. When a speaker wishes to encode a message for production, they consult their grammatical knowledge for all the structures that could be used to express that message. If this search turns up more than one syntactic option, the speaker then needs to make a choice, and this choice will be influenced by overall activation: all other things being equal, higher activation structures will be chosen more frequently.

For the sentences in Bock and Griffin (2000) (repeated from Chapter 1 as [41]-[43]) the Activation account would play out in a fashion familiar from many other psychological constructs. Producing (41) activates the syntactic rules associated with prepositional object datives and raises the activation well above the resting rate. This activation will have begun to decay by the time the speaker begins to formulate (43), but the overall level of activation will still be elevated. The higher levels of activation make the speaker more likely to choose the PO frame than they would have been without the experience with (41).

(41) The governess made a pot of tea for the princess.

(42) The lifeguard tossed the struggling swimmer a rope.

---

<sup>2</sup>Actual proposals for what constitutes a unit of abstract syntactic knowledge vary substantially across proposals, and many other papers are simply not explicit about what constitutes a unit in their framework. This is rarely a problem because most psycholinguistic theories can easily be made to accommodate a wide range of possible discretizations of syntactic structure.

- (43) a. The boy is handing the rock star a guitar.  
b. The boy is handing a guitar to the rock star.

Activation and its decay have been used to explain psychological patterns in many areas of cognition other than language. One of the appeals of Activation theories therefore rests with their simplicity and the fact that all the cognitive machinery is already independently needed for other tasks.

In many ways, Implicit Learning theories build on the conceptual framework of Activation-theories, but with some additional machinery and a slightly different starting perspective. These theories are arguably more diverse than Activation-based theories (Bock & Griffin, 2000; Chang, Dell, & Bock, 2006), but generally speaking priming as implicit learning suggests that priming in the lab is a reflex of a cognitive mechanism which allows speakers to adjust to the frequencies of their speech community and acquire new structures through repeated exposure.<sup>3</sup> The mechanism behind implicit learning is a sort of bookkeeping procedure: speakers track how many times they encounter a particular syntactic form and adjust their production to match the frequencies in the environment. Activation is in some sense a ready-made version of this bookkeeping mechanism, because baseline activation varies with frequency. However, other mechanisms for tracking syntactic frequency are possible and theories diverge on the extent to which they couch implicit learning in terms of activation or something else (cf. Chang et al., 2006; Fine et al., 2013). The critical mechanical difference between Activation and Implicit Learning theories is how “activation” is distributed. Where Activation theories typically assumed that more activation could be added to the system without affecting other elements, Implicit Learning theories can be seen as a closed system where all possible activation is already distributed. In this view, adding “activation” to one syntactic rule therefore necessitates taking it from another, competing structure. We will return to the details of this mechanism when we consider a specific Implicit Learning proposal in Section 4.3.2.

## 4.2 Missing Priming in Comprehension

The general theories that were discussed in Section 4.1 made a point of invoking mechanisms that are broadly available, not restricted to any single construction or task. This seemed justi-

---

<sup>3</sup>Theories vary on whether they claim priming is strictly an adult learning mechanism, leaving open the possibility of another learning mechanism for earlier acquisition in children (Bock & Griffin, 2000; Fine et al., 2013). Alternatively, acquisition by children could use the same process (Chang et al., 2006).

fied, even required, based on the evidence indicating that priming appeared in a wide variety of contexts.

However, syntactic priming is not quite universally reliable. While production tasks consistently find evidence of both abstract priming and the lexical boost, the classic consensus has been that in comprehension priming paradigms, in which the participant is never asked to say or write any part of the target, abstract priming disappears (Pickering & Ferreira, 2008, *inter alia*). The lexical boost, meanwhile, remains intact in comprehension.

Branigan et al. (2005) demonstrate this finding with a picture-matching task. A prime-target pair began with sentences that contained a PP-attachment ambiguity, such as the one in (44) where *gun* could either be the instrument of the teacher-hitting event (high-attachment) or the distinguishing feature that separates the swimmer from all other swimmers (low-attachment).<sup>4</sup>

- (44) a. The teacher hitting the swimmer with the gun.  
b. The teacher prodding the swimmer with the gun.
- (45) Target: The nun hitting the monk with the ball.

Participants then saw two pictures and were asked to choose which picture matched the meaning of the string they had just read. In prime trials, one picture had the same characters and objects but did not match any reading of the sentence and the other picture was compatible with either the high-attachment *or* the low-attachment but not both. This forced the prime to be interpreted with a particular attachment structure. These were immediately followed (although the participants were not told this) by a target trial which paired a separate sentence such as (45) with two pictures that could describe the event, one that was consistent with the high-attachment interpretation and one consistent with low-attachment. The resulting ambiguity allowed Branigan et al. to test a participant's attachment preferences.

Across two experiments, Branigan et al. (2005) found that when the prime and target shared a verb, as in (44a) and (45), then during the target trial participants were primed to choose a picture that matched the attachment preference in the prime trial. However, if the verb did not overlap, as when (44b) preceded (45), then no priming of the attachment occurred. Based on these results, it would appear that comprehension-to-comprehension priming is possible with the lexical boost but that abstract priming is no longer an option in the comprehension domain.

---

<sup>4</sup>Examples are slightly adapted from Branigan et al. (2005).

A possible objection to the Branigan et al. (2005) study is that prepositional phrase attachment choices by their nature are accompanied by meaning differences, and this alternation is therefore quite different from ‘meaning-neutral’ alternations originally targeted by the priming literature. Because meaning differences were so central to the manipulation, participants may have relied more on semantics than syntax, potentially leading to a different empirical pattern. As noted by Arai et al. (2007), this was compounded by similarities between the prime and target pictures, which could have led to a preference to choose pictures that were closely related, without any reference to the sentence structure.

However, results across a number of constructions and paradigms since have reinforced Branigan et al.’s conclusion that comprehension supports the lexical boost but not abstract priming. Similar findings have been reported for datives (Arai et al., 2007; Carminati, van Gompel, Scheepers, & Arai, 2008) and main verb/relative clause garden path ambiguities (Ledoux, Traxler, & Swaab, 2007; Traxler, Tooley, & Pickering, 2014). These studies also span a range of methodologies beyond Branigan et al. (2005)’s picture matching task, from visual world eyetracking (Arai et al., 2007; Carminati et al., 2008), eyetracking-while-reading (Tooley, Traxler, & Swaab, 2009; Traxler et al., 2014) and ERP (Ledoux et al., 2007; Tooley et al., 2009). Across all of these studies, the consistent conclusion was that abstract priming effects disappeared, even while lexical boost remained reliable.

In particular, Arai et al. (2007) pursued an alternative in the visual world paradigm which led to an interesting expansion of Branigan et al.’s original claim. Arai et al. ran two visual world experiments using the dative alternation, which differed from each other only in that the first experiment had lexical overlap between the target and the prime and the second experiment did not. In both experiments, the written primes were presented on screen and participants were asked to read them out loud. On the target trials, an image was presented on the screen simultaneously with auditory presentation of a sentence describing the image. The critical measure was looks to the recipient versus the theme of the ditransitive after the onset of the verb but prior to the onset of the first post-verbal noun. During this period, participants have enough information to know that the sentence could be ditransitive<sup>5</sup> and have the opportunity to make anticipatory saccades to the object they believe most likely to be mentioned next. In this paradigm, priming manifests as an increase in anticipatory looks to the recipient following a DO prime and increased looks to the theme after a PO prime. When the prime and target shared a verb, priming was realized as an

---

<sup>5</sup>Many fillers were intransitive or mono-transitives.

interaction of prime type by picture region in First Gaze durations and separately as an increased tendency to look at the theme after a PO prime in the period after the verb onset but before the onset of the first post-verbal noun. When the prime and target did not share a verb, differences were stark: all priming effects disappeared, and only the form of the target picture description itself had any effect. This was a clear example of prime structure persisting in comprehension, even appearing in predictive processing, but only when supported by the lexical boost. Furthermore, Carminati et al. (2008) replicated Arai et al.'s original study while also fixing a potential confound from animacy of the arguments.

The Arai et al. (2007) findings are informative because they highlight a critical divide between when abstract priming obtains and when it does not. In general the other studies on priming in comprehension in this section have used tasks in which both the prime and the target were comprehended, so called comprehension-to-comprehension priming. Arai et al.'s participants actually produced the prime by repeating it back after reading it. Only the target was never produced, making Arai et al.'s task an instance of production-to-comprehension priming. Because repeating a prime out loud has been known to cause priming since the original Bock (1986) study, the prime could not have been responsible for missing priming in Arai et al.'s study. This critically rules out any explanations of missing comprehension priming due to encoding of the prime and puts the explanatory onus for missing priming specifically on the comprehended target (see also Bock et al., 2007).

While the gap for priming in comprehension is well replicated, many authors have noted that there are still some remaining doubts about how to interpret this finding and consequently how to predict the full set of conditions that control whether priming obtains or not. First, comprehension and production studies of priming have tended to use different syntactic structures. Production studies have focused on short, simple structures which are easy to elicit and minimize semantic differences, particularly the dative and active/passive alternations. These structures have generally been assumed to be too simple to process to reliably evoke a measurable difference in typical comprehension methods such as reading times and ERPs. Comprehension studies have tended to use more complex structures which elicit clearer effects in comprehension, such as garden path sentences or attachment ambiguities and (consequently) have been more willing to tolerate semantic differences between structures (as in Branigan et al., 2005). Second, the two modes also almost necessarily have different tasks and measures. Production tasks have tended to focus on the proportion of utterances produced with a particular structure in elicitation, or more rarely latency of production, while comprehension has used looking preference in visual world (Arai et al., 2007;

Carminati et al., 2008), preference in picture matching (Branigan et al., 2005), reading times, and size of the N<sub>400</sub> ERP component (Ledoux et al., 2007). The discrepancies between priming studies on production versus comprehension leave open the possibility that methodological factors play a significant role in determining the conditions under which priming exists.

Complicating matters further is that the lack of comprehension priming is not entirely consistent. Even with the extensive replication of null comprehension priming, a few studies do find evidence that it exists. Among these is Pickering, McLean, and Branigan (2013), a follow-up paper to Branigan et al. (2005) which used the exact same critical stimuli but with more fillers and more than twice as many subjects (N=32 in 2005; N=72 in 2013). This would seem to indicate that the original finding in Branigan et al. (2005) might simply be a null result due to lack of power.

In a similar vein, Tooley and Bock (2014) found comprehension priming in a within-subjects comparison of comprehension and production. The Tooley and Bock study is particularly important because the authors made an effort to match the structure and the task for the comprehension and production studies as closely as possible. Their study used a variant of Potter and Lombardi (1990)'s three-part task with both dative and reduced relative/transitive target structures. Participants first saw a sentence presented in RSVP, then a number-list distractor task, and finally a third part which corresponded to the trial type. In production trials, whether prime or target, participants repeated back the sentence that they had just read, while in comprehension trials they were given a masked self-paced reading task with the critical sentence and asked whether it was identical to what they had seen in the first part of the trial (which it always was on the critical trials). Tooley and Bock do find comprehension priming, manifested as a difference in self-paced reading times on the post-verbal region. Moreover, there was no reliable difference between the size of the effects in production and comprehension.

However, although Tooley and Bock (2014) find comprehension priming where it had not been found before, other effects in this study are surprisingly weak. In production, the effect of dative priming was only marginal and in comprehension the effect was only present in datives, and was reversed in benefactives, unlike in previous studies. This is not the only way in which the findings from this study differed from previous well-replicated results: Tooley and Bock also find no reliable difference in the size of the priming effect for lexical boost trials, despite the fact this has otherwise been a fairly stable finding in the literature (McLean, Pickering, & Branigan, 2004; Pickering & Branigan, 1998; Rowland et al., 2012; Scheepers, Raffray, & Myachykov, 2017). Although the appropriate trend for the lexical boost is clear in Tooley and Bock's data, this might

bring into question the generality of the effect size estimates from this study. Not to mention that comparing effect sizes across production and comprehension methodologies can be tricky.

The question then is, how should we reconcile these contradictory findings? Is the variation in findings due to task/manipulation? Were the studies that found a lack of comprehension priming merely underpowered? Or could the Pickering et al. (2013) and Tooley and Bock (2014) findings be Type I errors?

Critically, comprehension priming is not just an empirical problem, but a theoretical one as well. The unstable nature of priming in comprehension is not *a priori* predicted by the historically prominent theories of priming developed from the original finding in production. In Activation theories, there is no clear reason why accessing a rule for production should add activation and accessing a rule for comprehension should not.<sup>6</sup> Likewise, since Implicit Learning theories are designed around adjusting language use to the local environment, they actually predict that there *should* be a comprehension priming effect. This makes it imperative to determine the full empirical status of comprehension priming in order to determine how these theories needed to be changed or updated.

### 4.3 Paths Forward

Unsurprisingly, there have been myriad attempts to address the gap in comprehension priming on both the empirical and theoretical fronts. This has led to a multitude of studies which have looked for phenomena that could be considered comprehension priming. Some studies have looked for ways to save a very prolific notion of comprehension priming, similar to production, and some have operated on the explicit premise that finding comprehension priming even under very special, limited circumstances would be important given its apparent disappearance. Any good, explanatory theory will, of course, be closely entangled with empirical outcomes and crucially will make predictions about where priming should be expected. In some cases, theoretical stance may also determine whether a particular data pattern is considered true priming or a similar, but distinct phenomenon like parallelism.

The remainder of this chapter addresses these issues by surveying a number of potential candidates for comprehension priming that have emerged over the years, but doing so under the auspices of two theories which have developed at least partially to accommodate the unreliability of comprehension priming. For the purposes of this dissertation, these theories will be referred

---

<sup>6</sup>Although the path of access might differ in crucial ways. We return to this point below.

to as Syntactic Adaptation and Lexically-Filtered Comprehension. For each theory, we provide an overview of the theoretical machinery and then consider potential instances of comprehension priming that already exist in the literature and that would be predicted by that theory.

One important note: many, in fact probably most, of the case studies that we consider in this section were not run with either Syntactic Adaptation or Lexically-Filtered Comprehension in mind. The authors of these studies proposed their own theoretical accounts for their results, many of which go a long way to filling gaps left open by Syntactic Adaptation and Lexically-Filtered Comprehension. The goal of this section is to take two, explicit theories which have much in common with other proposals in the literature but which contrast with each other, and evaluate the evidence that is consistent with them as a way to consider that state of the field overall. In several cases the different theoretical stances of the authors will be discussed alongside the studies themselves, but it also bears saying upfront that these differences exist, in order to avoid misrepresentation of the original authors.

#### **4.3.1 Special Cases**

We actually begin with two instances that are claimed to be comprehension priming, but which are not predicted by either of the theories.<sup>7</sup> They are nonetheless worth mentioning for several reasons. First, they were among the earliest examples of potential instances of comprehension priming to emerge after Branigan et al. (2005)'s findings. As such, they have had a significant impact on the subsequent literature. Second, because they do not fit clearly into either of the theoretical frameworks that we take up later in the chapter, they serve to frame an important issue which is not always discussed explicitly: namely, what types of effects in comprehension should be considered to be the correlates of the original production effects? Is any comprehension measure that is sensitive to prior syntactic experience sufficient, or is more required? If the former, the two proposed instances of comprehension priming in this sub-section might be seen as challenges to the updated theories. If the latter, then there is a need for principled and *detailed* criteria determining what is priming and what is not.

##### **4.3.1.1 Thothathiri & Snedecker (2008)**

Thothathiri and Snedecker (2008a) is widely credited as being one of the very first studies to have found indications of abstract priming in comprehension. Their experiment in many ways

---

<sup>7</sup>Critically, the term 'Special Cases' is only relevant to the particular grouping pursued in this paper and is not a judgement on these results overall.

mirrored the visual world task of Arai et al. (2007), with a few critical changes. First, the visual scene was physically constructed by placing toys on one of four shelves in the participant's field of view, rather than being presented as an image on a screen. After participants listened to an auditorily-presented dative sentence, they were asked to physically act out the event using the toys (e.g., *Show the horse the book*). Using this procedure, Thothathiri and Snedeker demonstrated a priming effect such that participants who were primed with a DO dative were more likely to anticipatorily look to the potential goal (*horse*), whereas participants who heard the PO prime looked more to the potential theme (*horn*) before the disambiguating phonological information.

The first experiment that demonstrated this pattern of results had participants act out both the prime and the target, while Thothathiri and Snedeker's experiments 2 and 3 included the act out procedure only for the target sentence. A second paper, Thothathiri and Snedeker (2008b), then went on to replicate the effect in children. All this held without any overlap in content words between the prime and the target, i.e., exactly the effect that Arai et al. had failed to find.

Thothathiri and Snedeker's results are a critical demonstration of how comprehension-to-comprehension effects might be realized for dative structures: that is, as anticipation of the identity of the first argument after the verb (similar in this sense to Scheepers & Crocker, 2004). However many authors, including Thothathiri and Snedeker themselves, have noted that the critical change from previous studies seems to be the addition of what appears to be a production-like task. Thothathiri and Snedeker's answer to this is to remove the act-out task from the prime trials in experiments 2 and 3. However, these experiments continued to use the act-out procedure for the target sentences. If the act-out task is seen as production, this would still make the task in Thothathiri and Snedeker comprehension-to-production priming, which has been well-attested elsewhere (Gurevich et al., 2010). While it would be interesting and relevant in its own right if acting out an event counts as production for the purposes of abstract priming, it remains unclear whether Thothathiri and Snedeker (2008a) should count as a true example of comprehension-to-comprehension priming.<sup>8</sup>

An intermediate approach is to suppose that the imperative construction is the critical feature of these experiments that leads to comprehension priming. If this were the case, then the action of acting out the task need not count as production *per se* (which might be an uncomfortable state of affairs for anyone who wanted to view abstract priming as being about syntactic knowledge

---

<sup>8</sup>An alternative is to view the task as more of a production-to-comprehension priming task, which would make an important counter-example to the Arai et al. (2007) study in particular.

directly). Instead, comprehending the imperative and using it to inform a motor plan may engage some additional processes beyond normal comprehension (potentially prediction, although there are other possibilities). If these additional processes were susceptible to priming, then Thothathiri and Snedeker's results would be true comprehension priming. On the other hand, this would not say anything about non-imperative comprehension and might or might not be priming in the sense of Bock (1986).

#### 4.3.1.2 Coordination Priming

A second case where comprehension priming has been claimed to exist is in the processing of coordinate structures. It is known that in order for coordination to be grammatical, the coordinated elements need to be sufficiently similar. While what counts as "sufficiently similar" remains a matter of debate, it is not, as was first thought, a requirement that the syntactic category of the coordinating constituents match (Munn, 1993). However, in an eyetracking while reading study, Frazier, Munn, and Clifton (2000) showed that there was still a processing advantage for the second coordinand when it matched the syntactic structure of the first relative to when the second phrase was processed outside of coordination. A representative example of their stimuli sentences is given in (46).

(46) Hilda noticed { a strange man/a man } and a tall woman when she entered the house.

Because syntactic category matching is not a grammatical requirement, Frazier et al. (2000) concluded that the facilitation must be the result of a processing advantage for like syntactic structures. Frazier et al. went on to show that subjects did not prime the structure of objects, indicating that coordination, and not mere proximity, was key. Apel, Knoeferle, and Crocker (2007) report very similar results for German.

While this could be seen as abstract priming, Frazier et al. (2000) point out that the relationship to coordination makes it an odd fit. On the one hand, it is coordination in particular that seems to make structural facilitation across matching structures possible. On the other hand, doing so is not actually about fulfilling the requirements for coordination. That is, the effect is highly construction specific, but does not actually engage with any of the requirements of that construction. It is not clear that one would expect this sort of grammatical specificity in a purely processing effect like priming. Nor can the effect be grammatical in origin because there is no grammatical requirement to enforce. Frazier et al. themselves suggest that their effect is perhaps better characterized as parallelism rather than priming. They base this hypothesis specifically on the fact that priming

should have been able to cross from subject to object while parallelism is more sensibly restricted to specific grammatical or discourse contexts.

A counter-argument comes from Sturt, Keller, and Dubey (2010), who suggested that priming might be blocked by the different case values of the prime and target NPs in Frazier et al.'s subject/object study. Sturt et al. instead tested for priming between main clause and embedded clause subjects, and found that the underlined region of (47a) was read faster than the corresponding region (47b), as they predicted.

- (47) a. A boss who was demanding said that a worker who was lazy did not do the job properly.  
b. A demanding boss said that a worker who was lazy did not do the job properly.
- (48) a. A demanding boss and a worker that was lazy did not do the job properly.  
b. A boss who was demanding and a worker that was lazy did not do the job properly.

A further advantage of Sturt et al.'s design was to include a direct manipulation of subordination versus coordination, (47), within the same experiment, where prior work had only compared between experiments. Again, the critical region in (48b) was read faster than in (48a).

Sturt et al. (2010) shows facilitation between non-coordinated elements, but they do not clearly answer *why* case mismatch should prevent priming. Priming is possible between other ways of differentiating arguments, for instance PO datives prime benefactives (Bock, 1986) and locatives despite different thematic roles (Bock & Loebell, 1990). Moreover, Frazier et al. (2000) gives no reason to believe that their parallelism account should necessarily be limited to coordination. It might well extend to other processing conditions which highlight grammatical or discourse similarity.

If the effect in these studies is ultimately to be considered priming, then it is far more restricted than production priming. More work is needed to determine the extent to which grammatical factors and/or proximity of the prime and target dictate why 'priming' appears in coordination and not freely elsewhere.

#### 4.3.1.3 Discussion of Special Cases

The Special Cases consist of Thothathiri and Snedeker (2008a) and coordination priming. These results stand out from other potential instances of comprehension priming not just because they fit less clearly into our two theoretical frameworks, but also because of the relationship between the

empirical findings and priming overall. For most of the studies that will be reviewed in this chapter, any confirmed effect is almost certainly an example of comprehension priming. Thothathiri and Snedeker (2008a) and coordination priming are the standout exceptions as both are currently open to alternative explanations. Moreover, they are notable in the sense that if coordination priming or the act-out task turned out to be the *only* true examples of comprehension priming, this would indicate that the distribution of priming in comprehension is vastly more limited than it seems to be in production. Consequently, it would argue in favor of substantially more restricted mechanisms behind priming in comprehension, although the act-out task and coordination differ in what the exact restrictions would be. For Thothathiri and Snedeker (2008a), the permitting factor seems to be the co-occurrence of the linguistic task with a specific non-linguistic one. For coordination priming, the exact nature of the restriction remains somewhat unclear. Section 4.3.1.2 argued that what has been called coordination priming has both grammatical and proximity constraints on its distribution. The more influence that proximity has over when “coordination priming” does and does not occur, the more that the effect looks like priming, which would simply be weaker and therefore need more support from recency in order to be realized than in production. If the grammatical component turns out to be key, such that facilitation only transfers between constituents that are sufficiently grammatically related, then this variety of facilitation is rather unlike production priming in that respect. In fact, it is under these circumstances that a parallelism account becomes more favorable.<sup>9</sup> It is the combination of grammatical and processing restrictions that make a parallelism account attractive to begin with, since production priming does not clearly have such grammatical restrictions. The members of the Special Cases class are just that: they imply that instances of comprehension priming are rare and in many ways quite unlike their production cousins.

#### 4.3.2 Syntactic Adaptation

The first of the two theories of comprehension priming considered in this chapter is Syntactic Adaptation. We will spend some time on the details of this theory, as it will be a critical part of the motivation for Experiments 3 and 4. The particular type of Adaptation discussed here is modeled on Fine et al. (2013)’s rapid, statistically-sensitive syntactic Adaptation. Adaptation lies squarely in the Implicit Learning family of priming theories. Generally speaking, the basic parsing

---

<sup>9</sup>Notably, Frazier et al. conceive of parallelism as a purely processing effect and this could still be true. There is a difference between taking advantage of the processing conditions which are rather accidentally restricted to just a few syntactic constructions versus the parser engaging with grammatical knowledge to fulfill grammatical requirements or preferences.

mechanism in Adaptation is incremental prediction of up-coming input, based on the input string that has been comprehended so far and on frequency in the environment: the more frequent a structure is, *as a continuation for the initial string*, the more expected it is.<sup>10</sup> Expectation, in turn, determines processing difficulty via two distinct but closely intertwined probabilistic processes.

One of these processes occurs as part of the prediction during online processing of individual sentences. Adaptation is a parallel processing theory, meaning that the parser entertains multiple predictions about the upcoming input at any one time, ranked according to their probability. For instance, if the parser encounters a dative-alternating verb that has a strong preference for the PO form, then it will entertain separate predictions for both the PO and DO forms, but will be more strongly committed to the PO prediction. If the most predicted structure is confirmed,<sup>11</sup> the parser has no additional load and processing is relatively easy. Processing difficulty occurs when new input favors a previously dispreferred analysis and forces the parser to change the relative probabilities assigned to potential parses in order to accommodate the demands of new information. Continuing the dative example, if the material following the verb turns out to be a DO frame, then the parser will be surprised and will have to expend effort to revise its parse at the point where it encounters evidence for the DO (certainly at the onset of the second NP, and sometimes during the initial NP after the verb, Bresnan, Cueni, Nikitina, & Baayen, 2007). This effort is realized as longer reading times, more pronounced ERP components, lower acceptability judgements, or other indices of processing difficulty.

The re-ranking of potential structures occurs during the online processing of a single sentence, and is not Syntactic Adaptation. Rather, Adaptation is the second process, which supplies the frequencies that determine the strength of expectation allocated to a parse. After a structure has been built and confirmed as the most likely final parse for a given sentence, there is a small increment in the overall probability associated with that structure in the comprehender's store of grammatical knowledge. The probabilities in the comprehender's knowledge of their grammar then go on to inform the input probabilities during processing of future sentences. This incremental updating of frequencies based on experience results in Adaptation over the long-term: more frequent struc-

---

<sup>10</sup>Technically, most of the theories in this family predict at the level of the upcoming lexical item, more than abstract structure (see e.g., Levy, 2008, for a widely-adopted example). For instance, predictions will not simply be for a member of the syntactic category *Verb*, but rather for a specific verb with all its attendant lexical-specific implications for following input. However, it is possible, and indeed very widely done, to discuss the predictions about abstract structure because abstract structure is a distinct component that contributes to how surprising (or not) a lexical item is in context. The syntactic component of how expected a word is is essentially factored out from all the other impactful properties (semantics, overall frequency, etc.).

<sup>11</sup>Or at least not contradicted, in cases of global syntactic ambiguity.

tures in the environment are allocated increasingly more expectation, and being expected makes them increasingly easier to process.

A final element of this system is competition between syntactic parses during Adaptation itself, not just during online comprehension. Because the currency of Adaptation is probability, which is a limited resource, the frequency that matters in Adaptation is not absolute frequency, but the relative frequency of the eventual structure versus all the other structures which could have been continuations of the same initial string.

To demonstrate how the whole ensemble works together, consider the dative sentences in (49):

- (49) Odessa gave...
- a. ... Quentin a cake.
  - b. ... a cake to Quentin.

When a parallel predictive parser reaches *gave*, it will then predict both the DO in (49a) and the PO in (49b). For the sake of this toy example, suppose that the parser allocates a probability of 0.6 to the DO parse and 0.3 to the PO parse.<sup>12</sup> If the parser ultimately encounters (49b), it will need to do the work of reallocating probabilities *during* the processing of this particular sentence. However, because the PO dative (specifically with *give*) is now just slightly more frequent in the parser's overall experience than it was before (49b), the parser will also update the frequencies for datives *beyond* this particular sentence. Specifically, it will allocate slightly more probability to PO parses and the increase in probability for PO parses will necessarily be taken from what had been probability allocated to the DO parse (or lower probability alternative parses). The new, resulting probabilities could now be 0.599 for DO datives and 0.301 for the PO frame.<sup>13</sup> These would become the new input probabilities for the next dative sentence, particularly one with *give*. If the parser then encounters another PO dative, it will be just slightly easier to process than if the comprehender had never seen (49b). On the other hand, if the next dative that the parser encounters occurs in the DO frame, that sentence will be very slightly *harder* to process than if the parser had never encountered (49b). If the original input had instead been (49a), the same probability-shifting competition would be in effect, except in the DO parse's favor. Because the

---

<sup>12</sup>The probabilities do not sum to 1 because some probability will be allocated to alternative parses that aren't of interest for this example.

<sup>13</sup>In point of fact, some of the shift will occur in the overall frequencies of DO and PO frames, and some will occur in the probability of *give* in particular occurring with either the PO and DO structures. Consequently, if the next dative sentence that the parser encounters has a different verb, the change will be slightly less, while another dative sentence with *give* would feel the full effect of the probability shift.

parser has quite a bit of experience before encountering (49), the amount of probability shift due to any individual sentence will be quite small, but systematically-biased input over time can add up to large shifts in the ease processing a particular structure. This is implicit learning of the favored structure, but always at the expense of the competing structures.

The actual magnitude of difficulty due to updating predictions during online comprehension in Adaptation, (and therefore the relevant scale for observable differences between structures), is typically determined on a Surprisal scale. The Surprisal of an event is the negative log probability of that event in context (Levy, 2008). Applied to sentences, the difficulty incurred during processing of the current word,  $w_i$ , is the negative log of the probability of  $w_i$  in the context of the string that the parser has already seen (i.e.,  $w_1...w_{i-1}$ ).

$$\text{difficulty} \propto -\log P(w_i | w_1...w_{i-1}) \quad (4.1)$$

(Levy, 2008, pp. 1130)

While the precise reasons for choosing Surprisal over any other distribution are beyond the central interests of this dissertation,<sup>14</sup> there is one additional consequence of this choice which is quite important to this chapter and the next. Because relative difficulty is computed on a log-scale, rather than, say, a linear one, means that the observed change in processing difficulty will be different at distinct points on the scale for equivalent increments in frequency. Specifically, less frequent structures incur a larger shift in difficulty per encounter, while more frequent structures shift less for the same change in absolute frequency. With this observation in hand, we finally have all the pieces necessary to evaluate what kinds of priming this theory will predict and what it will not.

Syntactic Adaptation essentially *is* comprehension priming. A true lack of comprehension priming would be fatal to this theory. Moreover, Adaptation predicts that comprehension priming is always occurring, with every construction, every time it or a related structure is encountered. However, Adaptation also predicts that the amount of priming from any single sentence should be very small, particularly with frequent structures like either of the dative frames. This means that Adaptation is still compatible with the apparent lack of comprehension priming in Branigan et al. (2005), Arai et al. (2007), and others, because the probability-shift in these studies could easily be too small to detect on a trial-to-trial basis. According to Adaptation theory, the way to find

---

<sup>14</sup>The interested reader is directed to the explanation in Levy (2008).

comprehension priming is to maximize the probability shift between the prime and the target. The remaining parts of this section look at cases in the literature that fit that description.

#### **4.3.2.1 Cumulative Priming**

The first potential case of comprehension priming in this section that we will consider is cumulative priming. Cumulative priming is a direct prediction of Implicit Learning theories, especially Adaptation. The core idea is that if comprehension effects are simply too weak to be measurable on a trial-to-trial basis, one way to detect an effect should be to measure the effect of many primes combined together. In a typical psycholinguistic experiment where the same structure is repeated many times, cumulative priming would manifest as a gradual effect of Order, making each instance of the structure slightly easier to process than the ones before it. The facilitation between any two tokens might be too slight to be reliable, but the difference between the first and last token could still be significant. In fact, the weakness of comprehension priming may be seen as an asset from an implicit learning point of view. Ideally, long-term learning should slowly adjust to new environments rather than dramatically changing behavior based on only a few examples, which could lead to overcorrection. Gradual adjustment due to the cumulative effects of many similar sentences is a more stable learning strategy.

One of the first studies to pursue the cumulative priming approach experimentally was Wells et al. (2009). Wells and colleagues made use of a well-known reading time penalty for object relative clauses (ORCs) compared to subject relative clauses (SRCs) (Gordon, Hendrick, & Johnson, 2001; King & Just, 1991). Wells et al. reasoned that implicit learning, if it exists, should ameliorate the ORC penalty with practice. This should be particularly true because ORCs are less frequent than SRCs and therefore it is easy to manipulate their frequency in context to shift expectation from SRCs to ORCs. During four sessions over the course of several weeks, participants were presented with an equal number of ORCs and SRCs using self-paced reading, and then reading times were compared from a pre-test on the first session to a post-test on the last. They found a decrease in reading times across the two blocks, indicating facilitation due to experience for both ORCs and SRCs but particularly pronounced for ORCs. They took this to be support for their version of the implicit learning theories and especially competition between rare and frequent structures. That is, since ORCs were the less frequent/practiced structure, the associated reading times had the most room to improve.

More recently, Fine et al. (2013) tested for Syntactic Adaptation to the MainVerb/RelativeClause garden path, given in (50).

- (50) a. **Main Verb:** The experienced soldiers warned about the dangers before the midnight raid.
- b. **Relative Clause:** The experienced soldiers warned about the dangers conducted the midnight raid.

Like Wells et al., Fine et al. used self-paced reading as the primary task, but Fine and colleagues were able report the reading time across all time points in the experiment. Also like Wells et al., they found an interaction of construction and Order, but this time with an important addendum: at the end of their experiment 2, Fine et al. saw the RT curves of the Main Verb and Relative Clause constructions cross, with the result that at the end of the experiment the Main Verb reading was actually harder than the Relative Clause reading. This is a crucial prediction that distinguishes statistically-sensitive Adaptation theories from Implicit Learning and Activation theories. The prediction arises from the competition for expectation between like-structures. As expectation for the Relative Clause continuation increases in the context of the experiment, Adaptation theories necessarily decrease the expectation allocated to Main Verb readings. Decreased expectation leads to increased reading times. If two competing constructions were primed in such a way as to flip the unexpected reading to become the expected reading, then the Order curves of the reading times would cross over each other: this is exactly the pattern that Fine et al. saw. The increase in reading times for the Main Verb reading would not be predicted because activation can only facilitate RTs and only on the activated structure. This effect makes Fine et al.'s finding among the best evidence for statistically-sensitive Adaptation to date.

On the other hand, each of these studies has its limitations. Wells et al. only compared the pre- and post-test RTs, and therefore were not in a position to evaluate whether adaptation occurs gradually and incrementally over time, as predicted by both Activation and Adaptation theories. In addition, they used self-paced reading, which is known to be vulnerable to task adaptation. Natural reading is among the most practiced tasks that typical experiment participants would share. By the time they reach college, most adults readers have reached a sufficient amount of expertise to converge on the most efficient reading strategies (Rayner, Pollatsek, Ashby, & Clifton, 2012). Self-paced reading, on the other hand, is a novel task and participants' strategies may vary between each other, between experiments and even between trials (e.g., Adams, Clifton, & Mitchell, 1998). While self-paced reading has been shown to approximate natural reading for many measures, task adaptation is a particular confound when the measure of interest is also adaptation over the course of the experiment. This means that while the Wells et al. findings are

certainly important, it would be reasonable to follow-up to ensure that alternative explanations are not at play.

Fine et al. (2013) come with another set of concerns. Recall that they also use self-paced reading and therefore confound the two types of adaptation. Because it was central to their predictions, Fine et al. do analyze the data throughout the experiment and do successfully demonstrate the predicted gradual cline in facilitation of reading times. However, their choice of construction is potentially problematic. MainVerb/RelativeClause garden paths are a particularly difficult structure for comprehenders. Anecdotal evidence suggests that the difficulty and subsequent repair can rise to the level of conscious strategy. It is certainly the case that comprehenders can encounter a grammatical Relative Clause garden path for which they fail to arrive at an acceptable parse (take Bever's famous *The horse raced past the barn fell* sentence; Bever, 1970, p. 316). This level of difficulty makes MainVerb/RelativeClause garden paths stand out from normal sentence processing. It may be the case that for extraordinarily difficult constructions like this, the learning process is not the same. In particular, the failure to arrive at a grammatical parse may indicate that the Relative Clause structure is not even reliably available as a repair structure. Comprehenders would first have to learn that this is a possible structure at all and *then* learn to use it, potentially as a completely conscious strategy. This means that the learning path, and therefore potentially the learning mechanism, for Relative Clause garden path readings may not be indicative of the learning path for easier structures.

These concerns are undoubtedly related to a series of high-profile failures to replicate Fine et al. in Dempsey, Liu, and Christianson (2020); Harrington Stack, James, and Watson (2018); McCann and Kaan (2017) or partial replications which argued that what appeared to be Syntactic Adaptation is better understood as a task effect (Prasad & Linzen, 2020).<sup>15</sup> As most of these publications included more than one experiment, between them they constitute nine replications that either failed or demonstrated an alternative explanation was more plausible. Harrington Stack et al. (2018) made a particular effort to replicate Fine et al. (2013) with the number of participants required for adequate power and still fail to find the critical facilitation effects. Meanwhile, Prasad and Linzen (2020) find an effect which is similar to Syntactic Adaptation, but argue that the bulk of the effect was better understood as adaptation to the experimental task.<sup>16</sup> These attempts at

---

<sup>15</sup>And see the following chapter of this dissertation for another.

<sup>16</sup>Prasad and Linzen do eventually find an effect which they identify as Syntactic Adaptation, on top of the task adaptation that they find, but they note that it required 800 participants to do so. This number of participants is well outside the range for most studies purporting to find Adaptation, and is a further indicator that even those studies which do replicate Fine et al. (2013) run the risk of identifying a task effect over true implicit learning of syntactic structure.

replication suggest that the concerns about vulnerabilities in past cumulative priming findings are not idle.

Because cumulative priming is such a central prediction of Implicit Learning theories—and rapid, statistically-sensitive Adaptation in particular—it is critically important to establish whether it is realized or not as a way to move the theory forward. As this section has shown, there is a considerable body of evidence in favor of Adaptation already, but also some remaining concerns.

#### 4.3.2.2 Mere Exposure and Syntactic Satiation

The next candidate for comprehension priming that would be predicted under Adaptation comes from a very different initial theoretical background. Luka and Barsalou (2005) called this type *mere exposure*, from the fact that experiments of this variety consist of just two parts: an exposure block and a test block. In the exposure block of their Experiments 2 and 3, participants silently read several types of rare and even ungrammatical sentences and in the test block they provided acceptability ratings for sentences with identical underlying syntactic structures. Luka and Barsalou found that acceptability for the critical sentence types increased for participants who had experienced them in the exposure trial over participants who had not: in other words, a priming effect that manifests in acceptability.<sup>17</sup>

This type of phenomenon seems to have correlates that reach well-beyond linguistic processing. Luka and Barsalou themselves take the term *mere exposure* from Zajonc (1968), who compared mere exposure for both words and images. Through this theoretical lens, Luka and Barsalou argue that the critical feature of their study that permitted priming was the mere exposure task, and that other comprehension tasks might not. This view is quite separate from Syntactic Adaptation, as Adaptation would not predict this kind of task-specificity.

However, while the mere exposure account/task specificity is not necessarily predicted by Adaptation theories, the results of Luka and Barsalou (2005) are. The critical stimuli used in Luka and Barsalou (2005) were either very rare or actually ungrammatical for participants, and therefore the critical strings contained transitions with extremely high Surprisal. If Luka and Barsalou's findings are recast in Fine et al.'s account, their results would be expected on the basis that exposure to rare constructions quickly makes those constructions easier to comprehend. There would also need to be some additional mechanism that makes constructions which are easier to comprehend more acceptable, but this mechanism is independently required for a large

---

<sup>17</sup>Luka and Choi (2012) showed that when participants read the prime sentence aloud (production-to-comprehension priming) the structural facilitation effect could persist up to a week, but did not test silent reading of the prime.

segment of acceptability-based research. The apparent task-specificity would have resulted from the fact that mere exposure studies have preferred acceptability judgments to other comprehension measures. To our knowledge, it remains unclear whether the deciding factor in Luka and Barsalou (2005)'s findings is the mere exposure task itself or the low frequency of their critical items. If frequency turns out to be the primary factor, Luka and Barsalou's results would provide additional support for an Adaptation account.

There are a few studies that use a procedure similar to Luka and Barsalou (2005), but which explicitly take a theoretical approach that is far more similar to Adaptation, such as Kaschak and Glenberg (2004). Kaschak and Glenberg ran four self-paced reading experiments exposing native English speakers to the *needs X'ed* construction. An example of this construction is given in (51a), with the Standard American English equivalent in (51b) and a control with a matched initial string in (51c).

- (51) a. The meal needs cooked given that dinner is in an hour.  
b. The meal needs to be cooked given that dinner is in an hour.  
c. The meal needs cooked vegetables to make it complete.

(Kaschak & Glenberg, 2004, p. 451)

According to Kaschak and Glenberg, the *needs* construction is a feature of the Northern Midlands dialect, and would have been entirely novel to the speakers from Wisconsin and Minnesota who participated in the experiment. Across their studies, they found evidence that participants who were exposed to the *needs* construction in the initial block of the experiment read the construction faster in the second block than participants who had no exposure to the *needs* construction before the second block. The exposure paradigms varied between the studies, and ranged from listening to recorded dialogues and rating them in an unrelated task (how "friendly" the conversation was) to reading them in the self-paced reading task itself. They also found some evidence that exposure to (51a) impacted reading times for the control condition in (51c), although in at least one study, reading times were actually faster for the control condition after exposure to the *needs* construction (the reverse of what would be predicted by a competition-based account like Adaptation).

Kaschak and Glenberg (2004) is interesting in part because methodologically and theoretically it lies in between the cumulative priming studies discussed in the last section, and the *mere exposure* study in Luka and Barsalou (2005). Given that Kaschak and Glenberg did not need much more exposure than Luka and Barsalou, but still found exposure-based facilitation outside of ac-

ceptability using a very rare construction, the Kaschak and Glenberg results may be an indication that the rarity was the key ingredient to priming effects in Luka and Barsalou rather than the acceptability task.

Mere Exposure may also be functionally identical to a phenomenon known elsewhere in the literature as “syntactic satiation” (Snyder, 2000), which originated from the observation that professional linguists tend to find initially unacceptable constructions more acceptable the longer they work on them.<sup>18</sup> Like Luka and Barsalou (2005)’s mere exposure, it relies on exposure and takes acceptability as the dependent measure, with the primary difference being that satiation studies tend to include an acceptability judgment task for all trials, with no separate ‘exposure’ and ‘test’ blocks.<sup>19</sup> Satiation has been studied experimentally a number of times, including a 2017 study by Do and Kaiser which specifically investigated whether satiation shares some of the classic characteristics of priming, specifically a decreasing size of the priming effect as distance between the prime and target increases and the existence of the lexical boost. Their results indicated that satiation did not exhibit truly priming-like behavior for either property. Additionally, not all syntactic violations appear to participate in satiation equally (Do & Kaiser, 2017; Snyder, 2000; Sprouse, 2009). It is unclear whether this distribution across structures would constitute a divergence from priming-like behavior, as the constructions that have been investigated in satiation research are rare in other areas of the priming literature.

Satiation is an interesting addition to the family of potential priming and priming-like effects, but a review in Sprouse (2009) found that syntactic satiation findings are not robust to replication. Moreover, Sprouse included a series of studies indicating that the original satiation findings are equally well explained by a task-based strategy. The particular task-based strategy implicated by Sprouse would not encompass all of the studies in this section; indeed he makes a point of saying that his task strategy is a only a property of certain acceptability study designs, not acceptability judgments in general. However, it is close enough to indicate that there may be other strategies that play a role. Thus syntactic satiation (and perhaps mere exposure by extension) suffers from two potential characteristics that make it a non-ideal candidate for comprehension priming: (i) a set of characteristics that may differ from the processing pattern associated with classic priming and (ii) an alternative explanation as a task-effect.

---

<sup>18</sup>Sometimes also affectionately called “judgmentitis”.

<sup>19</sup>In this sense, satiation is reminiscent of the design of many cumulative priming studies.

#### 4.3.2.3 Item Bias

The last potential case of comprehension priming in the Adaptation section is Item Bias. It has been placed last in the section specifically because it is also predicted by the second theory of comprehension priming that we will consider, starting in Section 4.3.3.

A 2014 paper by Kim et al. suggested that at least some previous attempts to find comprehension priming, such as Arai et al. (2007), may have failed because they did not take into account item-level/lexical biases at a sufficient level of detail. Kim et al. specifically target these biases as a means to reduce noise that can drown out the signal from priming. Kim et al.'s sentences used PP attachment ambiguities, and were therefore similar to prior studies that had *not* found comprehension priming. However, unlike most previous studies Kim et al. created ambiguities by attaching into one of two NPs, the direct or indirect object, rather than into the verb phrase (see the example prime and target in [52] and [53] respectively, cf. [44] from Branigan et al., 2005). Because verbs are a major source of lexical bias in most items, this was intended to mitigate some of the bias impacting the target manipulation. In addition, Kim et al. pre-normed their stimuli for attachment bias with a separate group of participants, and then included those norming scores directly into the mixed-effects model as a factor. Sentences were presented via self-paced reading during the experiment itself and the attachment that a participant had pursued in a target trial was directly probed by a comprehension question, such as the one in (54). Prime sentences made use of world-knowledge bias to force a high or low attachment reading.

(52) High Attachment Prime: The gardener watered the tree with the bird's nest with tangled roots.

(53) Target: The FBI agent noticed the mirror on the wall with the crack.

(54) Question: What had a crack? (a) the wall (b) the mirror

(Kim et al., 2014)

With item-level attachment bias controlled for in this way, Kim et al. show that there was a significant effect of prime type on the final interpretation as indicated by responses to the question in (54). However, when the Item Bias factor was removed from the model, the priming effect was no longer significant. A corresponding priming effect was also found in reading times for low-attachment primed trials. This, then, appears to be evidence in favor of theories which suggest that abstract syntactic priming exists in comprehension but is typically overwhelmed by lexical and item-specific biases. However, because of the way that Adaptation theories are implemented, i.e., based on word-by-word Surprisal and therefore highly susceptible to influence of lexical-level

variation between items, Kim et al. argue that their results are also consistent with Adaptation theory.

By accounting for the pre-existing attachment bias inherent in individual sentences, the Item Bias account maximizes the Surprisal associated with any given item and does not dilute the final results by aggregating across items with wildly different Surprisals for their target continuations. This is exactly the kind of intervention that is predicted to yield significant priming on a Syntactic Adaptation account.

### 4.3.3 Lexically-Filtered Comprehension

The second theory of comprehension priming to be considered in this chapter comes from a line of thinking that argues that the lack of comprehension priming may be derivable directly from known differences between processing in production and comprehension.<sup>20</sup> In this vein, Tooley and Bock (2014) propose that the asymmetry between production and comprehension priming may be explained entirely by how the two modes determine syntactic form. Tooley and Bock don't give a name to their theory, but for us it will be convenient to refer to it as the Lexically-Filtered Comprehension hypothesis.

Like many other theories, arguably including the broader Implicit Learning family of theories, Tooley and Bock (2014)'s account works well with a framework in which the general architecture that is shared between production and comprehension consists of at least a lexicon and a cache of syntactic knowledge, similar to the one outlined in Pickering and Branigan (1998). In systems like this, the syntactic cache is discretized into rules or treelets of essentially the same type as the ones discussed for Activation theories in Section 4.1. Moreover, there are links between the lexicon and the syntactic cache based on which lexical items can appear with which syntactic frames, and how frequently they actually do co-occur. A verb, for instance, would be connected to all of its subcategorization structures with stronger connections to more frequent frames.<sup>21</sup>

Tooley and Bock (2014) then base their account on the fact that, in production, lexical content and syntactic structure are co-determined in parallel and are only partially dependent on each other (Bock & Levelt, 1994). In comprehension, on the other hand, the comprehender must first

---

<sup>20</sup>Some of these theories are not intended to address all priming under the same umbrella. For instance, Hartsuiker, Kolk, and Huiskamp (1999) suggest that certain instances of priming may be due to linearization processes in production specifically.

<sup>21</sup>We use "works well with" to highlight the fact that Tooley and Bock (2014) does not spell out this part of the architecture in detail and are almost certainly not committed to all of the specifics that we have included in order to make the rest of the Lexically-Filtered Comprehension proposal clearer in context.

parse the string into lexical items and then use these as the basis for determining syntactic structure. Since in production the processor can make syntactic choices independent of lexical content, it is possible for the outcome to change the weights of rules within grammatical knowledge independent of the lexicon (whether weights are in terms of activation, probabilistic expectation, or something else makes little difference). In comprehension, access to grammatical knowledge is mediated through the lexicon. As a result, Tooley and Bock propose that re-weighting generally occurs over the links between lexical items and grammatical rules, rather than for pure syntactic representations, yielding the lexical boost without abstract priming. Intriguingly, current evidence indicates that the lexical boost persists longer in comprehension than in production (2-3 trials versus 1), potentially consistent with an increased role for lexical input (Branigan et al., 2005; Tooley, Swaab, Boudewyn, Zirnstein, & Traxler, 2014).

However, traditional priming constructions such as datives would be one of the worst places to look for comprehension priming, because each dative verb has its own unique level of preference for one frame or another. These preferences indicate that syntactic choices at this level are heavily arbitrated by lexical information. Put another way, knowing the particular dative verb provides considerable information about whether the sentence will ultimately turn out to use a PO or DO frame and this information is far more specific than what would be provided by the weightings of the two frames in the abstract grammatical representation.

At first blush, it might appear that Lexically-Filtered Comprehension's ban on abstract priming is absolute, but this is not necessarily so. In point of fact, the theory predicts that abstract priming can and should occur wherever lexical bias is weakened. Even in comprehension, the parser will need to make decisions about syntactic structure in the absence of clear lexical preference. To the extent that the empirical literature finds evidence of comprehension priming under conditions of reduced lexical influence, and not other situations, those findings would support the Lexically-Filtered Comprehension view of priming.

In the sense that both Adaption and Lexically-Filtered Comprehension rely on interference from lexical specific information to understand why abstract priming does not reliably obtain in comprehension, the two theories actually have quite a bit in common. The lexical-bias that Tooley and Bock (2014) refer to can arguably be seen as one way the statistical notion of Surprisal can be implemented in a cognitive architecture. For our purposes, the difference will be that Adaptation predicts that comprehension priming is on-going, all the time, with every construction. The reason that studies fail to find it is methodological: researchers simply lack tools that are sensitive enough to detect priming unless the Surprisal shift is unusually large. Lexically-Filtered Comprehension

predicts that abstract priming is highly restricted, occurring only in those conditions where lexical information does not provide the parser with sufficient guidance.

#### **4.3.3.1 Item Bias Reprise**

The same Item Bias results that were described in Section 4.3.2.3 are also predicted by Lexically-Filtered Comprehension, only adjusted to a different cognitive architecture. In Adaptation, Item Bias worked by avoiding noise from aggregating over items with different Surprisal in context. Under Lexically-Filtered Comprehension the idea is very similar, except the noise that can be factored out comes from the strength of individual lexical item biases. Depending on the particular model of the parser that one has in mind, these two things may be difficult to separate for Kim et al. (2014)'s findings. By including item-specific biases in their model, Kim et al. (2014) may also control the lexical bias the Lexically-Filtered Comprehension predicts would normally overwhelm abstract priming.

#### **4.3.3.2 Attachment Priming**

Not all grammatical restrictions on priming need be as limiting as the ones proposed for “co-ordination priming”. If the determining factor that blocks abstract comprehension priming is overwhelming lexical bias, as suggested by Tooley and Bock (2014), then a sensible strategy is to look for instances in the parse where lexical bias is generally weaker than it is for argument structure building from subcategorizations. A class of syntactic situations which fits this description is attachment ambiguity, of either a prepositional phrase or relative clause. In general, nouns are thought to exhibit somewhat less lexical bias than verbs, and attachment ambiguities need never be resolved, a promising sign that lexical content might have a weaker influence at these points in at least some examples.

Two sets of attachment priming findings have already been discussed in this chapter: Branigan et al. (2005), which did not find evidence of abstract priming, and Pickering et al. (2013) with the same stimuli and roughly twice as many subjects, which did (see Section 4.2). Lexically-Filtered Comprehension would predict that the Pickering et al. (2013) results would be the ones that would replicate, and that the original Branigan et al. (2005) findings were a Type II error of some kind. Moreover, the stimuli from Kim et al. (2014) reveal that this study would also qualify for attachment priming.

Additional evidence that attachment bias may be the right place to escape lexical bias comes from production results showing that it is impressively domain general. Scheepers et al. (2011)

demonstrated that correctly solving an equation with parentheses —either  $80 - (9 + 1) * 5$  or  $80 - 9 + 1 * 5$  —primed completion of a sentence like *The tourist guide mentioned the bells of the church that...* with a relative clause that was semantically consistent with the attachment in the prime equation. Scheepers and Sturt (2014) went on to show that such priming worked from sentences to equations just as well as from equations to sentences. Both studies argued that equation to attachment priming indicates that such priming occurs at a level of processing that was not specific to language per se. This would be an interesting enough claim in its own right, but it also serves as an indication that attachment may be the right place to look for reduced lexical bias.

A very different approach comes from Cuetos, Mitchell, and Coreley (1996), who had Spanish schoolchildren read stories with attachment-biasing sentences. Children were first sorted based on whether they had a high-attachment or low-attachment bias in an original questionnaire. They then read the stories over the course of two weeks, for a total of 60 attachment biased sentences. The critical sentences were always biased in the group direction (high-attachment sentences for the high-attachment bias group and low-attachment sentences for the low-attachment group). After an intervening week with no special exposure they again took a questionnaire that probed whether they tended to interpret ambiguous attachment sentences with high or low attachment. The new questionnaire showed a significant difference between the two groups and the high-attachment group had increased the strength of their high-attachment preference, although no corresponding increase in strength was found in the low-attachment group. Cuetos et al. (1996) actually take an implicit learning or cumulative priming approach to their own findings, in line with the impressive length of time that their effect persisted. However, Spanish speaking adults have a substantial high-attachment bias, meaning that the effect only held for the group that was exposed to the pre-experimentally more frequent structure, which is the opposite of what the implicit learning theories highlighted in this chapter would have predicted.

Lastly, Traxler (2008) found facilitated reading times in eyetracking with sentences such as (55) and (56) when both the prime sentence and target sentence were presented on the screen at the same time.

- (55) a. Same-structure prime: The chemist poured the fluid in the beaker into the flask earlier.  
b. Different-structure prime: The chemist poured the fluid into the flask earlier.
- (56) Target: The vendor tossed the peanuts in the box into the crowd during the game.

The critical finding was that reading times were facilitated when the prepositional phrase *in the box* was disambiguated to an attachment position inside the NP *peanuts* after (55a).

The combined interpretation of all of these results in the context of Lexically-Filtered Comprehension and Syntactic Adaptation is still somewhat unclear. Branigan et al. (2005) and Pickering et al. (2013) appear to be contradictory with respect to the existence of attachment priming. Traxler (2008) is promising as a potential instance of attachment priming, but on its own, it could be attributed either to priming or some version of the proximity condition already identified in the section on coordination priming, simply because the critical stimuli were presented so close together. The Kim et al. (2014) and Cuetos et al. (1996) findings are similarly predicted by both the Adaptation and Lexically-Filtered Comprehension accounts, which limits their utility in arbitrating between these two theories. However, all of these studies have played an important role in delimiting the shape of the two accounts that are available today.

#### 4.3.3.3 Verb Final

The predictions for attachment priming work by finding a place in the grammar where lexical bias can be reduced even in the final interpretation of the sentence. Another prediction of Lexically-Filtered Comprehension is that priming of argument structure, e.g., dative priming, should be possible if the bias of the verb could be removed from the parser's consideration at the moment when argument structure is first being assigned. Since the Lexically-Filtered Comprehension account suggests that abstract priming disappears in comprehension because lexical-syntactic preferences overwhelm more general, abstract syntactic biases, it predicts abstract priming should reappear when lexical information is impoverished or comes too late. Because the verb typically carries so much information about the role of other elements in the event/sentence, it is the verb that will provide the most substantial links between the lexicon and syntactic options for the sentence.

In an SOV language, unlike English, the verb comes sufficiently late that an incremental parser must make an initial decision about structure without the benefit of the argument structure preferences of the verb (Bader & Lasser, 1994; Kamide & Mitchell, 1999). To date, relatively few studies have addressed this prediction in SOV languages, partially because SOV languages are prone to have well developed case morphology which can cancel out the underspecification of argument structure that arises from the late appearance of the verb. What work there is, is reviewed in Arai (2012) and has a mixed character reminiscent of the rest of the comprehension priming literature. The most promising study is perhaps Arai and Mazuka (2014), which used a visual world proce-

ture with the active/passive alternation. In prime conditions participants saw an image of just two characters with one clearly the agent and another clearly the patient while listening to either an simple active or passive sentence describing the event. In the target condition, participants heard sentences with nearly identical structure (the examples in [57] are technically targets, but are identical in structure to the primes except that primes lacked an adverb). The target trials differed from the prime trials by having three characters in an image, depicting two events overall. For example, the image for the trial in (57) showed a pig, a monkey and a giraffe in a row, where the pig was grabbing the monkey and the monkey was poking the giraffe.

- (57) a. Saru-san-ga kyuuni kirin-san-o tsutsuita  
 monkey-HON-GA suddenly giraffe-HON-ACC poked.active  
*The monkey suddenly poked the giraffe.*
- b. Saru-san-ga kyuuni buta-san-ni tsukamareta  
 monkey-HON-GA suddenly pig-HON-DAT grabbed.passive  
*The monkey was suddenly grabbed by the pig.*

Because the two target sentences are identical up to the offset of the adverb *kyuuni*, Arai and Mazuka were able to use looks to the giraffe (patient of the active sentence) versus the pig (agent of the passive sentence) from the onset of *-ga* to the offset of the adverb as a measure of priming. Children developed a clear priming effect during the adverb, while adults developed the priming effect several hundred milliseconds earlier and showed an even greater looking preference based on prime. As there is was no lexical overlap between the prime and target in Arai and Mazuka (2014), this study appears to be a clear example of abstract comprehension priming in a verb final language without reliance on the lexical boost.

On the other hand, another paired set of visual world studies from Arai, Nakamura, and Mazuka (2015) using relatives clauses found the lexical boost but not abstract comprehension priming: exactly the pattern which is familiar from head-initial languages.

- (58) a. RC: Roujin-ga shinbun-o yondeiru bijinesuman-ni hanashikaketa.  
 old.man-NOM [newspaper-ACC reading] businessman-DAT talked.to  
*The old man talked to the businessman who was reading the newspaper.*
- b. Roujin-ga shinbun-o yondeiru toki, bijinesuman-ni hanashikaketa.  
 old.man-NOM newspaper-ACC reading when businessman-DAT talked.to  
*When the old man was reading the newspaper, he talked to the businessman.*

In the experiment that had verb overlap between the prime and target, looks to the relative clause head at the first verb *yondeiru* were increased when the participant had just encountered a relative clause prime relative to the subordinate clause prime, but no such effect held in the second experiment without verb overlap.

Arai (2012), commenting on earlier presentations of the same studies in Arai and Mazuka (2010) and Arai, Nakamura, and Mazuka (2011), attributes the difference between Arai and Mazuka (2014) and Arai et al. (2015) essentially to a difference in construction. That is, the difference in thematic/grammatical role assignment in (58) may crucially hinge on processing the verb because it amounts to changing the thematic roles in the event. The argument structure change in the passive need not be so dependent on the verb, perhaps because the passive is a regular operation that can be applied to the majority of verbs, or because the thematic roles remain stable. However, because the critical finding in Arai et al. (2015) amounts to a null result, an alternative way to reconcile these findings could be methodological, along the lines of Kim et al. (2014): that even without lexical overlap, item-specific biases swamp a true underlying abstract effect unless item-level biases can be accounted for directly in the analysis. Potentially consistent with this idea is the fact that the Prime x Experiment interaction that would have been strong support for the lack of abstract priming was marginal to non-significant. Thus, the status of abstract comprehension priming in verb-final languages is far from settled, but potentially promising.

#### 4.4 Discussion

There were three goals for this chapter: i) introduce the problem of missing priming in comprehension, ii) review potential instances of abstract comprehension priming in the literature and evaluate them, and iii) introduce the contrast between Adaptation theory and Lexically-Filtered Comprehension theory. This last goal is important because this theoretical contrast forms the basis for the experiments in the next chapter.

The central finding of interest is that abstract comprehension priming does not appear in all of the contexts in which it might have been expected, such as interpreting images of double object datives or attachment ambiguity sentences after having been primed with an unambiguous example, etc (Arai et al., 2007; Branigan et al., 2005; Ledoux et al., 2007; Tooley et al., 2009; Traxler et al., 2014).

This begs the rather important question of why comprehension priming was expected in these contexts in the first place. Especially early on, research on comprehension priming used both

notably different constructions and notably different methodologies than had been typically used in production priming. Some subsequent studies have been very conscious about narrowing this gap. Even so, a noticeable difference in the characters of the production and comprehension priming research literatures remains (though c.f. Tooley & Bock, 2014).

On the syntactic side, comprehension priming eschewed many of the primary criteria that had been used to pick constructions of interest in production. Production research focused on simple constructions that were easy to elicit reliably and which crucially minimized semantic differences between the alternatives. Minimizing semantic difference was a critical component of the argument that *syntax* was being primed, not some other element(s) of linguistic production. Meanwhile, comprehension work turned toward difficult-to-process syntactic alternations to meet the restrictions of comprehension methodologies. Dative constructions and passives are relatively fast and easy to read or comprehend. The available comprehension methodologies are more reliable with more exaggerated differences between the variants.

The methodologies themselves are a second source of critical difference. It is not immediately clear that an effect which began as an increased probability of production of a syntactic structure should necessarily have a correlate in reading times or visual world looking times. It would be easy to imagine that some of the current confusion about the distribution of priming arose from the difficulty of comparing across the two modes of processing.

Fortunately for the clarity of the results, there are good reasons to think that differences in construction or methodologies are not wholly responsible for the failure to produce comprehension priming. First, studies which directly addressed syntactic and methodological differences between the two modes conclude that there are still differences between comprehension and production priming, even when construction and methodology are controlled as closely as possible (Tooley & Bock, 2014). Second, the continued reliability of the lexical boost, even in environments that do not support abstract priming, suggests that comprehension methods and structures are not anathema to priming *in principle*. Rather, it really is *abstract* priming in comprehension that is the special case.

Nonetheless, there is no entirely consistent definition of what *should* count as comprehension priming. There are quite a number of potential instances of comprehension priming that have been reviewed in this chapter, but none of them are completely secure. At least some of the remaining uncertainty is because we still lack a consistent, cross-modal definition of priming to

go by.<sup>22</sup> Is it sufficient to say that priming is any instance in which prior experience impacts (facilitates?) processing of the same thing? The breadth of effects that have been put forward as comprehension priming —or indeed as any type of priming, not just syntactic —might seem to suggest that the answer in the literature is ‘yes’. Moreover, this the definition cited in a range of studies across a wide spectrum of the effects in this chapter, including Arai et al. (2007); Branigan et al. (2005); Pickering and Ferreira (2008); Sprouse (2009) and Tooley and Bock (2014). However, this definition leads no room for, for instance, a distinction between priming, parallelism, mere exposure, and syntactic satiation, which might in fact be useful. It also invites the very thorny issue of what counts as “same”. The different theories and approaches to comprehension priming in this chapter all have their own unique combination of answers to these questions.

The two instances of potential priming which this chapter labeled *Special Cases* particularly highlight the issue of what should (and should not) count as comprehension priming. In the case of Coordination, several authors argued that facilitation from one like-construction to another is not always priming. Instead, Frazier and colleagues argued for a second processing effect that facilitates between like-structures, which they identified as parallelism. Making this cut necessarily restricts the definition and distribution of priming to something less broad. Thothathiri and Snedeker (2008a) meanwhile found a facilitation effect that others have argued is not true comprehension priming, again suggesting a more precise distribution of priming than mere facilitation between like-structures. Assuming that both the Coordination effects and Thothathiri and Snedeker’s findings are reliable (which seems likely, given that both have been replicated across multiple experiments), the *Special Cases* stand out for having a clearly available interpretation that is not abstract syntactic comprehension priming. It could take considerable additional work to establish which interpretation is the right one. In the absence of further study addressing the *Special Cases* however, for the moment we set them aside to develop the theoretical contrast that will be critical to Chapter 6.

Adaptation (Section 4.3.2) is an instance of the much larger Implicit Learning family of theories which have frequently been brought to bear on priming. According to Adaptation, priming should be constantly on-going, with every construction, every time it is used. After all, in this theory priming is a reflex of the adjustment to the probabilistic frequency of that construction both globally and in the environment. Frequency information is not reserved for just a few, special con-

---

<sup>22</sup>Though not without considerable ink and time having been dedicated to this issue: see for instance the discussions of priming versus learning in Bock and Griffin (2000), Chang et al. (2006), and Luka and Choi (2012).

structions. It is a general property that can be associated with any unit of grammatical knowledge, regardless of its internal syntactic properties. Moreover, it changes every time a speaker uses or encounters that construction, if only by a little bit. The only reason that experiments fail to find priming, according to this theory, is that the probability shift is too small to detect using available methods or experimental designs.

For this reason, the potential priming cases reviewed in the context of Adaptation theory all focused in some way on either increasing the size of the frequency effect or reducing the unassociated noise. The former strategy is at the heart of cumulative priming. The goal of cumulative priming is to dramatically increase frequency in the local context, and let the decrease in Surprisal facilitate reading times. Looked at a different way, it increases the effect size by stacking priming from not just one but many primes over the course of an experiment. The other strategies in this family —what this chapter called Item-bias and Mere Exposure —work, according to Adaptation theory, by reducing the surrounding, noisy changes in frequency of unrelated material to measure more detailed shifts in Surprisal than would otherwise be detectable.

Lexically-Filtered Comprehension theory (Section 4.3.3) is quite different. It suggests that in any case where a priming effect can be assigned to the connection between a lexical item and its associated possible grammatical rules, it will. Abstract priming becomes the option of last resort. In comprehension, where much of lexical processing necessarily precedes syntactic analysis, this will leave abstract syntactic priming apparently non-existent. The only cases where abstract priming could surface are those in which a grammatical option cannot be associated with any individual lexical items. This predicts that any construction that relies on verbal subcategorization, (such as double object alternations), will be a subpar place to look for priming in comprehension. By contrast, grammatical options that are generally available for all members of a syntactic category or phrase, such as nominal modification with relative clauses or prepositional phrases, are much more likely candidates. Lexically-Filtered Comprehension thus allows abstract priming, but with a dramatically limited distribution compared to other theories.

While the environments that might permit comprehension priming under Lexically-Filtered Comprehension are limited, they can be somewhat diverse. Attachment priming, for instance, is predicted to allow priming because any noun has the (roughly equal) ability to take phrasal modification and therefore ambiguity arises when multiple host NPs are present. Although some nouns might take modification more frequently than others (e.g., *the idea that...*) most nominals will differ so little in their tendency toward modification that the parser typically cannot make a

substantial guess about attachment based on the identity of the lexical noun in isolation.<sup>23</sup> Attachment decisions like this can be globally ambiguous and relatively free of lexical bias across a whole sentence. The Verb-Final case study that was considered last addresses a separate option: that even if lexical information *could be* available to guide parsing decisions globally, when incremental parsing must make a decision before strong lexical information is available, then abstract priming may be able to influence online parsing. The delay of strong lexical information like subcategorization will be common enough in verb-final languages. However, detecting abstract priming is hampered by nominal case, which most SOV languages use as a verb-independent strategy to resolve much of the argument structure ambiguity well before the verb appears. That said, it is likely that more environments which allow priming due to a delay in lexical information will be found as psycholinguistics broadens the range of languages that it considers.

The cases of comprehensions priming reviewed in this chapter are not sufficient to rule out either Adaptation or Lexically-Filtered Comprehension. Both theories appeared to have evidence which supported their specific claims and was less suited to other, but definitive evidence is lacking. Because Adaptation predicts priming for all syntactic structures and Lexically-Filtered Comprehension for only a few, the critical test cases will be those that Adaptation predicts should show priming but Lexically-Filtered Comprehension does not. The next chapter intends to provide such a test, with a specific focus on whether previous evidence taken in support of Adaptation is reliable.

---

<sup>23</sup>Making an incremental decision about attachment based on the noun in the 'semantic' context of the whole sentence is a separate case.

## CHAPTER 5

### AN EMPIRICAL TEST OF ADAPTATION AS SYNTACTIC PRIMING

#### 5.1 Introduction

Chapter 4 set up the contrast between two potential models of priming in comprehension: rapid, statistically-sensitive Adaptation on the one hand and Lexically-Filtered Comprehension on the other. Because the instances when Lexically-Filtered Comprehension predicts that comprehension priming should obtain are largely a subset of the instances when Adaptation predicts it should, Chapter 4 concluded that the best way to decide between the two theories was to test the predictions of Adaptation that do not overlap with Lexically-Filtered Comprehension. This chapter provides a multi-layered empirical test of Adaptation in syntactic contexts that would not be predicted to allow priming under Lexically-Filtered Comprehension. It starts with an analysis of pre-existing data from Staub, Dillon, and Clifton (2017) and then goes on to present two novel experiments that further test for Adaptation.

#### 5.2 Order Analysis of Staub, Dillon, & Clifton (2017)

If comprehension priming in the form of Adaptation exists, it should be detectable in any experiment with enough tokens, even if the experiment was not originally designed to detect order effects. The data in this first section therefore comes from several previous eyetracking experiments on ORCs vs. SRCs, and will be used to look for evidence of syntactic Adaptation effects. The starting point is the data from Staub et al. (2017), which used eyetracking-while-reading to localize the ORC penalty within the relative clause. If Adaptation exists, it should appear in the Staub et al. data as a facilitatory interaction of Order and Sentence Type for tokens late in the experiment, just as it does in Fine et al. (2013).

##### 5.2.1 Materials and Procedure

The results of this re-analysis are based on the data from Staub et al. (2017) and were originally presented in Andrews, Staub, and Dillon (2017). Two eyetracking-while-reading experiments used stimuli based on the sentences in (59).

- (59) a. ORC: The chef [ that/ the waiter/ distracted/ \_\_\_ ] poured/ the flour onto the counter.  
b. SRC: The chef [ that \_\_\_ / distracted/ the waiter/ ] poured/ the flour onto the counter.

The first experiment included 74 subjects, nine of whom were rejected due to excessive incidental data loss, leaving 65 subjects for analysis. Experiment two initially ran 60 subjects and excluded 6 for data loss. Because both experiments used the same stimuli, the results were directly pooled, for a combined data set of 119 subjects. This gave the analysis a particularly high power, which functions as a partial guard against the possibility that any lack of priming in comprehension is merely due to effects that are too small to be reliable with typical sample sizes.

### 5.2.2 Overview of Staub et al.'s Original Results

Staub et al.'s original analysis looked at three regions of interest: the Relative NP, the Relative Verb and the Matrix Verb. Of these, the Relative NP showed the largest effects and tracked most closely with the ORC/SRC manipulation, as predicted by expectation-based accounts (Hale, 2001; Levy, 2008).<sup>1</sup> In this region, the ORC penalty was realized as longer reading times in both First Pass and Go Past measures. In addition, the effects at the Relative Verb were very small compared to the effects at the Relative NP, but were still significant in Go Past times. Accordingly, the Order analysis will also be focused primarily on the Relative NP and Verb in First Pass and Go Past.

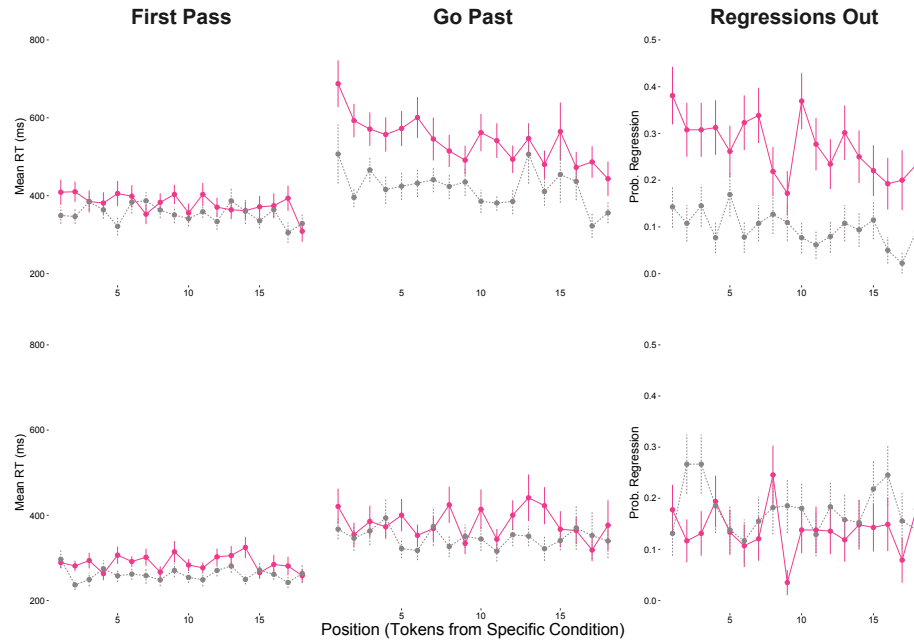
### 5.2.3 Results of the Staub et al. Order Analysis

Summary plots of the Order analysis are given in Figure (5.1). All analyses consisted of linear mixed effects models with the factors SENTENCETYPE and ORDER and a full random effects structure. SENTENCETYPE was coded as ORC (.5) and SRC (-.5). ORDER was defined as the number of tokens of that Relative Clause type that a participant had already seen.<sup>2</sup> All final models used centered predictors to avoid collinearity. The results reported below are based on the combined data for Staub et al.'s Experiments 1 & 2, however, the individual data for each experiment had very similar patterns (substantial divergences between the combined data and individual experiments are noted when appropriate). Following Gelman and Hill (2007), significance is assessed at the  $t > 2$  level.

---

<sup>1</sup>This was, in fact, one reason to look at ORCs as a target construction, in addition to Wells et al. (2009)'s prior study. Because other parts of the processing comport with the predictions of expectation-based theories, the ORC is an especially attractive construction to use in a test of expectation-based Adaptation.

<sup>2</sup>Two other definitions of ORDER were also considered: 1) the number of Relative Clauses of *either* type that a participant had already seen or 2) the number of total sentences a participant had seen (i.e., absolute position in the experiment). Both of these definitions of ORDER produced plots which were similar but had visibly more noise than the definition of ORDER that was ultimately adopted.



**Figure 5.1.** Reading Times and Probabilities of Regression for each position in the order analysis. Top: Relative NP; Bottom: Relative Verb

At the critical Relative NP region there was a main effect of `SENTENCETYPE` which was significant in both First Pass and Go Past, reflecting the expected penalty for ORCs (First Pass:  $\beta=18.66$ ,  $SE=9.14$ ,  $t=2.04$ ; Go Past:  $\beta=117.35$ ,  $SE=13.53$ ,  $t=8.67$ ). In addition, there was a significant effect of `ORDER` in Go Past times ( $\beta=-5.19$ ,  $SE=1.73$ ,  $t=-3.00$ ), indicative of reduced reading times for timesteps later in the experiment. However, the crucial interaction of `SENTENCETYPE` x `ORDER` was not significant in either measure (First Pass:  $\beta=-0.51$ ,  $SE=1.26$ ,  $t=-0.41$ ; Go Past:  $\beta=-3.09$ ,  $SE=2.59$ ,  $t=-1.19$ ).

At the Relative Verb region, the main effect of `SENTENCETYPE` was significant as early as First Fixation ( $\beta=29.28$ ,  $SE=3.55$ ,  $t=8.25$ )<sup>3</sup> and held through First Pass ( $\beta=26.38$ ,  $SE=6.11$ ,  $t=4.32$ )<sup>4</sup> and Go Past ( $\beta=35.99$ ,  $SE=15.59$ ,  $t=2.31$ ). The main effect of `ORDER` was not significant at First Pass ( $\beta=0.11$ ,  $SE=0.53$ ,  $t=0.21$ ) or Go Past times ( $\beta=-1.33$ ,  $SE=1.14$ ,  $t=-1.16$ ). Notably, the direction of the effect is not even in the expected direction in First Pass. Moreover, the crucial interaction never attained significance (First Pass:  $\beta=0.53$ ,  $SE=1.02$ ,  $t=0.52$ ; Go Past:  $\beta=0.63$ ,  $SE=2.18$ ,  $t=0.29$ ).

<sup>3</sup>This was the only significant effect in First Fixation throughout the analysis of the combined data.

<sup>4</sup>This effect was not significant in Experiment 2 when experiments were analyzed individually ( $\beta=17.32$ ,  $SE=11.19$ ,  $t=1.55$ ).

The Order-analysis results support Staub et al.'s conclusion that the sources of difficulty at the relative NP and relative verb are both qualitatively and quantitatively different: the difficulty at the relative verb showed no detectable sensitivity to Order, while the Order effects were in evidence at the NP (since the ORDER factor was centered, the  $\beta$  from the combined analysis can be interpreted as  $\sim 5$ ms of facilitation per new exposure at the mid-point of the experiment). The presence of a facilitation effect for Order at the NP does argue that experience with relative clauses matters, but there was no evidence for a difference in the magnitude of the Order effect between the ORC and SRC conditions, which was the crucial prediction of the Fine et al. Adaptation account from Chapter 4. Given the difference in methodology, this would preliminarily suggest that Fine et al. could have been observing some type of change other than Syntactic Adaptation, such as task adaptation.

#### 5.2.4 Discussion of Staub et al. Order Analysis

The results of the re-analysis indicate no discernible Syntactic Adaptation effect (that is, no interaction with syntactic structure), although there was a clear effect of adaptation to the experiment in the form of the main effect of ORDER. This seems to be in contrast with the findings of Fine et al. (2013) and particularly Wells et al. (2009). Recall that the findings from Wells et al. were especially compelling because there appeared to be a differential effect just for ORCs (see the reproduction of Wells et al.'s results in Figure [5.2]). The critical feature was that in the post-test, the reading times (RTs) for ORCs were substantially facilitated in the Experience Group over the RTs in the Control Group, but the effect of training for SRCs was much smaller.

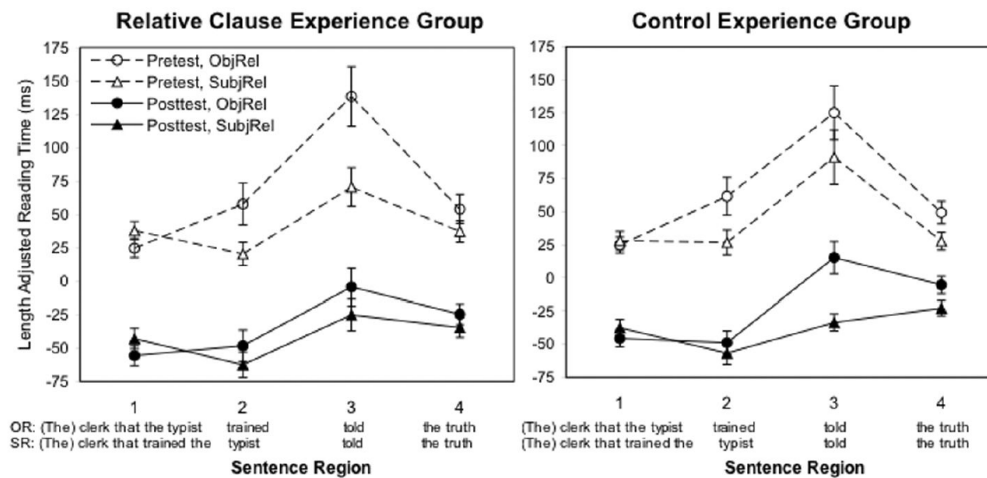


Figure 5.2. Reproduced from Wells et al. (2009). Pre- and post-test self-paced RTs for the relative clause conditions.

On the basis of two new eyetracking experiments, the remainder of this chapter argues that what presents as adaptation for a particular syntactic construction is instead a form of adaptation to the task or the context. Specifically, what appears to be Syntactic Adaptation does not actually follow the patterns predicted by a statistically-sensitive Adaptation account in that it is not well captured by exposure to any individual construction. Rather, the adaptation seen in the following two experiments, as well as Fine et al. and Wells et al., is better characterized as adaptation to the difficulty of an experiment as a whole, and based on the participant's experience with the experiment, not the particular construction.

Because the concept of task adaptation will be integral to the ultimate conclusions of the chapter, before moving on it is worth taking a moment to clarify what we mean by this term. The broadest relevant definition of task adaptation is: any adjustment in a participant's behavior due to the experimental context, rather than a reaction to change in the stimuli (though critically, it can be superficially sensitive to the difference between conditions, as we shall see). Task adaptation is frequently considered to be a large sub-type of what is sometimes known in the behavioral experiment literature as experimental artifacts, or changes in behavior that are caused by the experiment but which are auxiliary to the phenomenon under study (Orne, 1962; Rosenthal & Rosnow, 1969/2009), but in some cases task adaptation can be the effect of interest. This characterization of task adaptation is intentionally inclusive, highlighting how ubiquitous it is. For instance, many of the factors which affect the well-known Speed-Accuracy Tradeoff will fall in this category (for a review, see Heitz, 2014). Illustrations of this type of research within psycholinguistics include a demonstration by Swets, Desmet, Clifton, and Ferreira (2008) that comprehenders adjust the depth of attention to reading based on the difficulty of the comprehension questions used as a control task. Potts, Keenan, and Golding (1988) found similar depth of attention effects when comparing naming and lexical decision tasks. A related literature to the SAT has produced considerable research on the impact of incentives on decision making, for instance the consequences of differing sizes of rewards for correct responses to questions versus penalties for incorrect responses (e.g., Chen & Kwak, 2017). An entirely separate—but highly psycholinguistically relevant—example comes from Adams et al. (1998), who demonstrated that participants in self-paced-reading adjust the prosodic phrasing that they assign to a sentence to preferentially make prosodic boundaries align with the presentation “chunks”. Because the Adams et al. effect is a change in how comprehenders parse a sentence based on mode of presentation, this too is considered task adaptation, despite the fact that it is arguably quite different from adjusting to the difficulty of comprehension questions or the SAT. Furthermore, these examples are far from exhaustive. Their point is that

task adaptation effects are varied, pervasive, and can influence outcomes even in very carefully constructed research.

Moreover, the reason to ascribe change over an experiment to task adaptation as a whole, rather than to a specific source, is to suggest that multiple processes may have an effect both within and across experiments, and indeed, across participants within a single experiment. There is no reason to think that both the Swets et al. and Adams et al. types of adaptation could not exist side-by-side in the right methodological setting (in this case, a phrase-by-phrase self-paced reading task with comprehension questions) and indeed be compatible with still more concurrent task adaptation strategies. Relevant to the current questions about syntactic priming, both of these types of adaptation could easily manifest as a facilitatory Order effect if the comprehension questions are sufficiently simple and the prosodic phrasing is predictable enough. Moreover, some participants may be more sensitive to prosodic phrasing than question difficulty and vice versa, leading to within-subject variation in the impact of each type of task adaptation. Thus, it is not only possible but in fact likely that any task adaptation in these experiments is an amalgam of several different strategies in effect at the same time.

### 5.3 Experiment 3

The results of Wells et al. (2009), Fine et al. (2013), and the order analysis of Staub et al. (2017) in the preceding sections suggest that cumulative priming effects *may* exist in comprehension, but all three suffer from limitations that make interpretation difficult. For Fine et al., interpretation is hampered by the fact that self-paced reading confounds task adaptation with Syntactic Adaptation and that the target construction is so difficult to comprehend that it may not be representative of the learning strategies at work in normal language processing. Wells et al. has a more natural target structure, but only evaluates pre- and post-test values, and is still subject to the same concerns about self-paced reading. Critically, the concerns under discussion about self-paced reading are specific to the particular debate about adaptation. While self-paced reading results frequently agree with findings in eyetracking, the argument here is that task adaptation is a special concern when the effect of interest is also a form of adaptation. Task adaptation over the course of an experiment could easily mimic Syntactic Adaptation without the proper control conditions to tell them apart.

A distinct set of concerns applies to the Order analysis of Staub et al. (2017). The experiments in Staub et al. (2017) were not designed to test for priming and thus employ equal rates of pre-

sensation of ORCs and SRCs, which may not have been dramatic enough to evoke a clear reversal of SRC RTs as predicted by expectation-based theories. The amount of expectation shift for a construction over the course of an experiment is determined by the degree of difference between the pre-experimental probability of seeing a continuation and its probability within the experiment. Because the scale that determines the amount of expectation shift is Surprisal rather than a linear scale, every additional unit of change from the prior to the experimental context can have a substantial impact on the overall size of the effect.

Moreover, while eyetracking-while-reading is a much more practiced task than self-paced reading, and therefore less prone to changes in task-strategy over the course of the experiment, it is still possible/likely that there is adjustment to the particular sentence types in the experiment and an overall effect of experiment fatigue/adaptation that may affect sentence types differently based on their difficulty. This means that the results of the Order analysis of Staub et al. (2017) are neither a clear argument against the expectation-based theories of Fine et al. (2013) nor a clear indicator of the presence of the monotonically-increasing facilitation for both SRCs and ORCs predicted by activation-based theories. Finally, all three studies also lacked a lexically-matched baseline that was not predicted to compete for expectation with the target construction, making statistical comparison and interpretation more difficult.

Experiments 3 & 4 are a set of paired experiments designed to counter the limitations that hamper the interpretation of previous studies. Together, the experiments assess the adaptation effect for two target structures. The first of these is the ORC/SRC comparison of Wells et al. and the Order analysis of Staub et al. (2017), which is contrasted with adaptation to the NP/Z→Z construction, a garden path of comparable difficulty to the MainVerb/RelativeClause used in Fine et al. (2013), Tooley and Traxler (2018), and others.<sup>5</sup> The two experiments manipulate these contrasts to simultaneously test two set of predictions from the Adaptation account.

The first prediction being tested is the direct effect of frequency, applied to a given syntactic structure. Following the logic of Adaptation, a single structure should be facilitated when it is more expected in context relative to when it is contextually rare. To that end, there was a between-experiments manipulation of FREQUENCY for the two target structures. The ORC target construction was manipulated to be frequent in Experiment 3 and rare in Experiment 4; conversely, the NP/Z → Z garden path was rare in Experiment 3 and frequent in Experiment 4. The frequent

---

<sup>5</sup>It was not possible to use the MV/RR contrast in the same experiment as ORCs because MV/RR also involves a relative clause. The two structures would have impacted the expectation allocated to each other.

construction was seen 32 times in its experimental context, while the rare construction was seen 8 times. Not only did this mean that the structure was four times more numerous when it was frequent than when it was rare, but additionally the experiments were constructed so that the frequent construction was ~25% of all sentences in that experiment. This ratio of target to non-target structures was chosen to match the 1 in 4 rate of presentation that has been shown to lead to increased learning in memory studies (Glenberg, 1976).<sup>6</sup>

	Experiment 3	Experiment 4
ORC	32	8
NP/Z → Z	8	32

**Table 5.1.** Frequency of the two target structures across Experiments 3 and 4

The second set of predictions involves the comparison of the same two target structures to closely related baselines. Each target was accompanied by two baseline structures. For ease of exposition, we will use the ORC paradigm to introduce the two control types, but we will return to the NP/Z paradigm below. One of the two baselines was a Competing control, which, as the name suggests, was in competition with the target for expectation in an Adaptation framework. The Competing control for an ORC was a lexically-matched SRC, because at the complementizer *that* in (60a) and (60b) an adapting comprehender will have to choose how to allocate expectation across the possible upcoming input: a DP favors an ORC continuation while a verb favors an SRC.

- (60) a. Target: ORC - The marine biologist *that*/ the botanist/ consulted/ presented/ a paper at the conference last week.
- b. Competing Control: SRC - The marine biologist *that*/ consulted/ the botanist/ presented/ a paper at the conference last week.
- c. Non-Competing Control: Complement - The marine biologist believed *that*/ the botanist/ consulted/ a statistician at the conference last week.

The second control is a Non-Competing control. The Non-Competing control is lexically-matched to the target and Competing control at the critical region but has something syntactic that rules out a continuation with the target or the competing control early in the structure. For

---

<sup>6</sup>If priming is an implicit learning mechanism as Fine et al. suggest, then it may well follow the same learning rates as those in other learning tasks. While no particular psycholinguistic point in the present experiments rests on priming have the same learning characteristics as other memory tasks, matching the presentation rates to those in general memory studies certainly does no harm. If learning and priming characteristics are the same, it may increase the overall effect size that the study aims to detect.

the ORC paradigm, this extra property is the addition of the main verb *believed*. While (60c) still contains an embedded clause like (60a) and (60b) do, *believed* forces the embedded clauses to be a complement within the matrix verb phrase, rather than a relative clause modifier on the subject.<sup>7</sup>

The Competing and Non-Competing controls work in tandem to make Syntactic Adaptation as clearly identifiable as possible. Per the logic in Fine et al. (2013), as the target structure becomes more and more expected with exposure, less expectation should be allocated to the Competing structure. All other things being equal, this would predict longer RTs for the Competing control as the experiment goes on. To increase the effect size associated with this prediction, in hopes of making it more detectable, the design employed a skewed presentation rate: the controls were always presented at the same rate as the rare targets (8 tokens per experiment), making them one quarter as frequent in the experiment as their corresponding target.

Even with the skewed presentation, it is still possible that the upward pressure on the RTs of the Competing control is not enough to actually make the Competing structure harder to read at the end of the experiment than the beginning. Net facilitation over the course of the experiment could result if there is enough general task adaptation to outweigh the decreased expectation. Thus, there is the risk that the effect of decreased expectation for the Competing control would exist as a true force acting on the RTs, but would be undetectable due to being swamped by unrelated adaptation. The Non-Competing baseline serves as a stop-gap guard against this possibility. The Non-Competing control is subject to all the same facilitatory pressures as the Competing condition, but does not have pressure for longer RTs from expectation competition. Therefore, even if the effect of competition on the Competing condition does not result in those sentences becoming harder to read at the end of the experiment, the competition effect would still be detectable as an interaction of the two baseline conditions.

In fact, the Non-Competing condition serves as an independent measure of the general, additive task adaptation over the course of the experiment. Any generally applicable task/experiment adaptation will be applied toward the difference between the RTs of the first and last token a participant sees from this condition. And because participants see so few tokens of the Non-Competing condition, the Syntactic Adaptation/priming applied to it will be small.

An alternative, but also fruitful way to view the two control conditions is as the appropriate baselines for two different theories of priming/implicit learning. This feature also counters a

---

<sup>7</sup>It is technically possible to get a reduced relative clause reading from *The marine biologist believed* but the likelihood of this reading becomes vanishingly small with the addition of the complementizer *that*, rather than, say, *to* or *by*, such that by the time the reader actually encounters *the botanist* the possibility of a relative clause reading should be negligible.

problem otherwise introduced by the skewed presentation rates. The hallmark of statistically sensitive, log-based Adaptation was supposed to be an interaction, but with skewed presentation rates even an additive activation account would predict an interaction between the priming rates of ORCs relative to SRCs. But additive activation priming would not predict the interaction between the two controls any more than task adaptation would. Given evidence of true priming from the between-subjects frequency manipulation, the relationship of the two baselines determines whether a competition or additive activation account is a better fit. An interaction between the two controls supports competition-based theories. A main effect of ORDER with no interaction is more consistent with older activation-style theories.

The NP/Z paradigm has a comparable set of Competing and Non-Competing controls. Though they are derived from the target structure in a different way than in the ORC paradigm, the two controls serve the same overall function. The NP/Z  $\rightarrow$  NP control (which will sometimes be referred to simply as the NP condition) is the alternative and more accessible reading of the garden path sentence. NP/Z garden paths hinge on the ability of the verb in the subordinate clause to take either an NP object or no object at all (Zero object). This leads to an ambiguity about whether *the deer* should be parsed as the object of *sketched* (the NP reading) or the subject of the matrix clause. The ambiguity is only resolved by the appearance of either the matrix verb or its true subject, depending on the condition. Previous studies have shown that comprehenders may fail entirely to resolve a coherent parse of the Z-resolved condition (F. Ferreira & Henderson, 1991). The longer that the processor entertains the NP parse, the stronger the garden path effect is (“digging-in” effects, Frazier and Clifton (1998); Frazier and Rayner (1982); Tabor and Hutchins (2004)), thus the ambiguous NP always had a three-word PP modifier. The Non-Competing structure simply uses a comma to force an intransitive reading of the subordinate verb, thereby avoiding any ambiguity about the attachment of the NP in the first place (this will sometimes be referred to as the Comma condition).

- (61) a. Target: NP/Z  $\rightarrow$  Z - While the artist sketched the deer in the field munched late summer grass peacefully.
- b. Competing: NP/Z  $\rightarrow$  NP - While the artist sketched the deer in the field the herd munched late summer grass peacefully.
- c. Non-Competing: NP/Z Comma - While the artist sketched, the deer in the field munched late summer grass peacefully.

Having two full target paradigms with accompanying controls has a number of advantages. First there is the obvious advantage of automatic replication and extension to different constructions for any effect that the experiment finds. Second, it allows these experiments to compare across constructions with different starting difficulties (and therefore different starting expectation from a Syntactic Adaptation point of view). ORCs are rarer than SRCs, but not so rare or difficult that comprehenders fail to reach a fully specified parse. NP/Zs can and do lead to catastrophic parsing failure (F. Ferreira & Henderson, 1991), which according to Adaptation theory should be precisely because they are so rare and unexpected. This means that ORCs start with less possible facilitation to gain and should also achieve facilitation slower. And because expectation exists on the log-based scale of Surprisal, the difference between Adaptation for the target constructions could be considerable. The predicted adaptation effect size based on the rarity and difficulty of the hard garden path reading is one of the primary reasons that more recent adaptation researchers have tended toward the comparably difficult MainVerb/RelativeClause garden paths as their structure of choice. But the difficulty of these hard garden paths is also a potential weakness. The fact that comprehenders do not always succeed at parsing either MV/RR or NP/Z garden paths potentially makes them quite different from normal parsing, where difficulty rarely reaches the level of conscious awareness. It may be that learning applied to these structures is learning to find a grammatical parse at all, and that might make them not quite the right structure for understanding implicit learning in more natural comprehension. In contrast, learning for ORCs would be more akin to ‘practice’ than acquiring a brand new skill. This makes the comparison between ORCs and NP/Z→Z garden paths an informative lens through which to view the prior literature.

As a minor extension of the comparison between easier and more difficult constructions, the design also included a set of 8 NP/S garden paths (see [62]). NP/Z garden paths are much more difficult than all the other sentences in the experiment, and there was a concern that other kinds of adaptation, not just Syntactic Adaptation, might apply differently to syntactic structures across a spectrum of difficulty. NP/S garden paths were included as a control to help diagnose this possibility because they are slightly easier on average than NP/Z garden paths, but still reliably produce considerable parsing difficulty (Frazier & Rayner, 1982).

- (62) NP/S: The anthropologist remembered the pottery style was characteristic of an obscure Amazonian tribe deep in the rainforest.

The RC and NP/Z paradigms, together with the NP/S garden paths and 32 unrelated fillers formed the bulk of the sentences in the eyetracking portion of Experiments 3 and 4. However, not all of these sentences appeared in both experiments. Because the goal was to keep the frequent target in each experiment as close as possible to one fourth of the total sentences, and because the two controls were of most use when their corresponding target was frequent, the controls were sometimes dropped. In Experiment 3, neither of the NP/Z controls conditions were included. In Experiment 4 it was decided that dropping both controls was too extreme, and therefore the complement control was removed, while the SRC baseline remained.

Table (5.2) provides an overview of the changes between Experiment 3 and 4, along with the overall frequencies of each structure in each experiment.

	Experiment 3	Experiment 4
ORC	.25 (32)	.058 (8)
SRC	.0625 (8)	.058 (8)
Complement	.0625 (8)	0 (0)
NP/Z → Z	.0625 (8)	.235 (32)
NP/Z → NP	0 (0)	.058 (8)
NP/Z → Z comma	0 (0)	.058 (8)
NP/S	.0625 (8)	.058 (8)
PO Datives	.25 (32)	.235 (32)
Fillers	.25 (32)	.235 (32)
<b>Total:</b>	128	136

**Table 5.2.** Relative Frequencies of sentence types the eyetracking portions in Experiment 3 and Experiment 4. Absolute quantities are given in parentheses

### *Sentence Completion Task*

Lastly, even if no priming effect obtained in eyetracking once experimental fatigue is accounted for, it would still be possible that something about the experiment prevented a priming effect for reasons unrelated to the hypotheses. This concern is especially acute because the priming literature has tended to focus on a few select constructions which do not include ORCs or NP/Z garden paths. Even if the experiment were in principle capable of producing priming, there might be some quirk that stopped it from applying to ORC and NP/Z target structures. In order to allay these concerns, participants performed a sentence completion task before and after the eyetracking portion of the experiment.<sup>8</sup> Sentence completion tasks such as this are a type of production

<sup>8</sup>The sentence completion task was intentionally chosen to be quite close to the cloze task, which has frequently been used to establish relative Surprisal of potential continuations. In a cloze task, participants are given the beginning fragment of a sentence and then asked to provide a continuation, but classic cloze tasks ask only for the first word following the

task and have previously been shown to be sensitive to abstract priming on a trial-to-trial basis (Branigan et al., 1999).<sup>9</sup> Therefore the sentence completion task can be used to assess whether exposure during eyetracking was able to produce *any* priming, independent of the question of priming in comprehension.

To address the worry about ORCs and NP/Zs specifically being primeable, all of the target stimuli in the eyetracking experiment were matched by the same number of prepositional dative sentences and the target sentence completion fragments were likewise matched by an equal number of dative sentence completion fragments. An example of the prepositional dative sentences used in the eyetracking experiment is given in (63) and the corresponding sentence completion prompt is given in (64a). Examples of the RC and NP/Z sentence completion fragments are given in (64b) and (64c) respectively.

(63) The real estate mogul left a lot of money to his children after he died.

(64) a. Dative: The reclusive novelist gave \_\_\_\_\_

b. RC Target (Experiment 3 Only): The brilliant inventor that \_\_\_\_\_

c. Target (Experiment 4 Only): While the motorcycle rider parked \_\_\_\_\_

Participants saw 14 target fragments and 14 dative fragments before the eyetracking portion of the experiment and another 14 of each afterward. Each participant saw one of four lists that divided the fragments between the pre- and post- test blocks. The lists were subject to the requirement that each fragment had to appear in each block for at least one list. Within each block the order of presentation of the fragments was randomized. Although the lexical boost effect is usually quite short lived (Tooley & Traxler, 2010, *inter alia*), as an additional precaution the sentence completion task fragments did not have any lexical overlap with either the eyetracking sentences or other sentence completion task fragments. In order to ensure that there were enough verbs to fulfill this requirement, some benefactive verbs were used alongside dative verbs. Critically, the goal was not just to elicit ORCs—which is more easily done with slightly different prompts (Gennari, Mirković, & MacDonald, 2012; MacDonald & Montag, 2009)—but to elicit ORCs under conditions that were most like our eyetracking sentences. Likewise for the NP/Z garden paths.

---

fragment (Taylor, 1953). This experiment used a slightly modified version of the cloze task which asked for participants to complete the entire sentence instead.

<sup>9</sup>Trial-to-trial priming is generally considered the strongest type of priming because a single prime trial is sufficient. Recall that it still counts as trial-to-trial even if the prime and target are separated by several intervening filler trials.

### 5.3.1 Participants

Seventy-two participants of American English participated in Experiment 3 in exchange for course credit or \$10. Two subjects were removed for having <80% accuracy on the comprehension questions. We also rejected trials for which there was a blink on track loss on the critical region (the critical NP). We would have rejected any subject who lost more than a quarter of their trials on the RC stimuli in this manner, but there were none.

### 5.3.2 Procedure

The experiment took approximately one hour. Participants first completed the pre-test section of the sentence completion task. They were shown a practice item and told that they should complete the task in natural English without prescriptive rules ('rules that are taught in school that no one actually follows in everyday speech'). They were told that they should worry less about whether their sentences were entirely plausible<sup>10</sup> and that what really mattered was that the final string was a complete English sentence when read from beginning to end.

After completing the sentence completion pre-test, participants took the eyetracking portion of the experiment on an SR Research Eyelink 2000 with head stabilization.<sup>11</sup>

All participants began the sentence completion task post-test within one minute of finishing the eyetracking portion of the experiment and the experimenter tried to minimize the amount of talking. The one-minute cutoff was pre-determined as an exclusion criterion, although no participants were actually excluded based on this criterion because no participants exceeded the one-minute mark.

As an informal debrief after completing the post-test, participants were typically asked if anything stood out to them or if they noticed any sentences in particular. For those that chose to answer, the most common response seemed to be about the NP/Z which they described as "difficult to understand", "missing a word" or even "missing a comma", although relatively few were able to remember anything about an individual item. When asked to estimate how many of these they had seen, the most common answer was "3-5", consistent with the NP/Z group being problematic rather than the ORCs.

---

<sup>10</sup>For instance, the practice fragment which was about computer programmers searching for \_\_\_\_\_. The instructions gave the example that computer programmers could be searching for tortoises.

<sup>11</sup>The first thirty-nine participants were not encouraged to take a break and rarely did. However, all following participants were made to take a break including a short walk at approximately the half way point, which the experimenter also used as an opportunity to recalibrate. Analyses checked to make sure that there was no difference between these populations.

### 5.3.3 Coding Sentence Completions

Sentence completions of dative preambles were scored with two continuations of interest. The first set of continuations of interest were true prepositional object datives, which were judged to also have a corresponding grammatical dative object form with the relevant verb. These are the most obvious continuations for PO datives to prime. The second continuation of interest had the form Direct Object NP + Prepositional Phrase but where the prepositional phrase could not grammatically participate in a DO form. Bock and Loebell (1990) demonstrated that PO datives could prime non-alternating VP-attached prepositional phrases such as locatives, meaning that excluding such VP-attached PPs from the analysis would underestimate the amount of priming in the experiment from the eyetracking PO sentences. While the actual PPs of interest in this category are most likely to be those attached directly inside of the VP instead of inside an object noun phrase, this category included any continuation which contained a full post-verbal prepositional phrase. This choice was made for two reasons: (i) to avoid needing to judge whether a phrase was attached inside the VP or an NP when it was genuinely ambiguous, which could introduce bias into the coding and (ii) we know of no reason that the rate of NP-attached prepositional phrases would change between the pre- and post-test other than exposure to the PO datives in the eyetracking portion.

Critically, merely having a post-verbal preposition was insufficient. Continuations were coded separately and removed from the analysis if they were judged to use a particle verb, which indicates that the preposition forms a separate syntactic and semantic unit with the verb. The critical feature of particle verbs is that the preposition in question can appear either next to the verb or after the direct object. An example of this type of continuation from the data in response to the fragment *The astronaut brought* would be *back some samples for testing* (where the alternate form of the particle verb would have been *some samples back for testing*). Particle verbs are assumed to have their own internal subcategorization biases separate from the simple form of the verb. Moreover, to our knowledge there is no evidence that particle verbs participate in priming with PO datives, nor is it clear whether such priming would be predicted under most theoretical syntactic accounts.

The remaining dative completions were either coded as DO if they consisted of an NP-NP continuation (with any number of following adjuncts) or NotDative if the continuation took any other grammatical form. Sentences which were judged ungrammatical were removed from analysis.

For the RC fragments, there were only three types of continuations recognized by our coding schema. The first of these were ORCs and SRCs. A sentence completion was coded as an ORC if the complementizer *that* was followed by an NP-subject phrase, the rest of a relative clause and

then a matrix VP. An SRC was any continuation in which *that* was followed by a relative verb, the rest of the relative clause, and then a matrix VP. These were the two expected continuations. Unexpectedly, it was also very common for participants to continue the RC fragments with a single verb phrase that would have been grammatical without the complementizer *that*. For instance, participants completed the RC fragment *The scuba diver that* with *never used an oxygen tank* or *loved the coral reefs*. This continuation was so common that we coded it separately from other types of ungrammatical continuations and included it in the analysis, as a *Main Verb* continuation.

### 5.3.4 Results

For all analyses unless otherwise noted, position of a particular trial for ORDER was calculated based on the number of *dominant targets*, i.e., ORCs, that a participant had seen. For non-ORC trials, any tokens encountered before the first ORC were coded as position 0, trials between the first and second ORC were labeled position 1, between the second and third ORC tokens was position 2, and so on up to position 32 for tokens which appeared after all ORCs.<sup>12</sup> The ORDER factor was centered for all analyses to ameliorate multicollinearity in regression models (Iacobucci, Schneider, Popovich, & Bakamitsos, 2016).

The coding used for all factors for the within-subjects analysis of Experiment 3 is given in Table (5.3).

ORC	0.5
SRC	-0.5
Complement	-0.5
Order	continuous

**Table 5.3.** Coding used in models for Experiment 3.

Reading times (RTs) were analyzed using linear mixed effects regression with a maximal random effects structure unless otherwise stated. Significance was assessed at  $t=2.00$  (Gelman & Hill, 2007).

#### *Embedded Clause Conditions*

Inferential analyses compared ORCs to SRCs and the Complement condition in separate models. As Staub (2010) found the primary reflex of the ORC penalty at the relative NP, we focus most

<sup>12</sup>Note that this is slightly different from the Order analysis coding applied to the Staub et al. (2017) data in Section 5.2 because there the number of ORCs and SRCs was evenly balanced, so Order could be calculated entirely within a sentence type and still be on a comparable scale.

of the results and discussion there. Because previous authors have suggested that the relative verb may be a region of interest for the ORC penalty, we did run analyses at the relative verb region. These analyses supported Staub's conclusion that the primary reflex of the ORC penalty is on the critical NP rather than the relative verb. Moreover, they did not change the conclusions from the critical NP substantially. Therefore for reasons of space and time they are not reported here.

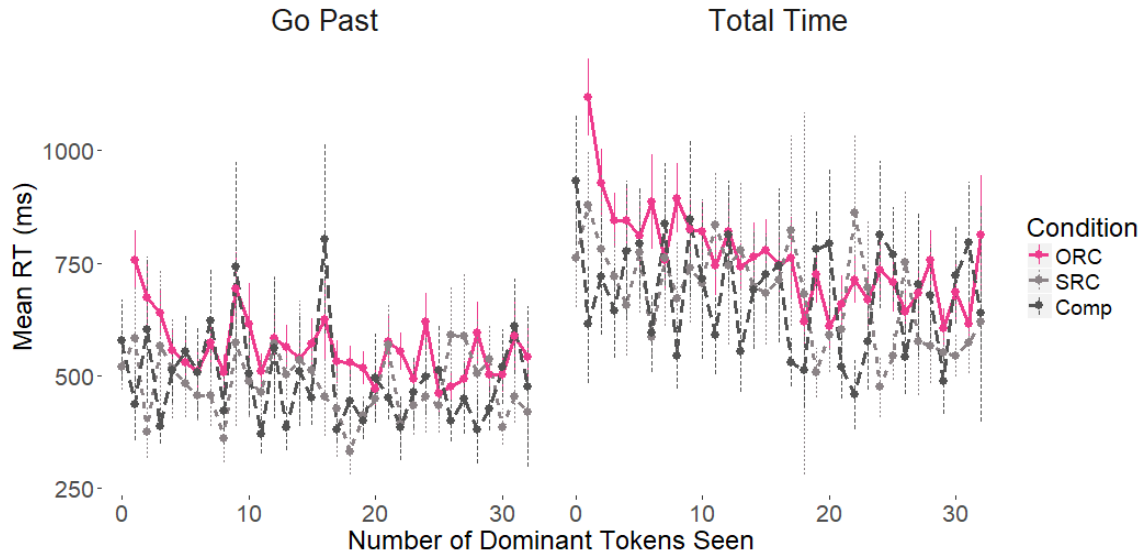
In the First Pass measure, the sole significant effect at the critical NP was an effect of ORDER in the ORC vs. SRC model ( $\beta=-1.14$ ,  $SE=0.57$ ,  $t=-2.02$ ). No other effects were significant in First Pass.

At the critical NP region in Go Past, the ORC vs. SRC comparison yielded both a main effect of ORDER ( $\beta=-2.35$ ,  $SE=1.17$ ,  $t=-2.01$ ) and ORC vs. SRC ( $\beta=68.30$ ,  $SE=21.68$ ,  $t=3.15$ ). However, the critical interaction of ORDER x SENTENCE TYPE predicted by statistically-sensitive Syntactic Adaptation accounts failed to reach significance ( $\beta=-2.65$ ,  $SE=2.00$ ,  $t=-1.33$ ). The ORC vs. Complement Clause model found a significant main effect of ORCs vs. Complement ( $\beta=137.26$ ,  $SE=41.91$ ,  $t=3.26$ ), replicating the critical ORC penalty that has been found in many studies previously with a new and different baseline construction. This model did not yield a significant effect of either ORDER ( $\beta=-1.035$ ,  $SE=1.77$ ,  $t=-0.58$ ) or interaction of ORDER x SENTENCE TYPE ( $\beta=-5.29$ ,  $SE=3.83$ ,  $t=-1.38$ ).

A logistic model comparing ORCs vs. SRCs for p(Regression) found a significant effect of RC type ( $\beta=0.728$ ,  $SE=0.127$ ,  $p < 0.0001$ ) reflecting the classic ORC penalty, but no effect of ORDER ( $\beta=0.004$ ,  $SE=0.006$ ,  $p=0.59$ ) and no interaction ( $\beta=-0.018$ ,  $SE=0.014$ ,  $p=0.18$ ). The ORC vs. Complement comparison found the same pattern. There was a reliable ORC penalty in the form of a significant ORC vs. Complement effect ( $\beta=1.456$ ,  $SE=0.255$ ,  $p < 0.0001$ ), but no ORDER effect ( $\beta=0.005$ ,  $SE=0.012$ ,  $p=0.661$ ) or interaction ( $\beta=-0.036$ ,  $SE=0.027$ ,  $p=0.183$ ).

Finally, a model of Total Times for ORCs vs. SRCs at the critical NP found a significant main effect of ORC vs. SRCs ( $\beta=69.63$ ,  $SE=21.93$ ,  $t=3.17$ ) and of ORDER ( $\beta=-7.70$ ,  $SE=1.43$ ,  $t=-5.38$ ), but did not find any interaction ( $\beta=-1.58$ ,  $SE=2.27$ ,  $t=-0.69$ ). The same held true for the ORCs vs. Complement model. There were significant effects for the ORC penalty ( $\beta=141.66$ ,  $SE=46.59$ ,  $t=3.04$ ) and ORDER ( $\beta=-6.97$ ,  $SE=2.06$ ,  $t=-3.39$ ), but not for the interaction ( $\beta=-3.04$ ,  $SE=4.87$ ,  $t=-0.624$ ).

There are a few consistent features of all of these models. First, the well-replicated ORC penalty was always significant, in all measures and against both baseline conditions (with the exception of First Pass, indicating that in this study the ORC penalty manifested primarily in regression-based measures). This is critical, because without this effect as a baseline the validity of any remaining null results would automatically be in question.



**Figure 5.3.** Change in Mean RT of the embedded clause conditions over the course of Experiment 3. Error bars represent standard error.

The pattern was less clear for main effects of *ORDER*, though in general *ORDER* effects were present in the SRC models and not in the Complement models. The exceptions were *p*(Regression) for SRCs, where *ORDER* was not significant, and Total Time for ORCs compared to the Complement conditions, where it was. Finally, the crucial interaction of the ORC penalty and *ORDER* predicted by statistically-sensitive Syntactic Adaptation never reached significance (or even marginality) regardless of measure or baseline condition.

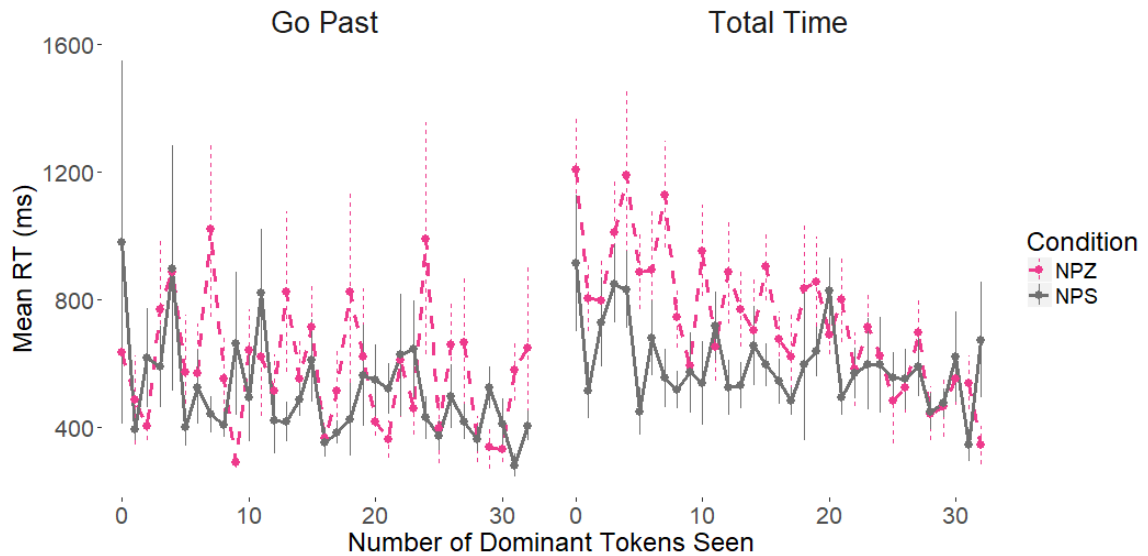
Note that a main effect of *ORDER* was *not* predicted for SRCs, because the anticipated data pattern for the ORC vs. SRC comparison was specifically a cross-over interaction.

#### *Garden Path Conditions*

In Experiment 3, neither NP/Z nor NP/S sentences had a baseline for comparison, therefore all models of the garden path conditions test only for the main effect of *ORDER*, with no interaction. All model statistics are given for the disambiguating verb region.

In Go Past, neither the NP/Z or NP/S conditions had a significant effect of *ORDER* (NP/Z:  $\beta=-2.73$ ,  $SE=2.95$ ,  $t=-0.93$ ; NP/S:  $\beta=-1.93$ ,  $SE=1.96$ ,  $t=-0.98$ ).

There was again no significant effect of *ORDER* for either the NP/Z or NP/S conditions in Probability of Regression (NP/Z:  $\beta= -0.013$ ,  $SE=0.011$ ,  $p=0.266$  ; NP/S:  $\beta=-0.014$ ,  $SE=0.014$ ,  $p=0.319$ ).



**Figure 5.4.** Change in Mean RT of garden path conditions over the course of Experiment 3. Error bars represent standard error.

Lastly, in Total Times, there was a significant effect of ORDER for NP/Z garden paths ( $\beta=-15.76$ , SE=2.22,  $t=-7.10$ ). However, the ORDER effect for NP/S garden paths remained non-significant ( $\beta=-4.44$ , SE=2.67,  $t=-1.67$ ).

These results are consistent with some adjustment in RTs/strategies over the course of the experiment, but primarily in *re-reading times*, specifically second pass. It is difficult to draw more conclusions from the garden path conditions with so little data, but these results will become central in the comparison of Experiment 3 to Experiment 4.

#### *Sentence Completion*

We now turn to the production data from the pre- and post-test sentence completion task. The counts and by-subject-adjusted percentages of each type of dative continuation are given in Table (5.4).

	Pre-test		Post-test	
	Raw Count	%	Raw Count	%
PO	202	20.30	238	23.9
VP-attached PP	73	7.34	106	10.7
Particle Verb	96	9.66	65	6.54
DO	157	15.80	225	22.6
Not Dative	473	47.60	364	36.6

**Table 5.4.** Sentence completion results for dative preambles in Experiment 3 as raw counts and by-subject percentages of tokens per block (each block included of 14 dative preambles).

Sentence completion for the dative fragments data was analyzed via two separate poisson regression models applied first to the PO-only counts and then to PO + VP-ATTACHED PP with before/after as the predictive factor. The results of these models are given in Table (5.5). The main effect of before/after eyetracking applied to PO continuations only was marginal ( $\beta=0.16$ ,  $SE=0.09$ ,  $p=0.0865$ ), while the model for the combined instances of all VP-attached PPs was significant for the same effect ( $\beta=0.22$ ,  $SE=0.08$ ,  $p < 0.01$ ). In other words, participants produced more continuations of the type which are primed by PO datives in the post-test sentence completion than they did in the pre-test. This significant effect represents a comprehension-to-production priming from the PO datives in the eyetracking portion of the experiment.

		$\beta(SE)$	$p$
<b>Dat PPs Only</b>	<i>Intercept</i>	1.13 (0.05)	<0.00001
	<i>Pre/Post</i>	0.16 (0.09)	0.0865
<b>Any VP-attached PP</b>	<i>Intercept</i>	1.42 (0.06)	<0.0001
	<i>Pre/Post</i>	0.22 (0.08)	<0.01

**Table 5.5.** Outcome of the mixed models applied to the dative sentence completion results of Experiment 3

For the RC fragments, raw counts and percentages per block are given in Table (5.6), and the results of the mixed effects models are given in Table (5.7). There was a clear and significant increase in the number of ORCs in the post-test relative to the pre-test ( $\beta=0.22$ ,  $SE=0.08$ ,  $p < 0.01$ ), although this conclusion is tempered by just how rare ORC continuations were overall.

	Pre-test		Post-test	
	Raw Count	%	Raw Count	%
ORC	4	0.47	19	2.22
SRC	729	85.40	723	84.70
Main Verb	272	15.50	262	14.30

**Table 5.6.** Sentence completion results for relative clause fragments in Experiment 3 as raw counts and by-subject percentages of tokens per block (each block included of 14 RC preambles).

		$\beta(SE)$	$p$
<i>Intercept</i>	-2.097 (0.28)	<0.0001	
<i>Pre/Post</i>	1.56 (0.55)	<0.005	

**Table 5.7.** Outcome of the mixed effects models applied to the relative clause sentence completion results.

The RC sentence completion had another function in this experiment, beyond establishing the ability to prime ORCs. The sentence completion task is a close cousin of the cloze task, which has

been used to estimate the Surprisal of potential input. While there are differing views on how closely the cloze task tracks with the Surprisal-link function applied to comprehension (c.f. Levy, 2008; N. Smith & Levy, 2011), it has been used in many previous studies and has the advantage of being specific to the precise syntactic form of the critical sentences in any given experiment. Extending the logic of Surprisal from lexical items to syntactic continuations, as Fine et al. (2013) did, suggests that the sentence completions from the pre-test can be used as a rough estimate of the pre-experiment expectation for ORCs. Calculating the aggregated “Surprisal” of ORCs from the pre-test yielded a value of 8.38 bits.<sup>13</sup> This is notable because it is actually higher than the pre-experiment Surprisal that Fine et al. (2013) calculated for the reduced relative continuation of the MainVerb/RelativeClause garden paths that they used in their experiment, which was 6.97 bits. This is important because a concern about the use of ORCs is that they might have too much expectation at the beginning of the experiment for Adaptation to be visible.

Finally, the prevalence of main verb continuations was unexpected as this led to an unambiguously ungrammatical final sentence. There were, in fact, 10 participants who never completed an RC fragment with anything other than a main clause continuation in either block, suggesting that these participants were not actually performing the task (several more participants completed whole blocks with only main clause continuations or only produced two or three tokens which were not main clause continuations across the whole experiment). Because the poisson models only compared production of ORCs across blocks and do not compare ORCs to the other two continuation types, the models should, if anything, be conservative with respect to the magnitude of ORC priming. That said, without exception every ORC continuation elicited by the sentence completion task in this experiment used a pronominal subject rather than the full NP subject that was used in the stimuli sentences. The ORC penalty has been observed to disappear with pronominal subjects (Gordon et al., 2001), which may indicate that pronominal and full-NP ORCs could have different behavior with respect to priming (although clearly overlapping, given our results). We leave this for further research. The relevant conclusion for the current purposes is that the sentence completion results probably underestimate the relative frequency of ORCs overall, but not necessarily the relative frequency of ORCs with full-NPs.

---

<sup>13</sup>Technically, Surprisal is a property of lexical continuations, not whole syntactic constructions (Levy, 2008). However, a component of the Surprisal of a lexical item is the information that it contributes about the syntactic structure. The calculation that we are applying here attempts to reverse engineer the syntactic component of expectation by applying the Surprisal calculation to all of the RC sentence completions in aggregate.

### *Bayesian Analysis*

The Bayesian analysis in this section, and in all sections in this chapter uses the `brms` R package (Bürkner, 2018). All models used the same formula of fixed and random effects as the NHST analysis, unless otherwise noted. The link function was a standard gaussian, as were the priors.

The NHST analysis yields a single decision about whether a parameter is significantly different from zero or not. But beyond this, linear models may or may not be clear about the full dimensions of the analysis, especially because many features of the data, such as the approximation of normality in the data distribution, can impact the interpretation of linear regression coefficient estimates in unpredictable ways.

One value of the Bayesian analysis is the ability to see the full range of credible values and how densely clustered they are. This information can help determine whether to treat a null result as a true null, or whether it may be more likely to be a Type II error (for instance, due to lack of power or an uncontrolled factor). If the probability mass of the posterior is tightly clustered around zero, then it is possible to have more confidence that the true value is within a range that is not theoretically-meaningfully different from zero. If the probability mass is more spread across a wide range of values, including some that are substantially different from zero, then a Type II error becomes more likely.<sup>14</sup> As the interpretation of this particular study leans heavily on the null result for the interaction, the Bayesian analysis stands to make a useful contribution.

As the interaction is where the Bayesian analysis stands to be most informative, we turn to it first. Table (5.8) shows that not only does the credible Bayesian interval for the interaction term include zero across both baseline comparisons, the interval is actually tightly clustered around 0. This indicates that the model was able to get a good estimate, and has a high degree of confidence that the true value for interaction parameter is quite minimal. This is also true of the posterior estimate of the `ORDER` term.<sup>15</sup>

The credible intervals are comparably tight in Total Time, as seen in Table (5.9). This is informative, because in the descriptive results in Figure (5.3) it looks as though the early facilitation for ORCs is more pronounced in Total Time than Go Past. The Bayesian model indicates that this is better accounted for as an increased main effect of `ORDER` applied to both ORCs and SRCs.

---

<sup>14</sup>Note that these are just statements about relative likelihood. A Type II error is always a possibility with a null result.

<sup>15</sup>Interestingly, the Go Past model's estimate of the standard deviation for the random slope of the ORC vs. SRC by-subject random effect when applied was quite large, although it also had one of the widest credible intervals, indicating a lack of certainty from the model. This suggests that there may be quite a bit of variability associated with the size of ORC penalty by subjects. This was apparently not substantially modulated by `ORDER`, as the corresponding by-subject slope for the interaction was quite small in comparison.

	$\hat{R}$	$N_{eff}$	Mean	SD	2.5%	97.5%
<i>ORC vs. SRC</i>						
Order	1.00	4000	-2.46	1.19	-4.78	-0.18
SentenceType	1.00	4000	70.80	22.48	27.23	114.28
Order x SentenceType	1.00	4000	-3.07	2.23	-7.46	1.25
<i>ORC vs. Complement</i>						
Order	1.00	4000	-2.72	1.19	-5.09	-0.41
SentenceType	1.00	4000	75.69	21.99	33.17	118.81
Order x SentenceType	1.00	4000	-2.41	2.41	-7.20	2.28

**Table 5.8.** Outcomes of Bayesian modeling for both embedded clause comparisons in Go Past at the critical NP.

	$\hat{R}$	$N_{eff}$	Mean	SD	2.5%	97.5%
<i>ORC vs. SRC</i>						
Order	1.00	4000	-8.04	1.56	-11.18	-4.95
SentenceType	1.00	4000	74.90	24.38	27.76	122.28
Order x SentenceType	1.00	4000	-2.16	2.64	-7.24	3.16
<i>ORC vs. Complement</i>						
Order	1.00	4000	-2.74	1.20	-5.07	-0.37
SentenceType	1.00	4000	75.95	22.42	31.32	119.23
Order x SentenceType	1.00	4000	-2.46	2.31	-6.99	1.96

**Table 5.9.** Outcomes of Bayesian modeling for the embedded clause comparisons in Total Time at the critical NP.

### 5.3.5 Discussion

A full discussion of the implications of Experiment 3 will require comparison with the results of Experiment 4. We will turn to the combined results of these two experiments in Section 5.5. This section summarizes the intermediate implications from Experiment 3 and sets the stage for Experiment 4.

The central finding from Experiment 3 was that the critical interaction predicted by statistically-sensitive Adaptation accounts was never significant under any measure, using either SRCs or the Complement condition as a baseline for the ORC penalty. This clearly contrasts with the findings of Fine et al. (2013); Tooley and Traxler (2018) and particularly Wells et al. (2009), all of whom found an interaction of ORDER x SENTENCETYPE.

To the extent that this is a true null finding, it argues against the application of implicit learning/Adaptation theories of priming to comprehension. The most telling comparison in this respect is the comparison of ORCs to SRCs. The fullest prediction of a Surprisal-based account would have been crossover interaction that resulted in SRCs actually becoming harder than ORCs later in the experiment. The crossover should have occurred when the comprehender’s representation of the frequencies for grammatical rules reached 50/50 for ORCs vs. SRCs and then continued to

shift in favor of more frequent ORCs. A less extreme interaction would also have been possible, if the adaptation rate parameter were somewhat slower and had combined with the facilitation from a general ORDER effect. Even in these circumstances, if SRCs never actually appeared to get more difficult over the course of the experiment, an interaction with ORCs would have been expected such that the improvement in ORCs over the course of the experiment should have been much greater than the facilitation for SRCs. Given that ORCs were four times more likely than SRCs in this experiment, if participants were sensitive to a frequency-based competition between the two structures, it would have been expected to appear in at least some measure. Without any such evidence, our preliminary conclusion based on the data from Experiment 3 alone would be that there is no evidence of Syntactic Adaptation.

There are, as always, some caveats about interpreting a null effect in this case. For one thing, although the number of participants in this experiment matched the number in Fine et al. (2013), it is entirely possible that we lacked the proper power to reliably find an adaptation effect.

Alternatively, it is possible that the Adaptation rate was too rapid for this study. In both Go Past and Total Time, Figure (5.3) shows an early and substantial speed up of RTs relative to both baselines, which then levels off. This facilitation occurs almost entirely within the first four ORC tokens that a participant encounters, while the two baseline conditions exhibit very little change in RT. Such an effect would appear to be consistent with the “rapid” part of “rapid” and statistically-sensitive Syntactic Adaptation as put forth in Fine et al. (2013). If the speed up over the course of these few initial trials reflects all the time that it takes for a participant to adapt to their syntactic environment and after that comprehenders reach an asymptote of performance beyond which improvement is *de minimis*, then the statistical significance of the ORDER interaction could be diminished by the inclusion of too many trials. In other words, trials at the asymptotic level of adaptation could swamp the trials in which there was meaningful adaptation.

Proponents of an expectation-based theory might equally note that towards the end of the experiment, SRCs appear to reach the same level of difficulty as ORCs and even briefly cross in later parts of the experiment (beginning after position  $\sim 12$ ), to have longer processing times than ORCs. It is difficult to tell from this data whether the temporary crossovers represent normal amounts of noise or the beginning of a real crossover interaction. If the latter half of the present experiment does indeed reflect performance close to an asymptote of adaptation, where any remaining change is slow, then a developing crossover interaction plus noise could produce the pattern in the results. The expectation would be that if the experiment went on much longer,

points where SRCs were more difficult than ORCs would become more and more frequent until noise was not sufficient to overwhelm the flip.

While we acknowledge these caveats and encourage further research, we argue that the most natural interpretation of these results is the true absence of an effect, for several reasons. First, as previously noted, is the consistency with which the interaction fails to reach significance across measures and comparison constructions, even when other relevant effects are fully significant (particularly the ORC penalty relative to the two control conditions). Second, the Adaptation account as just outlined relies on rates of adaptation being simultaneously too fast to detect for ORCs (in the early speed up) and too slow to detect for SRCs (in the long tail with some crossover). While the log-scale of Adaptation makes such rates technically possible, the opposite requirements for the adaptation rate across constructions makes the Adaptation account seem somewhat less plausible. Third, and more substantially, the presence of both dative and ORC priming in the sentence completion task suggests that the experiment did have the ability to prime ORCs, if such an effect existed in comprehension.<sup>16</sup>

Nevertheless, both the concern that adaptation may be too fast in ORCs or too slow in the Competing condition can be addressed by investigating a target construction which has less starting expectation overall and a greater spread/difference in starting expectation from the Competing control. The NP/Z garden path paradigm meets these requirements, so we now turn to that comparison in Experiment 4.

## 5.4 Experiment 4

Experiment 4 was the counterpart of the frequency manipulation begun in Experiment 3. Where ORCs were frequent target construction in Experiment 3, they were rare in Experiment 4. Instead, *NP/Z* → *Zs* were the frequent structure. This formed the basis for the between-subjects frequency manipulation.

The instructions for the sentence completion section of the experiment remained largely the same, except that participants were consistently told to include punctuation if they could,<sup>17</sup> as a

---

<sup>16</sup>If anything, the effect size of priming for ORCs is even larger than for PO datives, although we interpret the relative slopes with caution due to the small sample size. Regression slopes are generally flawed measures of effect size because they are still unit-based statistics (the exact same data given in milliseconds and seconds will not have precisely the same slope). In this case, because the ORC and dative sentence completion data have the same raw units from the same task, the two regression coefficients are sufficient for a comparison of relative effect sizes.

<sup>17</sup>In general, participants in Experiment 3 were given the same instructions, but the change was to be universally consistent about it in Experiment 4. There was a worry that mentioning punctuation would induce editing that would

way to estimate the extent to which participants thought commas were necessary to include in an NP/Z structure. The outcomes seem to indicate that participants were fairly good at following this instruction.

#### 5.4.1 Participants

A second set of 72 participants was recruited from the same population as Experiment 3. The rejection criteria for comprehension question accuracy was again set to 80%. In point of fact however, accuracy overall was slightly lower than for Experiment 3, owing most likely to the difficulty of comprehending NP/Z sentences, even if the questions never specifically targeted the garden path ambiguity. To avoid too much loss of data, participants in the range from 75% to 80% were replaced, while participants in the same range as those rejected in the original study (<70%) were removed without replacement. Six subjects were replaced and one subject was rejected under these criteria. One additional subject was removed due to file error. Finally, one subject that was included in the eyetracking analysis had no accompanying sentence completion data due to a technical error. As with Experiment 3, we would have rejected participants who lost more than 25% of critical trials due to blinks or track loss, however there were no such participants.

#### 5.4.2 Procedure

The procedure for this experiment was identical to the procedure for Experiment 3.

#### 5.4.3 Coding Sentence Completions

The dative sentence completions used the same coding schema as in Experiment 3.

New in this experiment was a coding system for the NP/Z sentence completions. Unlike the dative and RC codes, these had two parts. The first part simply noted whether or not the subordinate and matrix clauses were separated by a comma. The second part concerned what directly followed the subordinate verb, and these were broken into four main categories —*intransitive*, *transitive*, *adjunct*, and *particle verb* —each with its own subdivisions.

The intransitive category was the one that was most important for this experiment, as this is the type that is closest to the dominant Z target structure and therefore the type which would be primed if priming is possible in this experiment. Within this category, there were three separate sub-types. The first was a Z continuation, so called because it was a match to the eyetracking

---

interfere with automaticity of sentence completion, so this was balanced by explicitly telling participants not to worry about typos and to put down whatever came to them first.

stimuli: the subordinate verb was intransitive but was followed by an NP which could have grammatically been the object of the subordinate verb but was actually the subject of the matrix verb. For example, the initial fragment *While the vikings raided* was completed as *the non-vikings were terrified.* and the initial fragment *While the masked bandit dueled* was completed as *, the unmasked bandit also dueled.*<sup>18</sup> The second possible continuation type was the *incremental* type. In this case, the subject NP of the matrix clause could not ultimately have been a possible object for the subordinate verb, but was such that an incremental parser could have begun to parse the NP at least part way and still entertained the possibility that it could have been the subordinate object. For instance, the fragment *While the motorcycle rider parked* was completed by one participant with *the man watched from the window.* The NP *the man* is an impossible object of *parked*, but there is a brief period where an incremental parser does not yet know whether it is parsing *the man* or *the man's bike*, the latter of which would be a potential object of *parked*. These responses were distinguished from the third type of *intransitive* continuation in which it was immediately obvious that the matrix subject NP could not be the object of the initial verb. In practice, this third type was reserved for continuations with clearly case-marked pronouns, such as continuing *Despite the fact that the scouts patrolled with , they did not find any clues.*

The other two main categories of interest were the *transitive* and *adjunct* continuations. *Transitive* continuations included any time the subordinate verb was followed by a NP, PP, or CP argument. *Adjunct* completions involved a subordinate verb followed by an adverb or adjunct PP.<sup>19</sup> If multiple arguments/adjuncts were present, the arguments always took precedence over the adjuncts and the first argument took precedence over other arguments. For instance, if the subordinate verb were followed by an adverb, an NP argument, and then a PP argument, this would be classified as a transitive NP continuation. In some cases, it was unclear whether a PP in the subordinate clause should be considered an argument or a PP, in which case the coder made the best judgment they could. Fortunately, relatively little of the interpretation rests on these tokens, since the primary category of interest was the *intransitives*. Moreover, the lion's share of the *transitive* continuations were realized as bare NPs, not PPs. There were also relatively few adjunct

---

<sup>18</sup>Comma usage is preserved as given by the participants in all examples in this section.

<sup>19</sup>There was some concern about whether tokens labeled as adjuncts here should be considered as intransitive. They were excluded from the inferential statistics. However, the descriptive statistics are presented separately so that it is possible to compare them to the other categories. In point of fact, this is the least numerous category, and the descriptive statistics make it clear that they do not differ substantially from the patterns of the *intransitive* continuations, so we believe that this should not have an outsized impact on the analysis.

continuations, meaning that the argument/adjunct ambiguity applied to a small number of the total continuations.

Lastly, the *particle verb* continuations for the NP/Z fragments had the same criteria as for the dative fragments. As before, it is assumed that the particle verb complex has different subcategorization properties than the simplex verb, and therefore were considered a separate category entirely, although they were included in the calculation of by-subject means and percentages.

#### 5.4.4 Results

The analysis of Experiment 4 was exactly like Experiment 3, except that the dominant token which determined position for the ORDER factor was the NP/Z → Z condition instead of the ORC condition. The coding of the target and controls in each model mirrors that used in Experiment 3. This variable coding is given in Table (5.10).

NP/Z → Z	0.5
NP/Z → NP	-0.5
NP/Z → Z comma	-0.5
Order	<i>continuous</i>

**Table 5.10.** Variable coding used in models for Experiment 4.

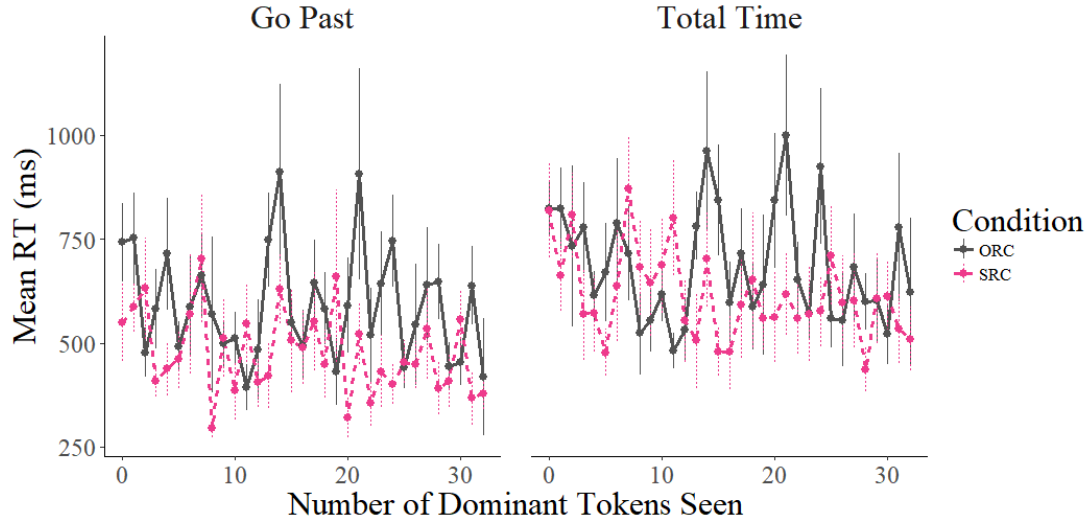
#### *Embedded Clause Conditions*

There are fewer embedded clause comparisons in Experiment 4 because the Complement condition was not included in this experiment. In addition, the models in this subsection did not converge if they included any random slopes by item. This is most likely due to the fact that each subject saw a different subset of 8 ORC sentences from the original 32 presented in Experiment 3. The absolute number of observations per item was therefore reduced, and the mixture of ORC lists across subjects in particular hampers the ability of the model to assign variability to items. The random intercept by item and the full random slope structure by participants remained in the models.

In *Go Past*, the model yielded a main effect of ORDER ( $\beta=-2.58$ ,  $SE=1.22$ ,  $t=-2.12$ ) and a main effect of SENTENCETYPE (ORC vs. SRC) ( $\beta=98.51$ ,  $SE=33.33$ ,  $t=2.96$ ). The interaction did not reach significance ( $\beta=1.54$ ,  $SE=2.42$ ,  $t=0.63$ ).

In probability of regression, only a main effect of SENTENCETYPE reached was reliable ( $\beta=1.24$ ,  $SE=0.18$ ,  $p < 0.0001$ ). The critical interaction remained non-significant ( $\beta=-0.001$ ,  $SE=0.017$ ,  $p=0.939$ ).

In *Total Time*, again, both main effects were significant (ORDER:  $\beta=-3.44$ ,  $SE=1.52$ ,  $t=-2.26$ ; SENTENCETYPE:  $\beta=68.98$ ,  $SE=26.27$ ,  $t=2.63$ ), but the interaction was not ( $\beta=-0.06$ ,  $SE=2.77$ ,  $t=-0.02$ ).



**Figure 5.5.** Change in Mean RT of the embedded clause conditions over the course of Experiment 4. Error bars represent standard error.

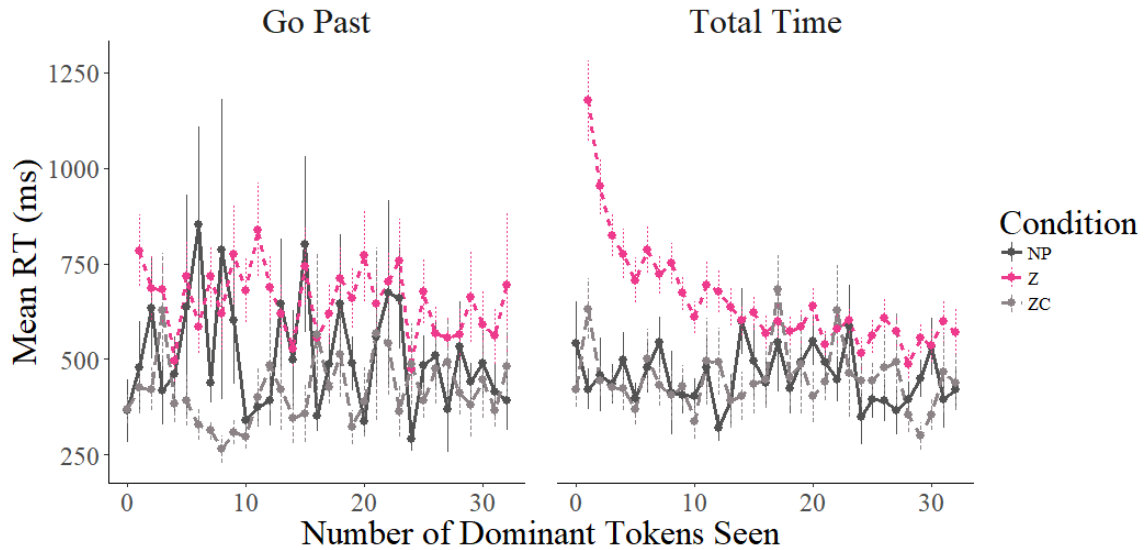
The outcomes of all of these models are consistent with the overall pattern of two additive effects from ORDER and the ORC penalty, which do not interact.

#### *Garden Path Conditions*

In Go Past, the comparison of Z to the NP baseline had a main effect of SENTENCETYPE ( $\beta=141.04$ ,  $SE=45.81$ ,  $t=3.08$ ), but no other significant effects. The same held true when the Z condition was compared to the Comma condition (SENTENCETYPE:  $\beta=227.16$ ,  $SE=46.54$ ,  $t=4.88$ ).

Probability of regression found a main effect of SENTENCETYPE for both Z vs. NP ( $\beta=0.39$ ,  $SE=0.15$ ,  $p < 0.01$ ) and Z vs. Comma ( $\beta=0.95$ ,  $SE=0.16$ ,  $p < 0.0001$ ). The interaction was non-significant for both comparisons (Z vs. NP:  $\beta=0.0002$ ,  $SE=0.02$ ,  $p=0.98$ ; Z vs. Comma:  $\beta=-0.02$ ,  $SE=0.02$ ,  $p=0.16$ )

The pattern changed in Total Time. Here, the Z vs. NP model had both a significant main effect ORDER ( $\beta=-6.19$ ,  $SE=1.39$ ,  $t=-4.45$ ) and a significant main effect of SENTENCETYPE ( $\beta=202.14$ ,  $SE=26.16$ ,  $t=7.73$ ). More importantly, the critical interaction was significant in this measure ( $\beta=-12.02$ ,  $SE=2.28$ ,  $t=-5.27$ ), reflecting a massive speed up in RTs for the Z conditions over the first  $\sim 10$ -15 Z tokens, visible in Figure (5.6). Again, the same pattern held regardless of baseline. The Z vs. Comma versions of ORDER ( $\beta=-7.07$ ,  $SE=1.44$ ,  $t=-4.89$ ) and SENTENCETYPE ( $\beta=208.49$ ,  $SE=24.15$ ,  $t=8.63$ ) were significant, as was the interaction ( $\beta=-10.24$ ,  $SE=2.10$ ,  $t=-4.87$ ).



**Figure 5.6.** Change in Mean RT of the NP/Z conditions over the course of Experiment 4. Error bars represent standard error.

#### Sentence Completion Results

PO datives were once again primed in this experiment. Interestingly, in this experiment the priming effect now manifests only in the statistical test with PO datives themselves ( $\beta=0.255$ ,  $SE=0.094$ ,  $p < 0.01$ ). The statistical comparison over PO dative + VP-attached PPs was not significant in this experiment ( $\beta=0.071$ ,  $SE=0.071$ ,  $p=0.32$ ), and in fact the trend for VP-attached PPs alone goes in the opposite direction (although arguably this is *because* priming for PO responses shifted responses into the PO category out of all the others).

	Raw Count	<i>Pre-test</i> Mean Count per Block	Mean %	Raw Count	<i>Post-test</i> Mean Count per Block	Mean %
PO	200	2.74	19.57	258	3.53	25.24
VP-attached PP	183	2.51	17.91	153	2.09	14.97
DO	189	2.59	18.49	158	2.16	15.46
Not Dative	354	4.85	34.64	348	4.77	34.05

**Table 5.11.** Descriptive results for dative fragments in the sentence completion task of Experiment 4. Mean per Block and Mean % were both calculated by-subjects.

While the shift from priming VP-attached PPs to priming just PO datives is worth noting, in these experiments there was no reason to expect one primed structure or another, only that priming from PO datives would exist. It would be of general interest to know whether the slightly different environments between the two experiments was enough to cause the shift, and if so, what the cause was. However, for the present purposes of the experiment what matters is that there

		$\beta(SE)$	$p$
<b>Dative PPs Only</b>	<i>Intercept</i>	1.135 (0.047)	<0.00001
	<i>Pre/Post</i>	0.255 (0.094)	< 0.01
<b>Any VP-attached PP</b>	<i>Intercept</i>	1.693 (0.036)	<0.00001
	<i>Pre/Post</i>	0.071 (0.071)	0.320

**Table 5.12.** Outcome of the mixed models applied to the dative sentence completion results of Experiment 4. Pre/Post test was coded as pre-test = -0.5 and post-test = 0.5.

was priming from the PO datives in the eyetracking to the sentence completion results. We do not see a way, in either experiment, to explain a significant increase in the type of responses primed by PO datives without appealing to the influence of the 32 datives in the eyetracking sentences. This demonstrates once again that the experiment had the raw ability to induce priming when the target was produced and therefore could potentially have shown comprehension priming if there was any such effect. We leave the question of what, if anything, caused the shift between the two types of primed responses open to further research.

		Raw Count	Mean Count per Block	Mean %
<b>Pre-test</b>				
<i>Intransitive</i>	Comma	95	1.30	9.3
	No Comma	81	1.11	7.93
<i>Transitive</i>	Comma	682	9.34	66.73
	No Comma	30	0.41	2.94
<i>Adjunct</i>	Comma	55	0.75	5.38
	No Comma	3	0.04	0.29
<b>Post-test</b>				
<i>Intransitive</i>	Comma	57	0.78	5.58
	No Comma	74	1.01	7.24
<i>Transitive</i>	Comma	732	10.03	71.62
	No Comma	35	0.48	3.42
<i>Adjunct</i>	Comma	41	0.56	4.01
	No Comma	3	0.04	0.29

**Table 5.13.** Descriptive statistics for Mean per Block and Mean % were both calculated by-subjects.

Poisson regression models for the NP/Z sentence completions took raw counts of intransitive completions in the pre- and post-test as input (no other completions were tested). There was a significant effect of block on production of intransitive ( $\beta=-0.301$ ,  $SE=0.116$ ,  $p < 0.01$ ), but curiously the effect goes opposite of the predicted direction: participants were *less* likely to produce an intransitive completion after seeing the 32 intransitive tokens in the eyetracking portion of the experiment. Even more curiously, there was no significant modulation of the decrease by comma usage (although the interaction was marginal), but numerically the trend seems to be driven largely by a decrease in the number of intransitive tokens with commas. This argues against

participants applying a conscious strategy of having learned that intransitives are easier to parse with commas. Such a strategy would have predicted a massive shift to preferring intransitives with commas in the post-test. Instead, the data are more consistent with a strategy that avoids intransitives, but which may also associate intransitives with a lack of comma usage. Note that the results do not appear to support an account where participants were simply too rushed at the end of the experiment to be bothered with punctuation, because they had no problem including commas with transitive completions.

	$\beta(SE)$	$p$
<i>Intercept</i>	-1.065 (0.058)	<0.0001
<i>Comma</i>	0.051 (0.116)	0.662
<i>Pre/Post</i>	-0.301 (0.116)	<0.01
<i>Comma x Pre/Post</i>	0.420 (0.232)	0.070

**Table 5.14.** Outcome of the mixed effects models applied to the intransitive NP/Z fragment sentence completion results.

To the extent that the trend to use fewer commas for intransitive sentence completions is real, it seems to extend across all three sub-categories of intransitive in the coding schema, as shown in Table (5.15).

			<i>Pre-test</i>			<i>Post-test</i>	
	Comma	Raw Count	Mean Count per Block	Mean %	Raw Count	Mean Count per Block	Mean %
<i>Z</i>	Comma	32	0.44	3.13	11	0.15	1.08
	NoComma	29	0.4	2.84	23	0.32	2.25
<i>Incremental</i>	Comma	20	0.27	1.96	14	0.19	1.37
	NoComma	14	0.19	1.37	14	0.19	1.37
<i>Pronoun</i>	Comma	43	0.59	4.21	32	0.44	3.13
	NoComma	38	0.52	3.72	37	0.51	3.62

**Table 5.15.** Descriptive results for the sub-types of the intransitive completions of NP/Z fragments. Mean count per Block and Mean % are both adjusted by-subjects.

### *Bayesian Analysis*

As with Experiment 3, there was an additional Bayesian analysis applied to the main within-subject comparisons of the target against its two baselines. As before, the Bayesian models do not change the significance decisions from the NHST analysis, so much as provide more nuance.

In many ways, the outcomes of these models are quite similar to the within-subject Bayesian models from Experiment 3.<sup>20</sup> The credible intervals for SENTENCETYPE are quite wide, indicating a high degree of uncertainty about the value of this parameter. As the estimate of the by-subject random slope for SENTENCETYPE (not given) was comparably wide in both models, it is reasonable to think that the width results from a great deal of variation within and between subjects in how they react to the NP/Z garden paths. Again, the credible intervals are nowhere near zero, indicating that the NP/Z contrast remains robust.

	$\hat{R}$	$N_{eff}$	Mean	SD	2.5%	97.5%
<i>Z vs. NP</i>						
Order	1.00	4000	-2.75	1.98	-6.66	1.16
SentenceType	1.00	4000	139.53	44.32	52.74	226.80
Order x SentenceType	1.00	4000	0.25	4.33	-8.19	8.53
<i>Z vs. Comma</i>						
Order	1.00	4000	-0.89	1.97	-4.77	2.83
SentenceType	1.00	4000	227.39	48.70	133.90	322.54
Order x SentenceType	1.00	4000	-3.36	3.92	-11.01	4.35

**Table 5.16.** Outcomes of the Bayesian models for both garden path comparisons in Go Past (at the disambiguating verb)

	$\hat{R}$	$N_{eff}$	Mean	SD	2.5%	97.5%
<i>Z vs. NP</i>						
Order	1.00	3821	-6.22	1.41	-9.04	-3.53
SentenceType	1.00	3439	201.92	26.79	148.83	254.89
Order x SentenceType	1.00	4000	-12.00	2.41	-16.64	-7.24
<i>Z vs. Comma</i>						
Order	1.00	4654	0.30	1.85	-3.28	3.94
SentenceType	1.00	2293	398.32	58.79	284.71	511.45
Order x SentenceType	1.00	3776	-25.00	4.64	-34.21	-15.83

**Table 5.17.** Outcomes of the Bayesian models for both garden path comparisons in Total Time (at the disambiguating verb)

The credible intervals for ORDER and the interaction are much more compact and are either much closer to zero or include zero. Especially in Total Time, the estimates trend distinctly negative, indicating a greater degree of facilitation over the course of the experiment for Zs than the two baselines. Recall that the ORDER factor was centered for all analyses, meaning that the parameter estimate for the interaction reflects the amount of facilitation for Zs, above and beyond

<sup>20</sup>The *Z vs. Comma* model at Total Time required additional iterations beyond the other models. Although  $\hat{R}$  reached 1.00 for the three reported parameters in the original model that we ran,  $N_{eff}$  was as low as 1502. The reported model had 3000 iterations (1500 warmup) over 4 chains, while all other models used 2000 iterations (1000 warmup) also over 4 chains.

the main effect of ORDER, at the 16<sup>th</sup> Z token. Visual inspection of Figure (5.6) suggests that by this point much of the apparent adaptation has already occurred, which means that an estimated 10-30ms of continuing facilitation for each Z token is impressive.<sup>21</sup>

#### 5.4.5 Discussion

The NP/Z comparisons in Total Time in Experiment 4 appear to be exactly the pattern predicted by rapid, statistically-sensitive Adaptation. The interaction of ORDERXSENTENCETYPE driven by a dramatic speed up for the more difficult construction is, in fact, precisely the data pattern that has been used to argue for Syntactic Adaptation in other studies such as Wells et al. (2009), Fine et al. (2013) and Tooley and Traxler (2018). Critically it is not just an ORDER effect, which would be too easily written off as task adaptation.

There is now a tension between the conclusions from Experiment 4, and those from Experiment 3 which did not find evidence of Adaptation. However, Adaptation theory predicts that the effect size should be larger for the NP/Z paradigm, given that the Z target starts off more difficult than ORCs. The increased difficulty means that it has more potential facilitation to gain. And indeed, the early speed up over the first ~4 ORC tokens that was noted in Experiment 3 has the appearance of a weaker version of the NP/Z interaction seen in this study. The ORC version also develops over the same timeframe in the eye movement record as the NP/Z interaction, starting with a hint in Go Past that becomes more prominent in Total Time. Therefore, based on the ORDER x SENTENCETYPE interaction, it would seem that the most likely way to resolve the tension between the two experiments is by concluding that ORCs and Zs are subject to the same basic facilitation interaction, but that the overall effect is much stronger for NP/Z garden paths. And because all these facts align with the predictions of statistically sensitive Syntactic Adaptation, based on the within-subject comparisons, true Syntactic Adaptation would seem to be the most likely explanation.

Given the RT results, the lack of production NP/Z priming in the sentence completion task is surprising. If Syntactic Adaptation is intended to be the comprehension correlate of abstract priming in production, then the fact that the production priming and adaptation results for ORCs and NPZs flip is mysterious. We return to this fact in the discussion of the between-subjects results.

---

<sup>21</sup>Though this term is likely inflated somewhat by the simple fact that this is a linear model trying to fit what appears to be a curvilinear interaction. However, even adjusting for this, as much as 30ms of learning for a single exposure to a structure is substantial.

The next section turns to the between-subject frequency manipulation to see if the additional comparisons are in agreement with the within-subject conclusions.

## 5.5 Combined Analysis of Experiments 3 & 4

The results of the within-subjects comparisons from Experiments 3 and 4 in the previous section on the whole appeared to lean toward a Syntactic Adaptation account, but do so based primarily on the ORDER  $\times$  SENTENCE TYPE interaction in Total Time of Experiment 4. This section presents the between-subjects comparisons for the two experiments. The between-subjects results are in a position to critically inform the interpretation of the within-subjects interactions, because while the within-subjects interactions address the cross-constructional competition that distinguishes Adaptation from other theories, only the between-subjects results are in a position to adjudicate whether the within-subject results can actually be attributed to *frequency*, as Adaptation suggests. The models which test this prediction are those that include both ORDER and FREQUENCY as factors and apply these to one construction at a time. Recall that in this case, FREQUENCY is equivalent to EXPERIMENT in the actual design. As before the critical term is the interaction, which now indicates whether the magnitude of the facilitation due to ORDER is modulated by FREQUENCY.

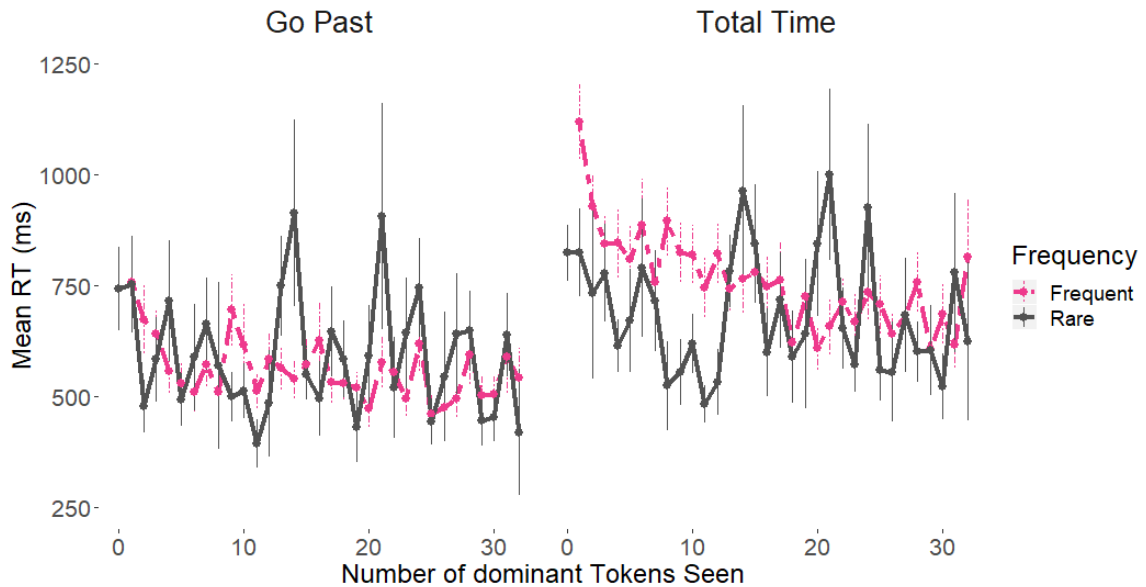
The critical regions remain the same. Because subject and item did not overlap across experiments, the models did not contain random effects for subject and item by experiment. However, the models retained the general random effects for subject and item, as well as additional random effect for subject and item by ORDER. In addition, the ORDER factor was always calculated relative to the dominant target in the experiment from which the trial originated, i.e., ORCs in Experiment 3 and NP/Z  $\rightarrow$  Z readings in Experiment 4. As before,  $t=2$  is taken as the threshold for significance (Gelman & Hill, 2007).

### 5.5.1 Embedded Clause Results

The results of models for between-subjects on ORC sentences are given in Table (5.18). In Go Past times, there was a significant effect of ORDER ( $\beta=-3.09$ ,  $SE=1.13$ ,  $t=-2.73$ ), but neither the main effect of FREQUENCY nor the interaction of ORDER  $\times$  FREQUENCY approached significance. The same pattern held in both Probability of Regression and Total Time: again, the main effect of ORDER was significant (p(Regression):  $\beta=-0.021$ ,  $SE=0.006$ ,  $p < 0.001$ ; Total Time:  $\beta=-6.48$ ,  $SE=1.43$ ,  $t=-4.53$ ), but the main effect of FREQUENCY was not, nor was the interaction.

	Order	Frequency	Order x Frequency
Go Past	-3.09(1.13)	28.49(44.79)	1.84(2.26)
p(Regression)	-0.021(0.006)	0.283(0.198)	-0.009(0.012)
Total Time	-6.48(1.43)	-75.99(67.63)	4.33(2.86)
	Order = continuous	Frequent = -0.5	Rare = 0.5

**Table 5.18.** Between-subjects model coefficients for ORCs. Standard errors are in parentheses. Significant outcomes are highlighted in gray.



**Figure 5.7.** Between-subject plots of change in Mean RT for ORCs sentences. Error bars represent standard errors.

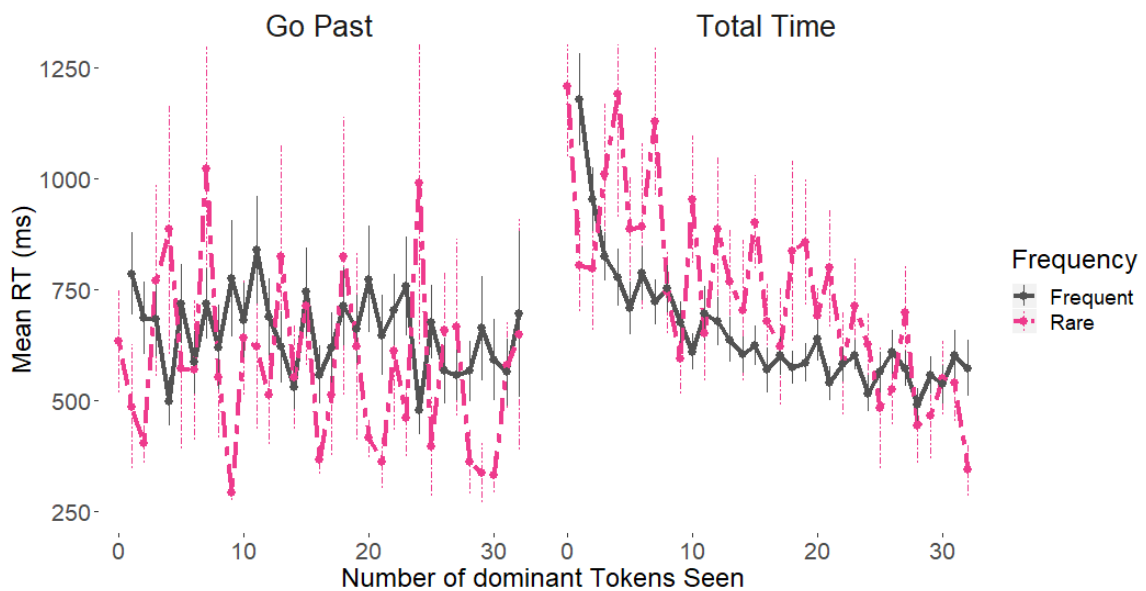
The presence of a main effect of ORDER without a corresponding main effect of FREQUENCY or an interaction indicates that the rate of “adaptation” to ORCs was indistinguishable when they were rare from when they were four times more frequent in context. This is inconsistent with expectation-based views of comprehension priming, but given that the within-subjects interaction of ORDER with the ORC penalty did not come out significant in Experiment 3, there is a worry that this is an issue of insufficient power.

Descriptively, the ORC RTs across-experiments track each other quite closely in Go Past, except for an odd jog in the middle of the Experiment 4 results. In Total Time the pattern is more complicated. In the first half of the ORDER variable there is a substantial separation between the two experiments. However a closer inspection reveals the slope is in fact fairly similar over the early positions, indicating that this is likely not Syntactic Adaptation either, but simply an

additive difference in RTs between experiments. We will return to a discussion of this somewhat mysterious data pattern later, after considering the remaining between-subject tests.

### 5.5.2 NP/Z Results

The within-subjects comparisons for the NP/Z paradigm were consistent with true Syntactic Adaptation, which makes the between-subjects comparisons potentially even more informative than the embedded clause conditions. However, the descriptive results in Figure (5.8) show that the Z conditions appear to overlap across the frequency manipulation even more than the ORC RTs.



**Figure 5.8.** Plots of change in Mean RT based on ORDER for NP/Z→Z garden paths

Accordingly, the between-subjects comparison for NP/Z→Zs found no significant effects in Go Past. In Total Time there was a significant main effect of ORDER ( $\beta=-15.11$ ,  $SE=2.47$ ,  $t=-6.13$ ), but as with the ORC comparisons, the effect of FREQUENCY and the interaction failed to reach significance.

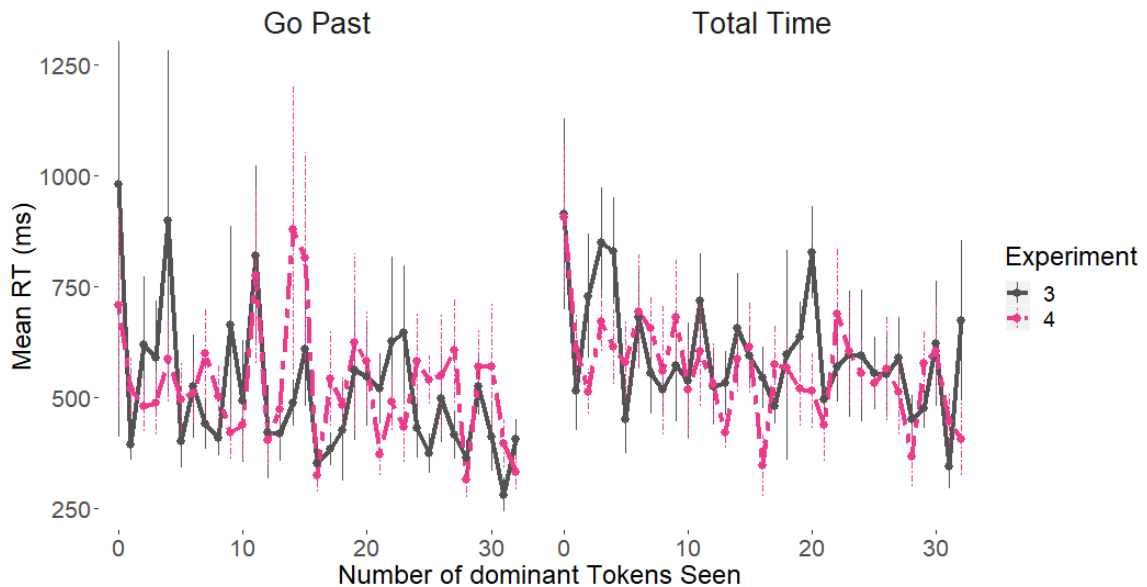
This pattern of results strongly indicates that the frequency of NP/Zs in the experimental context had little to no impact on the rate of “adaptation” found in the within-subject results.

		$\beta$	SE	$t/p$
<i>Go Past</i>	ORDER	-2.46	2.00	-1.23
	FREQUENCY	11.48	62.34	0.18
	ORDER x FREQUENCY	-0.82	4.01	-0.21
$p(\text{Regression})$	ORDER	-0.007	0.008	0.391
	FREQUENCY	-0.089	0.213	0.674
	ORDER x FREQUENCY	-0.001	0.017	0.939
<i>Total Time</i>	ORDER	-15.11	2.47	-6.13
	FREQUENCY	-92.10	59.39	-1.55
	ORDER x FREQUENCY	3.63	2.85	1.27
Order = continuous		Frequent = 0.5	Rare = -0.5	

**Table 5.19.** Between-subjects comparisons for NP/Z→Z garden paths. For the linear models for Go Past and Total Time, the significance statistic is  $t$  and it is  $p$  for the logistic model applied to probability of regression.

### 5.5.3 NP/S Results

The conclusion that frequency is not the driving force behind facilitation over the course of these experiments finds some further corroboration in the data pattern from the NP/S garden paths. Recall that these garden paths were rare across both experiments, yet Figure (5.9) shows that they displayed roughly the same magnitude ORDER effect as the NP/Z conditions (on the order of  $\sim 400$ ms from the first position to the last).



**Figure 5.9.** Change in Mean RT of the NP/S → S garden paths over the course of both experiments. Error bars represent standard error.

Furthermore, the NP/S garden paths act as a check on any extraneous differences between the two experiments. While Figure (5.9) certainly demonstrates variability between the two experiments, it does not appear to demonstrate a difference of the size needed to have spuriously produced the ORDER x SENTENCE TYPE interaction that was apparent in the NP/Z conditions of Experiment 4.

Both of these aspects of the NP/S data are furthered by the fact that the interaction is not significant in the models for the NP/S condition, either for Go Past ( $\beta=1.44$ ,  $SE=3.28$ ,  $t=0.44$ ) or Total Time ( $\beta=-0.67$ ,  $SE=2.54$ ,  $t=-0.26$ ). The only effect which did achieve significance across the two models was a main effect of ORDER in Total Time ( $\beta=-3.90$ ,  $SE=1.27$ ,  $t=-3.08$ ).

Unfortunately, as the NP/S were not lexically matched with the NP/Z garden paths, there is no direct comparison between the two types of garden paths. As a result, the conclusions in this subsection are somewhat tentative. Nevertheless, it would have been a cause for concern if the NP/S results had suggested a different account than the ORC and NP/Z sentences. These results from the NP/S conditions are certainly consistent with the conclusions from the more central conditions.

#### 5.5.4 Bayesian Analysis

Once again, a Bayesian analysis was performed for the critical predictions from this section, which this time focus on the ORDER x FREQUENCY interaction. In addition, the Bayesian analysis for the combined data looks at two of the control structures for estimates of the difference between the two experiments independent of the frequency manipulation and the magnitude of the ORDER effect across different levels of sentence difficulty. As before, the base models for the Bayesian analysis are the linear models that were reported in the NHST analysis. The linking function and priors were all set to standard Gaussians.

Before going on to discuss the bulk of the results, there is one trend in these models which is independent of construction and deserves a note up front. One of the immediately noticeable features of the models in this section is how wide the credible intervals are for the FREQUENCY term. Some of this is almost certainly attributable to actual variation in the data, but there are at least two features of the model that contribute to the sizeable width. Recall that these models did not have random slopes relative to FREQUENCY, since this was a between-subjects manipulation. This means that there are fewer random variables available to take up variation specifically related to FREQUENCY. Moreover, subject and item are nested under FREQUENCY in these models,

and parameters higher in the hierarchy are known to frequently have reduced  $N_{eff}$  relative to lower parameters.<sup>22,23</sup> A reduced  $N_{eff}$  may lead to less certainty in the estimation and therefore a wider credible region. We tried ameliorating the  $N_{eff}$  by increasing the number of iterations and chains, but the effect was minimal and in some cases even detrimental. Given that the Gelman-Rubin statistic,  $\hat{R}$ , has reached 1.00 for all the reported parameters, (indicating that all four chains converged on the same estimate for the mean) (Brooks & Gelman, 1998; Gelman & Rubin, 1992), we accept the  $N_{eff}$ s and the FREQUENCY posterior estimations in these models as sufficient.

*Object Relative Clauses*

As has been the case for several previous models, the credible interval for the critical interaction is quite narrow, indicating that the model did not have a lot of uncertainty about the value of the interaction, relative to, say, the estimate of FREQUENCY. This makes a true null result more credible than it otherwise would have been. For instance, all three of the terms in the models in Table (5.20) have credible intervals that contain zero, yet it would be far more of a stretch to assume that the true value of the parameter was only trivially different from zero for FREQUENCY than for ORDER or the interaction.

	$\hat{R}$	$N_{eff}$	Mean	SD	2.5%	97.5%
<i>Go Past</i>						
Order	1.00	4000	-3.09	1.19	-5.42	-0.72
Frequency	1.00	2322	29.33	45.12	-56.81	119.76
Order x Frequency	1.00	4000	1.83	2.40	-2.86	6.55
<i>Total Time</i>						
Order	1.00	4000	-6.48	1.49	-9.39	-3.55
Frequency	1.00	1753	-74.07	70.61	-216.17	64.38
Order x Frequency	1.00	4000	5.26	3.05	-0.57	11.27

**Table 5.20.** Summary of the Bayesian between-subjects models for the Z condition of garden paths.

*NP/Z → Zs*

The Bayesian models on the between-subjects Z target comparison corroborates the findings of the NHST and other Bayesian analyses. As before, the credible interval for the critical interaction

<sup>22</sup>Even though this problem is actually better in brms’s STAN-based system than in a traditional Gibbs sampler (Hoffman & Gelman, 2014; Stenberg, 2007)

<sup>23</sup>Although the majority of items were the same across the experiments, the models had no way to identify this.

has a probability mass closely clustered around values which are close to zero, indicating a trivial impact of FREQUENCY ON ORDER.

	$\hat{R}$	$N_{eff}$	Mean	SD	2.5%	97.5%
<i>Go Past</i>						
Order	1.00	4000	-2.47	2.24	-6.82	1.83
Frequency	1.00	1822	50.08	72.95	-95.59	194.51
Order x Frequency	1.00	4000	-0.23	4.52	-9.11	8.59
<i>Total Time</i>						
Order	1.00	2928	-13.91	1.54	-17.00	-10.97
Frequency	1.00	1480	-95.31	59.01	-211.65	16.79
Order x Frequency	1.00	3166	3.32	3.05	-2.55	9.41

**Table 5.21.** Summary of the Bayesian between-subjects models for the Z condition of garden paths.

The Bayesian models increase the likelihood that the lack of ORDER x FREQUENCY interaction is a true null and therefore that NP/Z garden paths are not subject to Syntactic Adaptation.

*Control Conditions: NP/S garden paths and SRCs*

Because the interactions of ORDER x SENTENCE TYPE are significant even when the between-subjects ORDER x FREQUENCY interactions are not, the results present something of a conundrum. There is apparently *some* kind of “adaptation” which can apply differentially across sentence types, but which is not about the frequency of exposure to any single structure. In order to say more about the properties of this *non*-syntactic adaptation, it is useful to have a sense of how it applies beyond the two target constructions. It is also useful to have some sense of how any general differences between the experiments might impact the RTs for individual constructions. One experiment or the other could have slightly slower RTs over all the sentences for many reasons, ranging from an accident of drawing particularly slow subjects from the population to the impact of changing the mix of sentences across the two experiments.

This section looks at two constructions, NP/S garden paths and SRCs, which were present in both experiments but had no frequency manipulation. NP/S garden paths are quite difficult, and the section on the NHST analysis of RTs for this construction noted that the ORDER effect for NP/Ss was roughly the same magnitude as the target Z condition. SRCs, on the other hand, are easier than either of the target constructions and provide an estimate of non-syntactic adaptation at the lower end of the difficulty scale.

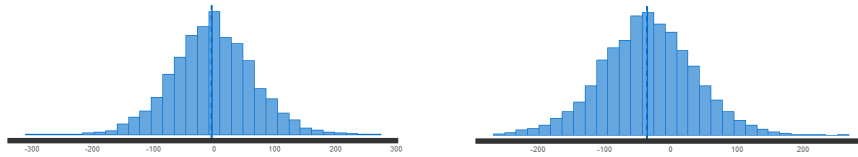
The NP/S garden path models are in Table (5.22). They demonstrate strikingly wide credible intervals for the main effect of EXPERIMENT, which indicates that the models retained quite a bit

of uncertainty about the most likely population value.<sup>24</sup> Also notable is that the means of the EXPERIMENT estimates are minimal, from 35.54ms in Total Time to only 3.40ms in Go Past. While 35ms might make a difference if it indicated a population-level difference across the experiments, it is a comparatively small effect for Total Times. Meanwhile, Figure (5.10) indicates that the posteriors are normal and symmetrical enough that the mean is a reasonable statistic to rely on, as opposed to the median.

Neither of these characteristics points to large, systematic differences between the two experiments that could have interfered with the primary effects of interest.

	$\hat{R}$	$N_{eff}$	Mean	SD	2.5%	97.5%
<i>Go Past</i>						
Order	1.00	2298	-3.65	2.42	-8.23	1.18
Experiment	1.00	1506	-3.40	67.53	-135.36	132.06
Order x Experiment	1.00	2203	0.27	4.72	-9.12	9.27
<i>Total Time</i>						
Order	1.00	4000	-5.07	1.64	-8.34	-1.88
Experiment	1.00	1687	-35.54	70.28	-179.17	103.78
Order x Experiment	1.00	4000	-1.19	3.28	-7.52	5.53

**Table 5.22.** Summary of the Bayesian between-subjects models for NP/S garden paths.



**Figure 5.10.** Posterior distributions of the EXPERIMENT parameter for the NP/S Bayesian between-subjects models. *Left: Go Past; Right: Total Time*

For SRCs, Table (5.23) reports both the results from the relative verb and the relative NP. All the results reported so far focused on the relative NP, because this region was lexically matched to the critical NP that signalled the start of an ORC. However, a comprehender learns that they are reading an SRC when they encounter a relative verb *first*, and then the relative NP. An Adaptation story would have predicted the bulk of the adaptation at the relative verb, and possibly only some spillover at the relative NP. Because our analyses and those of Staub (2010) found very little of note at the relative verb, the thought that any adaptation would end up on the NP spillover

<sup>24</sup>Note that EXPERIMENT is underlyingly the same factor as FREQUENCY in the two target models. The change in nomenclature reflects the slightly different interpretation of the factor in the models.

seemed justifiable. But in this case, where the question is about adaptation of SRCs themselves, it is important to include the relative verb region as well.

Critically, Table (5.23) indicates that there don't seem to be substantial differences in the estimates for either ORDER or the interaction between the two different regions.

	$\hat{R}$	$N_{eff}$	Mean	SD	2.5%	97.5%
<b>Relative Verb</b>						
<i>Go Past</i>						
Order	1.00	4000	-0.33	1.15	-2.66	1.95
Experiment	1.00	2620	1.91	30.38	-58.23	61.34
Order x Experiment	1.00	4000	3.24	2.27	-1.24	7.76
<i>Total Time</i>						
Order	1.00	4000	-3.00	1.27	-5.52	-0.53
Experiment	1.00	1752	-62.98	45.68	-152.64	32.42
Order x Experiment	1.00	4000	2.08	2.54	-2.83	6.96
<b>Relative NP</b>						
<i>Go Past</i>						
Order	1.00	4000	-2.32	1.23	-4.76	0.08
Experiment	1.00	2478	-5.28	42.15	-88.62	77.31
Order x Experiment	1.00	4000	-3.33	2.41	-8.02	1.49
<i>Total Time</i>						
Order	1.00	4000	-5.60	1.42	-8.47	-2.92
Experiment	1.00	1784	-67.06	53.44	-172.80	40.60
Order x Experiment	1.00	4000	1.85	2.76	-3.52	7.28

**Table 5.23.** Summary of the Bayesian between-subjects models for SRCs.

As with the other models, the estimate for the EXPERIMENT parameter is quite wide and includes zero as a credible value. However, in this case both models prefer a much larger estimate for the difference between the two experiments in Total Time relative to Go Past. This is consistent with the trend throughout the results for effects of interest to show up later, typically in re-reading times. Moreover, as in the NP/Z models, Experiment 4 is slightly faster overall, although this effect is not reliable.

The key finding from the analysis of the two controls is that there isn't even a consistent estimate of the main effect of EXPERIMENT when FREQUENCY is factored out. This makes it less likely that even basic differences between RTs can be attributed to something simple like arbitrarily slower readers in one experiment versus another. To the extent that this analysis is reliable, it seems to lean toward all effects being impacted by the particular structure, in a way that may track with the difficulty of the construction.

## 5.6 General Discussion

The primary finding from Experiments 3 and 4 is that neither ORCs nor NP/Zs show the proper constellation of effects to be considered true statistically-sensitive Syntactic Adaptation.

Crucially, the experiments did replicate the types of effects which previous work has used to argue for Syntactic Adaptation. Both NP/Zs and ORCs at times show descriptive and inferential patterns which would seem to be consistent with Syntactic Adaptation. Generally this comes in the form of facilitation over the first few tokens of the experiment for the within-subject comparisons. This is most dramatic for the NP/Zs, but also still a clear trend for ORCs. In fact, the relative magnitude of facilitation for NP/Zs vs. ORCs is consistent with Adaptation accounts, in that NP/Zs begin as a harder structure, and therefore presumably less expected under expectation-based Adaptation. This would suggest that NP/Zs should have much more expectation to gain, and because the linking function for Adaptation is a negative log, should make those gains faster. These are all data points that have been taken as evidence for Syntactic Adaptation in prior work, and it is important that they are replicated here so that the additional controls in this experiment can be evaluated against them.

Despite the presence of data patterns which have previously been used to argue in favor of Syntactic Adaptation, the actual findings of Experiments 3 and 4 argue against implicit learning of syntax in these contexts. Neither the critical within or between subject interactions for ORCs ever reached significance. Meanwhile, the within-subjects interaction of NP/Zs with the control structures did reach significance, but only in very late measures and not accompanied by the between-subjects interaction of `FREQUENCY x ORDER` that was necessary to argue against a task adaptation account.

Therefore the conclusion that we adopt is that there is some type of adaptation present in these experiments, but that it is not directly dependent on the number of times that a comprehender has encountered a specific structure. The insensitivity to frequency means that this type of “adaptation” cannot be *Syntactic* Adaptation. However, it is still interestingly sensitive to the properties of syntactic structures in the experimental context, particularly how difficult each structure is to process. On this basis, we propose that the type of “adaptation” in these experiments, and in the prior literature, is better characterized as adaptation to the overall difficulty of the experimental environment. This proposal has a lot more explanatory power than it might seem at first blush. Adapting to difficulty fits the production, strategic, timing, and cross-construction results better than Syntactic Adaptation.

### *Comprehension-to-Production Priming*

The majority of evidence against Syntactic Adaptation in these experiments comes from eye movement data. The combination of the within-subject ORDER  $\times$  SENTENCE TYPE interactions and the between-subjects ORDER  $\times$  FREQUENCY indicate ongoing adaptation, just not to specific syntactic structures. However, the production results from the sentence completion task also play a highly informative role.

Syntactic Adaptation was intended to be the comprehension correlate of abstract priming in production. At least one of the advantages of having clearly identified priming in both production and comprehension was supposed to be a unified implicit learning mechanism across the whole processing system. This view was bolstered by the existence of comprehension-to-production priming, replicated in this study for both PO datives and ORCs. The fact that the structure in this experiment that shows the best evidence for some kind of adaptation, NP/Z  $\rightarrow$  Zs, is also the structure for which comprehension-to-production priming does not obtain, strongly suggests that the type of adaptation applied to Z garden paths is not related to syntactic priming.

Interestingly, ORCs do seem to be susceptible to true abstract priming, at least in production. While this is not the focus of these current experiments, to our knowledge comprehension-to-production priming with ORCs specifically is a novel finding. What makes this finding particularly intriguing is that the structures produced by our participants were not exact syntactic matches for the eyetracking primes, in that eyetracking sentences always had full Determiner + Noun Phrase structures and ORCs in the sentence completion responses universally used pronouns. Although this is a small difference, the fact that Gordon, Hendrick, and Johnson (2004) found that the ORC penalty disappears with pronominal ORCs indicates that it can be an important one. Like Bock and Loebell (1990)'s finding that locatives and PO datives prime each other, this may have implications for the level of detail in the structural representation relevant to abstract priming. We leave this for further research.

### *Conscious Strategies*

One possible interpretation of the NP/Z data in these experiments is that having the NP/Z - Comma conditions alongside the difficult NP/Z  $\rightarrow$  Z sentences gave participants a chance to realize that the NP/Z  $\rightarrow$  Z sentences were improved by a conscious strategy of mentally inserting a comma or similar. There is good evidence to suggest that this is the case for at least a sizeable portion of our participants. Furthermore, we would argue that this is exactly the point.

Evidence that this is a reasonable interpretation of the NP/Z within-experiment interaction comes from the informal debriefings that we did with participants in Experiment 4 when possible. When participants provided an answer (which they typically did), the majority of them mentioned encountering sentences that were particularly difficult. Approximately a third of the participants for whom we have responses semi-spontaneously mentioned commas.<sup>25</sup> Responses at this stage ranged from “the punctuation was in the wrong place”, to ‘some sentences weren’t very clear and seemed like they would have been better with a comma’, to specifically stating that they realized that they just needed to figure out where to put a comma in order to make the sentence comprehensible. Other descriptions that were telling include describing the difficult sentences as run-ons, which may also indicate a conscious, non-comma-based strategy of separating the clauses.

This strategy of learning to parse NP/Z garden paths using a conscious strategy, perhaps involving comma placement but perhaps focused on another way to separate the clauses, is exactly in line with our conclusions about our data. The argument that this chapter has put forward is that there is some kind of learning or ‘adaptation’ that occurs for NP/Zs, it is just that this is not *Syntactic Adaptation*, as it usually is understood. A conscious strategy that allows participants to make sense of otherwise very difficult sentences is a kind of learning which is outside of normal linguistic parsing, in this case *because* it is conscious.

#### *Timecourse*

Another feature of this data that fits adaptation to context difficulty better than Syntactic Adaptation is the timecourse of the effects of interest in the eye movement record. Syntactic Adaptation draws on a family of processing mechanisms heavily based on prediction. The competition that was supposed to drive the interaction between ORDER x SENTENCE TYPE was grounded on shifting expectation from the contextually rarer structure to the contextually frequent one (e.g., ORCs → SRCs). At the relative complementizer *that* the comprehender would make predictions about the identity of the subsequent incoming lexical item, and a portion of this prediction —the part about syntactic category —addresses how likely the next word is to be consistent with an ORC versus an SRC. Any processing difficulty then comes from surprise if the prediction is not con-

---

<sup>25</sup>Although this debriefing was informal (it was important that participants felt that the question was optional), efforts were made to standardize the questions and to avoid asking anything too leading. The first question that participants were asked was always “Was there anything that stood out to you or any sentences that you noticed in either in the sentence completion or the eyetracking?”. If the response to this was affirmative but vague, participants would be asked “What about that/[their own wording] stood out to you?”

firmed. Although expectation-based models typically refrain from making explicit claims about eye movements (with the notable exceptions of Bicknell & Levy, 2010, 2011), a predictive mechanism would very naturally imply that Adaptation would appear in eye movements as part of lexical recognition, including syntactic category identification. Lexical recognition is when the parser has the information to determine that a prediction has or has not been confirmed.

The eye movement record is quite complex and it is not typically practical to determine exactly when higher order processing effects (syntactic or semantic) “should” manifest. But there is good evidence that basic lexical identification occurs in the earliest measures. For some properties (e.g., frequency), measurable differences may even be detectable as parafoveal-on-foveal effects (Schotter, Angele, & Rayner, 2012). Readers are sensitive to other lexical properties by First Fixation and certainly by First Pass. And yet in the present results, all effects manifest very late in the eye movement record. In some cases, traces of what appears to be adaptation appear in Go Past, but the bulk of the effect and statistical significance never show up until Total Times (see e.g., Figure [5.6] and the corresponding inferential statistics). Given that fixations on the region of interest itself up to Go Past time are incorporated into Total Time<sup>26</sup>, this means that the driving measure of adaptation is most likely in re-reading. While re-reading is a part of natural, unconscious reading, it is also late enough in the eye record that more conscious strategies can begin to have an influence. Thus, adaptation this late in the eye movement record is somewhat uncomfortable for expectation-based theories, but it is entirely consistent with adaptation to context difficulty, both in the realization that some sentences may be tricky and in the adjustment to the level of understanding required by the comprehension questions and other demands of the task.

---

<sup>26</sup>Note that only part of Go Past is factored into Total Time, as Total Time does not include the fixations on prior regions that are due to regressions, but Go Past does.

## CHAPTER 6

### CONCLUSION

The central theme of this dissertation has been the interaction of syntactic structure with various forms of memory. The experiments in the thesis have investigated this relationship across three different timescales. Chapter 3 looked at the relationship between syntax and memory on very short time scales, within the processing of a single sentence, or alternately what might be seen as the transition from Focus of Attention to working memory. Chapter 2 investigated a time duration that was only slightly longer but covers the timespan traditionally associated with working/short-term memory<sup>1</sup> and the proposed transition period from working memory to long-term memory (LTM). The second half of the dissertation, Chapters 4 and 5, addressed syntax on a long-term memory scale, over the course of several sentences or whole experiments. The resulting picture captures the syntax-memory relationship as it develops in a comprehender's (or speaker's) experience.

Understanding the relationship of syntax and memory is critical both for theories in psycholinguistics and for theories of memory itself. Many theories in sentence processing invoke memory as an explanatory mechanism and as a crucial limiting factor on capabilities of the system. These range from theories indexing difficulty based on the number of open dependencies (e.g., Dependency Locality Theory, Gibson, 2000) to detailed theories of agreement errors based on interference within cue-based retrieval architectures (Wagers, 2008; Wagers et al., 2009). Indeed, Lewis and Vasishth (2005) largely captures the work of sentence comprehension within a domain general cue-based memory architecture. While memory theories have had considerable success providing explanatory accounts in sentence processing, at the very highest levels there are open questions about whether these models capture ways the cognitive architecture *could* work or the way it *does* work.

Memory theories themselves also have something to gain from an understanding of how to integrate syntax into memory. Syntax is a different type of object than previous work has focused

---

<sup>1</sup>There is partial overlap in the terms short-term memory and working, as well as terminological differences between theories (Baddeley, 1986; Cowan, 2008).

on because it is so inherently hierarchical. Current cue-based models of memory are best-suited to flat information structure, and this has been sufficient for much of the stimuli in memory studies. If syntax ultimately needs to be accommodated into memory at some level<sup>2</sup>, this would likely require a different base data structure, which could in turn make new predictions for how the system handles flat structures as well. Moreover, most ways of accommodating syntax into memory would require implicit memory well-beyond implicit/procedural memory for motor-planning. That's if syntax should be incorporated into memory structures. If it should not, then memory models face a host of additional questions, largely about defining what memory truly is and principled ways to determine what information it is responsible for. Common current formulations of what memory is do not consider the possibility that whole classes of information would be incompatible with it (Baddeley, 1986). Additionally, online sentence processing seems to be in need of some "backwards-looking" processes (Allentoff & Lorimor, 2014). Under current definitions a backwards looking cognitive process would have no place in the system other than being part of memory. The solution would need a principled definition of what kinds of backwards looking processes are and are not memory. Or perhaps the response is to proliferate many types of memory and say that syntax only interacts with one. Such a move would require criteria to distinguish one memory domain from another. Clearly, either incorporating syntax into memory or rejecting it has implications for how memory functions at a basic level and fits into the overall cognitive architecture.

Chapter 3 argued that the memory mechanism which is involved in very short term sentence processing is highly sensitive to syntactic dependency types, because it appears to use different memory retrieval strategies for agreement versus the binding relationship in Principle A reflexive anaphors.<sup>3</sup> This would suggest a very close relationship between memory and syntax at this time point, in which memory mechanisms have access to grammatical details and tailor retrieval to them. This is hardly what would be expected if syntax is "incompatible" with longer-term memory.

---

<sup>2</sup>Even if there is "no memory" for syntax during online-to-offline sentence processing, there may still be a need to adjust memory models for a speaker's stable knowledge of their language's grammar.

<sup>3</sup>There is a running debate about the extent to which memory is involved in attraction errors (Bock, Eberhard, Cutting, Meyer, & Schriefers, 2001; Eberhard, Cutting, & Bock, 2005; Wagers et al., 2009). However, the debate centers around whether memory mechanisms are *responsible* for the errors or whether something else causes the error. This is separate from whether memory mechanisms play *any* role in resolving these dependencies, as the competitor to the memory based account is still compatible with memory being used, just not be the mechanism that causes the error (Bock et al., 2001). While the clearest interpretation of the Chapter 3 results could be in a memory account of attraction errors, there are arguments from both main theories that attraction errors are "backward-looking", in that they look back to the subject upon reaching the verb (Allentoff & Lorimor, 2014).

Yet Chapter 2 provided evidence that syntactic memory decays equally quickly in grammatical tasks (ellipsis antecedent matching) as it does in an explicit sentence recognition task. This finding goes against an implicit or procedural memory account of syntactic storage and corroborates prior evidence that syntax does not survive the transition to long-term memory (Potter & Lombardi, 1990; Sachs, 1974).

The second half of the thesis addressed an issue in the abstract priming literature. Abstract priming is of interest because it has long been considered some of the best evidence for long-term implicit syntactic memory, but has the problem that it is not equally available in comprehension as it is in production. The lack of general availability is unexpected because most theories of priming draw on very general cognitive mechanisms that should be available by default. That is memory is generally available, so implicit memory is only a viable explanation for priming if it does not disappear 50% of the time without a clear reason. Chapter 4 suggested that despite the classic claim that abstract priming does not exist in comprehension, that there was evidence to support at least two theories that predict a limited distribution of abstract comprehension priming: a Surprisal-based implicit learning mechanism known as Syntactic Adaptation and a theory from Tooley and Bock (2014) that appealed to the increased role of the lexicon in shaping the process of comprehension, which in this dissertation was called the Lexically-Filtered Comprehension hypothesis. Chapter 4 ultimately concluded that the predictions of these theories were too overlapping and the evidence in the literature too unclear to definitively decide between the two using the pre-existing data. Therefore, Chapter 5 went on to provide a direct test of Syntactic Adaptation. Experiments 3 and 4 found the same facilitation over the course of a single experiment that had previously been used to argue for Syntactic Adaptation. However, across experiments it was found that frequency of the target construction did not impact the rate of facilitation, despite the fact that this was a key prediction of Adaptation. Based on these findings, Chapter 5 concluded that effects which had previously been called Syntactic Adaptation were better characterized as task adaptation within the experiment. This left Lexically-Filtered Comprehension as the best current theory of priming in comprehension.

While the results of the individual experiments in this dissertation are fully useful and interpretable at their own levels, together they still form a puzzling story. To see why, consider two possible worlds that could have been supported by these experiments:

First, consider a world in which syntactic structure and memory are utterly incompatible. Whatever type of object syntax is, it is never encoded or accessed by memory. In this world, it is expected that sentences or clauses which differed primarily on syntactic traits would be indistin-

guishable in memory, even for the purpose of computing syntactic dependencies such as ellipsis. This would rule out all forms of implicit memory as well, including priming and Adaptation. Additionally, if syntax and memory are incompatible, then memory operations should be blind to syntactic considerations, even during the brief period of online sentence processing when they might co-exist.

As it has been described, this world has a number of oddities. First and foremost, it is odd that there is a type of information that is incompatible with memory at all. Human memory is incredibly flexible in terms of the amount and type of information it can store, especially when one considers that there seem to be different types of memory which each specialize in their own domain, such as the split between declarative and motor-procedural memory or the phonological loop of Baddeley and Hitch (1974) and Baddeley (1986). To our knowledge, there is no other instance where an entire domain of information has been proposed to be incompatible with memory like this. To make matters worse, it also seems unlikely that all syntactic information is equally resistant to memory encoding, especially as some lexico-syntactic properties like syntactic category do not seem to encounter the same problems as hierarchical structure.

This world runs into the problem of how a speaker/comprehender stores their grammatical knowledge in the first place. The mere existence of stable grammatical knowledge within and across speakers would seem to require memory, under current definitions of what memory is. This world either necessitates an alternate explanation for how a stable grammar is maintained that does not involve memory or an explanation for how syntactic structure can be stored and accessed in long-term memory (and possibly very short-term memory as well if memory-based online processing models are to be believed, Lewis & Vasishth, 2005) but be unavailable for individual sentences that have just been processed.

A third, though lesser oddity is that syntax would have benefitted from a closer relationship with memory. Grammatical constraints such as the ellipsis antecedent conditions in Chapter 2 are a lot of complexity to specify, only to fail to apply them because the memory representation isn't accurate enough. If syntactic operations could have taken advantage of memory resources, they should have done so.

So while this World One is arguably the model that is supported by Chapter 2 of this dissertation, as well as Sachs (1974) and Potter and Lombardi's Reconstruction Hypothesis, it is not clear that the cognitive architecture that it suggests would be fully internally consistent when integrating with current models of the lexicon and long-term knowledge of grammatical rules. On the other hand, one impressive implication of this world is that syntax provides such an enormous

organizational benefit that it is worth it even if it can only be partially applied within the cognitive system.

World Two is a model at the other end of the spectrum, where syntactic structure is fully integrated with memory systems. This world is the one implied by default in commonly used memory models such as Lewis and Vasishth (2005), where all of sentence processing is modeled in a domain general memory architecture. As forgetting still applies, speakers will not be burdened by perfect recall of every sentence they had experienced, just as they are not burdened by perfect recall of other quotidian details. But reliable retrieval would almost certainly occur with greater frequency and over longer timescales. When reliable retrieval did occur, syntactic structure would be better retrieved for its own sake and would serve as a further cue for more successful retrieval of lexical and semantic information as well. Moreover, the parser would have no trouble implementing syntactic constraints at a distance, such as cross-clausal ellipsis or Wh-dependencies, making parsing less error-prone. Memory mechanisms would be expected to have access to details of grammatical operations and the ability to tailor retrieval patterns to match the needs of specific dependencies/constructions for more efficient online parsing. In short, World Two seems to have a number of advantages above and beyond those in our own world.

*A priori*, World Two may be more easily made internally consistent than World One, thanks to the ability to adopt domain general memory mechanisms. Any anomalies present in World Two are largely dependent on whether syntactic memory is integrated into conscious, declarative memory or a type of implicit memory. If syntactic memory for sentences makes use of domain general declarative memory, then all forms of non-conscious syntactic memory, such as priming or a speaker's long-term grammatical knowledge, become uncomfortable parts of the system because they should have been available as explicit memory. It is unclear whether they would persist in the forms we know them in.<sup>4</sup> Implicit memory implementations of World Two engender questions about what kind of implicit memory is used and more importantly how many kinds of there are to choose from. There may be a proliferation of types of implicit memory in such a system, or syntactic memory might be one of very few. Both declarative and implicit versions of this world would need to integrate the extremely hierarchical nature of syntax into memory encoding abilities. Unlike world One, most of these concerns are not true world-internal inconsistencies,

---

<sup>4</sup>Though see Reitter, Keller, and Moore (2011)'s implementation of priming in an ACT-R framework as a possible way these issues could go.

but rather statements about how cognitive science's understanding would need to advance in order to achieve an explanatory model of the cognitive architecture.

Slight oddities aside, both of these worlds would be relatively internally consistent and reasonable ways for the actual world to work. However, the evidence in this thesis is split and therefore not fully consistent with either one. The ellipsis processing results in Chapter 2 sided firmly with previous work from verbatim memory in arguing for that syntactic structure decays from memory quite quickly ( $< 1$  minute). It also argued against the possibility of appealing to procedural memory as a work around. Additionally, the combined results of Experiments 3 and 4 rule out Adaptation, which would have been a clear form of syntactic bookkeeping in memory, and a possible way to save abstract priming in comprehension. Both of these results are very much in keeping with World One and the incompatibility of syntax and memory. On the other hand, Chapter 3's suggestion that sentence-internal memory is sensitive to grammar-internal dependencies is at odds with the supposed incompatibility. Furthermore, the combined conclusion of Chapters 4 and 5 is that comprehension priming likely does exist, just that it is better captured by Lexically-Filtered Comprehension than Adaptation. The resulting state-of-affairs is that neither of the most obvious and internally-logical worlds are available to be the world that *we* live in.

## APPENDIX

### LOG-TRANSFORMED READING TIME ANALYSIS FROM CHAPTER 3

#### *Agreement Results*

Raw results for agreement are given in Table (A); models are given in Table (A.2).

In First Pass, there was a main effect of grammaticality at the verb region ( $\beta=-0.09$ ,  $SE=0.03$ ,  $t=-3.15$ ), but no other effects were significant in either the verb region or the spillover.

#### **Mean RTs for the Agreement Conditions**

	<i>First Pass</i>	<i>Go Past</i>	<i>Regressions Out</i>	<i>Total Time</i>
<i>Predicate Region</i>				
Agreement-Gram-Intrusion	446 (19)	606 (28)	0.205	692 (28)
Agreement-Gram-NoIntrusion	447 (20)	578 (32)	0.142	698 (36)
Agreement-Ung-Intrusion	480 (21)	694 (42)	0.216	765 (33)
Agreement-Ung-NoIntrusion	517 (24)	850 (50)	0.289	890 (45)
<i>Spillover Region</i>				
Agreement-Gram-Intrusion	469 (17)	624 (27)	0.166	673 (26)
Agreement-Gram-NoIntrusion	467 (18)	613 (37)	0.159	698 (32)
Agreement-Ung-Intrusion	448 (16)	616 (29)	0.188	670 (28)
Agreement-Ung-NoIntrusion	461 (15)	738 (43)	0.228	734 (31)

**Table A.1.** Mean RTs at the Critical Region for subject verb agreement (the verb region). Standard errors are given in parentheses.

In Go Past at the verb region, there were significant main effects of both grammaticality ( $\beta=-0.19$ ,  $SE=0.03$ ,  $t=-5.60$ ) and lure number ( $\beta=-0.10$ ,  $SE=0.03$ ,  $t=-2.97$ ) but the interaction did not reach significance. In the spillover region, lure number was marginally significant ( $\beta=-0.06$ ,  $SE=0.03$ ,  $t=-1.79$ ), but nothing else approached significance.

Finally, in Total Times at the verb region, both main effects of grammaticality ( $\beta=-0.16$ ,  $SE=0.03$ ,  $t=-5.69$ ) and lure number ( $\beta=0.08$ ,  $SE=0.03$ ,  $t=2.81$ ) were significant, and furthermore the interaction was also significant ( $\beta=0.13$ ,  $SE=0.07$ ,  $t=2.01$ )\*. In the spillover region, neither main effect reached significance but the interaction was ( $\beta=0.12$ ,  $SE=0.05$ ,  $t=2.18$ ).

\*Appears non-significant due to rounding

### Agreement Model Summaries

	Intercept	Grammaticality	Lure Number	Grammaticality x Lure Number
<i>Verb Region</i>				
First Pass $\beta$	6.01(0.04)	-0.09(0.03)	-0.03(0.03)	0.07(0.06)
Go Past $\beta$	6.29(0.04)	-0.19(0.03)	-0.10(0.03)	0.09(0.07)
Total Time $\beta$	6.46(0.04)	-0.16(0.03)	-0.08(0.03)	0.13(0.07)*
<i>Spillover Region</i>				
First Pass $\beta$	6.01(0.03)	0.04(0.03)	-0.00(0.03)	0.03(0.05)
Go Past $\beta$	6.28(0.04)	-0.03(0.03)	-0.06(0.03)	0.09(0.06)
Total Time $\beta$	6.39(0.04)	-0.01(0.03)	-0.03(0.03)	0.12(0.05)

**Table A.2.** Summary of the Linear Models applied to the Agreement results as  $\beta$  coefficients. Standard Errors are in parentheses.  $t$ -values can be obtained by dividing the  $\beta$  value by the associated Standard Error. With the exception of intercepts, dark gray shading indicates significance and light gray shading indicates a marginal effect.

### Reflexive Anaphor Results

Raw reading times for the reflexive conditions are given in Table (A) and reflexive model summaries are given in Table (A.4).

In First Pass there were no significant effects in either region.

In Go Past, there were no significant effects in the reflexive region, but in the spillover there was a marginal effect of grammaticality ( $\beta=-0.06$ ,  $SE=0.03$ ,  $t=-1.84$ ) and a significant effect of lure number ( $\beta=-0.06$ ,  $SE=0.03$ ,  $t=-2.04$ ).

Total Times found a significant main effect of grammaticality in both the reflexive region ( $\beta=-0.09$ ,  $SE=0.03$ ,  $t=-3.27$ ) and the spillover ( $\beta=-0.07$ ,  $SE=0.03$ ,  $t=-2.85$ ), but no other significant factors.

### Mean RTs for the Reflexive Conditions

	<i>First Pass</i>	<i>Go Past</i>	<i>Regressions Out</i>	<i>Total Time</i>
<i>Reflexive Region</i>				
Reflexive-Gram-Intrusion	264 (9)	340 (16)	0.124	381 (18)
Reflexive-Gram-NoIntrusion	259 (8)	332 (18)	0.132	365 (16)
Reflexive-Ung-Intrusion	273 (11)	385 (32)	0.134	427 (27)
Reflexive-Ung-NoIntrusion	272 (8)	364 (21)	0.135	422 (22)
<i>Spillover Region</i>				
Reflexive-Gram-Intrusion	469 (20)	562 (27)	0.114	640 (27)
Reflexive-Gram-NoIntrusion	454 (19)	551 (35)	0.089	639 (30)
Reflexive-Ung-Intrusion	459 (19)	568 (30)	0.109	687 (32)
Reflexive-Ung-NoIntrusion	479 (18)	645 (46)	0.132	705 (35)

**Table A.3.** Mean reading times at the Critical Region for reflexive anaphors. Standard errors are given in parentheses.

Notably, the critical interaction of GRAMMATICALITY X LURE NUMBER was not significant or marginal for any comparison in the reflexive analyses. Even though none of the signature tests for intrusion reach significance, there is still substantial evidence in this data that reflexive intrusion might still exist. Every one of the analyzed measures at the spillover region has a clear trend for the *ungrammatical-intrusion* condition to be noticeably facilitated relative the *ungrammatical-no\_intrusion* condition.

#### Reflexive Model Summaries

	Intercept	Grammaticality	Lure Number	Grammaticality x Lure Number
<i>Reflexive Region</i>				
First Pass $\beta$	5.51(0.02)	-0.03(0.02)	-0.01(0.02)	-0.02(0.04)
Go Past $\beta$	5.67(0.03)	-0.05(0.03)	-0.01(0.03)	-0.02(0.06)
Total Time $\beta$	5.81(0.04)	-0.09(0.03)	-0.01(0.03)	-0.04(0.06)
<i>Spillover Region</i>				
First Pass $\beta$	6.02(0.04)	-0.02(0.02)	-0.03(0.02)	0.02(0.05)
Go Past $\beta$	6.18(0.04)	-0.06(0.03)	-0.06(0.03)	0.04(0.06)
Total Time $\beta$	6.34 (0.05)	-0.07(0.03)	-0.02 (0.03)	0.01(0.05)

**Table A.4.** Summary of the Linear Models applied to the Reflexive results as  $\beta$  coefficients. Standard Errors are in parentheses.  $t$ -values can be obtained by dividing the  $\beta$  value by the associated Standard Error. With the exception of intercepts, dark gray shading indicates significance and light gray shading indicates a marginal effect.

#### Aggregate Results

A key prediction of Dillon et al. (2013) was not just that reflexive intrusion effects shouldn't hold for the within-dependency comparisons, but that there should be a three-way interaction confirming the lack of intrusion for anaphors relative to agreement. This section reports the 2x2x2 models for each of our reported measures.

In First Pass times at the critical region, there was a significant main effect of dependency ( $\beta=0.516$ ,  $SE=0.034$ ,  $t=14.73$ ). This effect is unsurprising, since the critical region for the two dependencies was different, and on average the critical verb region for agreement was longer than the critical reflexives region. Because the interpretation of the dependency main effect is not central to the interpretation of the results, we will leave off reporting it for the remaining measures, although it will appear in the model summaries given in the tables. There was also a significant main effect of grammaticality ( $\beta=-0.057$ ,  $SE=0.019$ ,  $t=-3.07$ ). No other effects in First Pass at the

critical region were significant, although there was a marginal interaction of dependency x grammaticality ( $\beta=-0.061$ ,  $SE=0.033$ ,  $t=-1.87$ ). This may be an indication of an overall greater penalty for unacceptability for agreement in early measures, but again, interpretation of this particular effect is confounded by the unmatched-critical regions.

In Go Past at the critical region, all three main effect were significant: dependency ( $\beta=0.631$ ,  $SE=0.035$ ,  $t=17.44$ ), grammaticality ( $\beta=-0.119$ ,  $SE=0.024$ ,  $t=-4.97$ ), intrusion ( $\beta=-0.056$ ,  $SE=0.023$ ,  $t=-2.45$ ). Furthermore the dependency x grammaticality interaction was significant ( $\beta=-0.149$ ,  $SE=0.051$ ,  $t=-2.92$ ) and the interaction of dependency x intrusion was marginal ( $\beta=-0.096$ ,  $SE=0.053$ ,  $t=-1.81$ ). The three-way interaction did not reach significance ( $\beta=0.114$ ,  $SE=0.097$ ,  $t=1.17$ ).

The Total Time model also found significant effects for all main effects (dependency ( $\beta=0.651$ ,  $SE=0.031$ ,  $t=20.91$ ; grammaticality  $\beta=-0.124$ ,  $SE=0.019$ ,  $t=-6.32$ ; intrusion  $\beta=-0.048$ ,  $SE=0.019$ ,  $t=-2.42$ ). None of the interactions attained significance, including the three-way interaction ( $\beta=0.164$ ,  $SE=0.096$ ,  $t=1.72$ ).

There were no significant effects in First Pass at the spillover, but in Go Past, all three of the main effects were significant (Grammaticality:  $\beta=-0.047$ ,  $SE=0.021$ ,  $t=-2.24$ ; Intruder Number:  $\beta=-0.060$ ,  $SE=0.020$ ,  $t=-2.99$ ). The three-way interaction, however, did not reach significance ( $\beta=0.046$ ,  $SE=0.080$ ,  $t=0.58$ ).

Finally in Total Times at the spillover, there was a significant main effect of grammaticality ( $\beta=-0.043$ ,  $SE=0.021$ ,  $t=-2.06$ ), but again the three-way interaction was not significant ( $\beta=0.105$ ,  $SE=0.075$ ,  $t=1.41$ ).

## REFERENCES

- Adams, B. C., Clifton, C., & Mitchell, D. C. (1998). Lexical guidance in sentence processing? *Psychonomic Bulletin & Review*, 5(2), 265–270.
- Allentoff, A., & Lorimor, H. (2014). *Delaying verb production changes what matters in subject-verb agreement*. (Poster presented at the 27th annual CUNY Human Sentence Processing Conference, Columbus, Ohio)
- Anderson, M. C., & Neely, J. H. (1996). Interference and inhibition in memory retrieval. *Memory*, 22, 586.
- Andrews, C., Staub, A., & Dillon, B. (2017). *Syntactic adaptation vs task adaptation: The case of object relative clauses*. (Poster presented at Architectures and Mechanisms of Language Processing (AMLaP), Lancaster, UK)
- Apel, J., Knoeferle, P., & Crocker, M. W. (2007). Processing parallel structure: Evidence from eye tracking and a computational model. In S. Vosniadou, D. Kayser, & A. Protopapas (Eds.), *Proceedings of the 2nd European Cognitive Science Conference* (pp. 125–131).
- Arai, M. (2012). What can head-final languages tell us about syntactic priming (and vice versa)? *Language and Linguistics Compass*, 6(9), 545–559.
- Arai, M., & Mazuka, R. (2010). *Syntactic priming as an index of children's syntactic knowledge: evidence from visual world eye-tracking study*. (Poster presented at the CUNY Human Sentence Processing Conference, New York)
- Arai, M., & Mazuka, R. (2014). The development of Japanese passive syntax as indexed by structural priming in comprehension. *The Quarterly Journal of Experimental Psychology*, 67(1), 60–78.
- Arai, M., Nakamura, C., & Mazuka, R. (2011). *Predicting a dispreferred structural alternative as a result of syntactic priming in comprehension*. (Poster presented at the annual conference on Architectures and Mechanisms for Language Processing (AMLaP), Paris, France)
- Arai, M., Nakamura, C., & Mazuka, R. (2015). Predicting the unbeaten path through syntactic priming. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 41(2), 482.
- Arai, M., van Gompel, R. P., & Scheepers, C. (2007). Priming ditransitive structures in comprehension. *Cognitive Psychology*, 54(3), 218–250.
- Arregui, A., Clifton, C., Frazier, L., & Moulton, K. (2006). Processing elided verb phrases with flawed antecedents: The recycling hypothesis. *Journal of memory and language*, 55(2), 232–246.
- Baddeley, A. D. (1986). *Working memory*. Oxford: Clarendon Press.
- Baddeley, A. D., & Hitch, G. (1974). Working memory. In *Psychology of learning and motivation* (Vol. 8, pp. 47–89). Elsevier.
- Bader, M., & Lasser, I. (1994). German verb-final clauses and sentence processing: Evidence for immediate attachment. In C. Clifton, L. Frazier, & K. Rayer (Eds.), *Perspectives on sentence processing* (pp. 225–242). Hillsdale, NJ: Lawrence Erlbaum Associates Inc.
- Bernolet, S., Hartsuiker, R. J., & Pickering, M. J. (2012). Effects of phonological feedback on the selection of syntax: Evidence from between-language syntactic priming. *Bilingualism: Language and Cognition*, 15(3), 503–516.
- Bever, T. G. (1970). The cognitive basis for linguistic structures. In J. R. Hayes (Ed.), *Cognition and the development of language* (pp. 279–362). New York: Wiley and Sons.
- Bicknell, K., & Levy, R. (2010). A rational model of eye movement control in reading. In *Proceedings of the 48th annual meeting of the Association for Computational Linguistics* (pp. 1168–1178).
- Bicknell, K., & Levy, R. (2011). Why readers regress to previous words: A statistical analysis. In *Proceedings of the annual meeting of the Cognitive Science Society* (Vol. 33).

- Bock, J. K. (1986). Syntactic persistence in language production. *Cognitive Psychology*, 18(3), 355–387.
- Bock, J. K., Dell, G., Chang, F., & Onishi, K. H. (2007). Persistent structural priming from language comprehension to language production. *Cognition*, 104(3), 437–458.
- Bock, J. K., Eberhard, K. M., Cutting, J. C., Meyer, A. S., & Schriefers, H. (2001). Some attractions of verb agreement. *Cognitive Psychology*, 43(2), 83–128.
- Bock, J. K., & Griffin, Z. M. (2000). The persistence of structural priming: Transient activation or implicit learning? *Journal of Experimental Psychology: General*, 129(2), 177.
- Bock, J. K., & Levelt, W. J. (1994). *Language production: Grammatical encoding*. Academic Press.
- Bock, J. K., & Loebell, H. (1990). Framing sentences. *Cognition*, 35(1), 1–39.
- Bock, J. K., & Miller, C. A. (1991). Broken agreement. *Cognitive Psychology*, 23(1), 45–93.
- Branigan, H. P., & McLean, J. F. (2016). What children learn from adults' utterances: An ephemeral lexical boost and persistent syntactic priming in adult-child dialogue. *Journal of Memory and Language*, 91, 141–157.
- Branigan, H. P., Pickering, M. J., & Cleland, A. A. (1999). Syntactic priming in written production: Evidence for rapid decay. *Psychonomic Bulletin & Review*, 6(4), 635–640.
- Branigan, H. P., Pickering, M. J., & McLean, J. F. (2005). Priming prepositional-phrase attachment during comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31(3), 468.
- Bresnan, J., Cueni, A., Nikitina, T., & Baayen, R. H. (2007). Predicting the dative alternation. *Cognitive foundations of interpretation*, 69–94.
- Brooks, S. P., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7(4), 434–455.
- Bürkner, P.-C. (2018). Advanced Bayesian multilevel modeling with the R package brms. *The R Journal*, 10(1), 395–411.
- Carminati, M. N., van Gompel, R. P., Scheepers, C., & Arai, M. (2008). Syntactic priming in comprehension: The role of argument order and animacy. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34(5), 1098.
- Chang, F., Dell, G., & Bock, J. K. (2006). Becoming syntactic. *Psychological Review*, 113(2), 234.
- Chen, X.-j., & Kwak, Y. (2017). What makes you go faster?: The effect of reward on speeded action under risk. *Frontiers in psychology*, 8, 1057.
- Clark, S. E., & Gronlund, S. D. (1996). Global matching models of recognition memory: How the models match the data. *Psychonomic Bulletin & Review*, 3(1), 37–60.
- Cowan, N. (2008). What are the differences between long-term, short-term, and working memory? *Progress in Brain Research*, 169, 323–338.
- Cuetos, F., Mitchell, D. C., & Coreley, M. M. B. (1996). Parsing in different languages. In M. Carreiras, J. Garcia-Albea, & N. Sebastián-Gallés (Eds.), *Language processing in spanish* (pp. 145–187). Hillsdale, NJ: Erlbaum.
- Dempsey, J., Liu, Q., & Christianson, K. (2020). Convergent probabilistic cues do not trigger syntactic adaptation: Evidence from self-paced reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.
- Dillon, B. (2011). *Structured access in sentence comprehension* (Unpublished doctoral dissertation). University of Maryland.
- Dillon, B., Mishler, A., Sloggett, S., & Phillips, C. (2013). Contrasting intrusion profiles for agreement and anaphora: Experimental and modeling evidence. *Journal of Memory and Language*, 69(2), 85–103.
- Dillon, B., & Wagers, M. (2019). Approaching gradience in acceptability with the tools of signal detection theory. *OSF Preprints*.
- Do, M. L., & Kaiser, E. (2017). The relationship between syntactic satiation and syntactic priming: A first look. *Frontiers in Psychology*, 8, 1851.
- Dommergues, J.-Y., & Grosjean, F. (1981). Performance structures in the recall of sentences. *Memory & Cognition*, 9(5), 478–486.
- Eberhard, K. M., Cutting, J. C., & Bock, J. K. (2005). Making syntax of sense: number agreement in sentence production. *Psychological review*, 112(3), 531.

- Engelmann, F., Jäger, L. A., & Vasishth, S. (2015). The determinants of retrieval interference in dependency resolution: Review and computational modeling. *Manuscript submitted*.
- Ferreira, F., & Henderson, J. M. (1991). Recovery from misanalyses of garden-path sentences. *Journal of Memory and Language*, 30(6), 725–745.
- Ferreira, V. (2003). The persistence of optional complementizer production: Why saying “that” is not saying “that” at all. *Journal of Memory and Language*, 48(2), 379–398.
- Fine, A. B., Jaeger, T. F., Farmer, T. A., & Qian, T. (2013). Rapid expectation adaptation during syntactic comprehension. *PloS One*, 8(10).
- Fodor, J. A. (1983). *The modularity of mind*. MIT Press/Bradford Books.
- Fodor, J. A., Bever, T. G., & Garrett, M. F. (1974). *The psychology of language: an introduction to psycholinguistics and generative grammar*. New York; Montreal: McGraw-Hill.
- Frazier, L., & Clifton, C. (1998). Sentence reanalysis, and visibility. In J. D. Fodor & F. Ferreira (Eds.), *Reanalysis in sentence processing* (pp. 143–176).
- Frazier, L., Munn, A., & Clifton, C. (2000). Processing coordinate structures. *Journal of Psycholinguistic Research*, 29(4), 343–370.
- 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.
- Garnham, A., & Oakhill, J. (1987). Interpreting elliptical verb phrases. *The Quarterly Journal of Experimental Psychology*, 39(4), 611–627.
- Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models* (Vol. 1). Cambridge University Press New York, NY, USA.
- Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4), 457–472.
- Gennari, S. P., Mirković, J., & MacDonald, M. C. (2012). Animacy and competition in relative clause production: A cross-linguistic investigation. *Cognitive Psychology*, 65(2), 141–176.
- Gibson, E. (2000). The dependency locality theory: A distance-based theory of linguistic complexity. In A. Marantz, Y. Miyashita, & W. O’Neil (Eds.), *Image, language, brain* (Vol. 2000, pp. 95–126).
- Glenberg, A. M. (1976). Monotonic and nonmonotonic lag effects in paired-associate and recognition memory paradigms. *Journal of Verbal Learning and Verbal Behavior*, 15(1), 1–16.
- Gordon, P. C., Hendrick, R., & Johnson, M. (2001). Memory interference during language processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 27(6), 1411.
- Gordon, P. C., Hendrick, R., & Johnson, M. (2004). Effects of noun phrase type on sentence complexity. *Journal of memory and Language*, 51(1), 97–114.
- Gries, S. (2003). New perspectives on old alternations. In *Proceedings from the 39th annual meeting of the Chicago Linguistic Society* (Vol. 2, pp. 311–329).
- Gries, S. (2005). Syntactic priming: A corpus-based approach. *Journal of Psycholinguistic Research*, 34(4), 365–399.
- Grosjean, F., Grosjean, L., & Lane, H. (1979). The patterns of silence: Performance structures in sentence production. *Cognitive Psychology*, 11(1), 58–81.
- Gurevich, O., Johnson, M. A., & Goldberg, A. E. (2010). Incidental verbatim memory for language. *Language and Cognition*, 2(1), 45–78.
- Hale, J. (2001). A probabilistic earley parser as a psycholinguistic model. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies* (pp. 1–8).
- Hammerly, C., Staub, A., & Dillon, B. (2019). The grammaticality asymmetry in agreement attraction reflects response bias: Experimental and modeling evidence. *Cognitive Psychology*, 110, 70–104.
- Hardt, D. (1993). *Verb phrase ellipsis: Form, meaning, and processing* (Unpublished doctoral dissertation). University of Pennsylvania.
- Harrington Stack, C. M., James, A. N., & Watson, D. G. (2018, Aug 01). A failure to replicate rapid syntactic adaptation in comprehension. *Memory & Cognition*, 46(6), 864–877.

- Hartsuiker, R. J., Bernolet, S., Schoonbaert, S., Speybroeck, S., & Vanderelst, D. (2008). Syntactic priming persists while the lexical boost decays: Evidence from written and spoken dialogue. *Journal of Memory and Language*, 58(2), 214–238.
- Hartsuiker, R. J., Kolk, H. H., & Huiskamp, P. (1999). Priming word order in sentence production. *The Quarterly Journal of Experimental Psychology Section A*, 52(1), 129–147.
- Hartsuiker, R. J., Pickering, M. J., & Veltkamp, E. (2004). Is syntax separate or shared between languages? Cross-linguistic syntactic priming in Spanish-English bilinguals. *Psychological Science*, 15(6), 409–414.
- Hartsuiker, R. J., & Westenberg, C. (2000). Word order priming in written and spoken sentence production. *Cognition*, 75(2), B27–B39.
- Heitz, R. P. (2014). The speed-accuracy tradeoff: history, physiology, methodology, and behavior. *Frontiers in Neuroscience*, 8, 150.
- Hoffman, M. D., & Gelman, A. (2014). The No-U-Turn Sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15(1), 1593–1623.
- Iacobucci, D., Schneider, M. J., Popovich, D. L., & Bakamitsos, G. A. (2016). Mean centering helps alleviate “micro” but not “macro” multicollinearity. *Behavior Research Methods*, 48(4), 1308–1317.
- Jarvella, R. J. (1971). Syntactic processing of connected speech. *Journal of Verbal Learning and Verbal Behavior*, 10(4), 409–416.
- Johnson, K. (2001). What VP ellipsis can do, and what it can't, but not why. In M. Baltin & C. Collins (Eds.), *The handbook of contemporary syntactic theory* (p. 439-479). Blackwell Publishers.
- Johnson, N. F. (1965). The psychological reality of phrase-structure rules. *Journal of Verbal Learning and Verbal Behavior*, 4(6), 469–475.
- Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological review*, 87(4), 329.
- Jäger, L., Mertzen, D., Van Dyke, J., & Vasishth, S. (2018). *Contrasting facilitation profiles for agreement and reflexives revisited: A large-scale empirical evaluation of the cue-based retrieval model*. (Poster presented at Architectures and Mechanisms of Language Processing (AMLaP): Berlin, Germany)
- Kahana, M. J. (2012). *Foundations of human memory*. OUP USA.
- Kamide, Y., & Mitchell, D. C. (1999). Incremental pre-head attachment in Japanese parsing. *Language and Cognitive Processes*, 14(5-6), 631–662.
- Kaschak, M. P., & Glenberg, A. M. (2004). This construction needs learned. *Journal of Experimental Psychology: General*, 133(3), 450.
- Kehler, A. (2002). *Coherence, reference, and the theory of grammar*. Stanford, CA: CSLI publications.
- Kennison, S. M., & Trofe, J. L. (2003). Comprehending pronouns: A role for word-specific gender stereotype information. *Journal of Psycholinguistic Research*, 32(3), 355–378.
- Kim, C. S., Carbary, K. M., & Tanenhaus, M. K. (2014). Syntactic priming without lexical overlap in reading comprehension. *Language and Speech*, 57(2), 181–195.
- King, J., Andrews, C., & Wagers, M. (2012). *Do reflexives always find a grammatical antecedent for themselves?* (Poster presented at the 25th annual CUNY Human Sentence Processing Conference, New York, NY)
- King, J., & Just, M. A. (1991). Individual differences in syntactic processing: The role of working memory. *Journal of Memory and Language*, 30(5), 580–602.
- Kirchner, W. K. (1958). Age differences in short-term retention of rapidly changing information. *Journal of Experimental Psychology*, 55(4), 352.
- Kreiner, H., Sturt, P., & Garrod, S. (2008). Processing definitional and stereotypical gender in reference resolution: Evidence from eye-movements. *Journal of Memory and Language*, 58(2), 239–261.
- Kush, D. W. (2013). *Respecting relations: Memory access and antecedent retrieval in incremental sentence processing* (Unpublished doctoral dissertation). University of Maryland.
- Lau, E., Wagers, M., Stroud, C., & Phillips, C. (2008). *Agreement and the subject of confusion*. (Talk given at 21st annual CUNY Human Sentence Processing Conference. Chapel Hill, NC)

- Ledoux, K., Traxler, M. J., & Swaab, T. Y. (2007). Syntactic priming in comprehension: Evidence from event-related potentials. *Psychological Science*, *18*(2), 135–143.
- Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177.
- Lewis, R. L. (1996). Interference in short-term memory: The magical number two (or three) in sentence processing. *Journal of Psycholinguistic Research*, *25*(1), 93–115.
- Lewis, R. L., & Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science*, *29*(3), 375–419.
- Lombardi, L., & Potter, M. C. (1992). The regeneration of syntax in short term memory. *Journal of Memory and Language*, *31*(6), 713–733.
- Luka, B. J., & Barsalou, L. W. (2005). Structural facilitation: Mere exposure effects for grammatical acceptability as evidence for syntactic priming in comprehension. *Journal of Memory and Language*, *52*(3), 436–459.
- Luka, B. J., & Choi, H. (2012). Dynamic grammar in adults: Incidental learning of natural syntactic structures extends over 48 hours. *Journal of Memory and Language*, *66*(2), 345–360.
- MacDonald, M., & Montag, J. (2009). Word order doesn't matter: relative clause production in English and Japanese. In *Proceedings of the annual meeting of the Cognitive Science Society* (Vol. 31).
- Macmillan, N. A., & Creelman, C. D. (2004). *Detection theory: A user's guide*. Psychology press.
- Mahowald, K., James, A., Futrell, R., & Gibson, E. (2016). A meta-analysis of syntactic priming in language production. *Journal of Memory and Language*, *91*, 5–27.
- McCann, D., & Kaan, E. (2017). *Non-adaptation to garden paths*. (Poster presented at the CUNY Human Sentence Processing Conference, Boston, MIT)
- McElree, B. (2000). Sentence comprehension is mediated by content-addressable memory structures. *Journal of Psycholinguistic Research*, *29*(2), 111–123.
- McElree, B. (2006). Accessing recent events. In *The Psychology of Learning and Motivation* (Vol. 46, pp. 155–200). Elsevier Academic Press.
- McElree, B., & Doshier, B. A. (1989). Serial position and set size in short-term memory: The time course of recognition. *Journal of Experimental Psychology: General*, *118*(4), 346.
- McElree, B., Foraker, S., & Dyer, L. (2003). Memory structures that subserve sentence comprehension. *Journal of Memory and Language*, *48*(1), 67–91.
- McLean, J. F., Pickering, M. J., & Branigan, H. P. (2004). Lexical repetition and syntactic priming in dialogue. In *Approaches to studying world-situated language use: bridging the language-as-product and language-as-action traditions*. MIT Press.
- Mehler, J. (1963). Some effects of grammatical transformations on the recall of English sentences. *Journal of Verbal Learning and Verbal Behavior*, *2*(4), 346–351.
- Miller, G. A., & Isard, S. (1963). Some perceptual consequences of linguistic rules. *Journal of Verbal Learning and Verbal Behavior*, *2*(3), 217–228.
- Miller, G. A., & Isard, S. (1964). Free recall of self-embedded English sentences. *Information and Control*, *7*(3), 292–303.
- Mitchell, D. C., & Green, D. W. (1978). The effects of context and content on immediate processing in reading. *Quarterly Journal of Experimental Psychology*, *30*(4), 609–636.
- Munn, A. B. (1993). *Topics in the syntax and semantics of coordinate structures* (Unpublished doctoral dissertation). University of Maryland.
- Nicol, J. L. (1988). *Coreference processing during sentence comprehension* (Unpublished doctoral dissertation). Massachusetts Institute of Technology.
- Nicol, J. L., & Swinney, D. (1989). The role of structure in coreference assignment during sentence comprehension. *Journal of Psycholinguistic Research*, *18*(1), 5–19.
- Orne, M. T. (1962). On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American Psychologist*, *17*(11), 776.
- Osborne, J. W. (2010). Improving your data transformations: Applying the box-cox transformation. *Practical Assessment, Research & Evaluation*, *15*(12), 1–9.
- Parker, D., & Phillips, C. (2017). Reflexive attraction in comprehension is selective. *Journal of Memory and Language*, *94*, 272–290.

- Pazzaglia, A. M., Dube, C., & Rotello, C. M. (2013). A critical comparison of discrete-state and continuous models of recognition memory: Implications for recognition and beyond. *Psychological Bulletin*, 139(6), 1173.
- Pickering, M. J., & Branigan, H. P. (1998). The representation of verbs: Evidence from syntactic priming in language production. *Journal of Memory and Language*, 39(4), 633–651.
- Pickering, M. J., & Ferreira, V. (2008). Structural priming: A critical review. *Psychological Bulletin*, 134(3), 427.
- Pickering, M. J., McLean, J. F., & Branigan, H. P. (2013). Persistent structural priming and frequency effects during comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 39(3), 890.
- Potter, M. C., & Lombardi, L. (1990). Regeneration in the short-term recall of sentences. *Journal of Memory and Language*, 29(6), 633–654.
- Potter, M. C., & Lombardi, L. (1998). Syntactic priming in immediate recall of sentences. *Journal of Memory and Language*, 38(3), 265–282.
- Potts, G. R., Keenan, J. M., & Golding, J. M. (1988). Assessing the occurrence of elaborative inferences: Lexical decision versus naming. *Journal of Memory and Language*, 27(4), 399–415.
- Prasad, G., & Linzen, T. (2020). Rapid syntactic adaptation in self-paced reading: detectable, but requires many participants. *PsyArXiv Preprints*.
- Ratcliff, R., Sheu, C.-F., & Gronlund, S. D. (1992). Testing global memory models using roc curves. *Psychological Review*, 99(3), 518.
- Rayner, K., Pollatsek, A., Ashby, J., & Clifton, C. (2012). *Psychology of reading*. Psychology Press.
- Reitter, D., Keller, F., & Moore, J. D. (2011). A computational cognitive model of syntactic priming. *Cognitive Science*, 35(4), 587–637.
- Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., & Müller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, 12(1), 1471–2105.
- Rosenthal, R., & Rosnow, R. L. (1969/2009). *Artifacts in behavioral research: Robert Rosenthal and Ralph L. Rosnow's classic books*. Oxford University Press.
- Rowland, C. F., Chang, F., Ambridge, B., Pine, J. M., & Lieven, E. V. (2012). The development of abstract syntax: Evidence from structural priming and the lexical boost. *Cognition*, 125(1), 49–63.
- Sachs, J. S. (1967). Recognition memory for syntactic and semantic aspects of connected discourse. *Attention, Perception, & Psychophysics*, 2(9), 437–442.
- Sachs, J. S. (1974). Memory in reading and listening to discourse. *Memory & Cognition*, 2(1), 95–100.
- Scheepers, C. (2003). Syntactic priming of relative clause attachments: Persistence of structural configuration in sentence production. *Cognition*, 89(3), 179–205.
- Scheepers, C., & Crocker, M. W. (2004). Constituent order priming from reading to listening: A visual-world study. In M. Carreiras & C. Clifton (Eds.), *The on-line study of sentence comprehension: Eyetracking, ERPs, and beyond* (pp. 167–185). New York: Psychology Press.
- Scheepers, C., Raffray, C. N., & Myachykov, A. (2017). The lexical boost effect is not diagnostic of lexically-specific syntactic representations. *Journal of Memory and Language*, 95, 102–115.
- Scheepers, C., & Sturt, P. (2014). Bidirectional syntactic priming across cognitive domains: from arithmetic to language and back. *The Quarterly Journal of Experimental Psychology*, 67(8), 1643–1654.
- Scheepers, C., Sturt, P., Martin, C. J., Myachykov, A., Teevan, K., & Viskupova, I. (2011). Structural priming across cognitive domains: From simple arithmetic to relative-clause attachment. *Psychological Science*, 22(10), 1319–1326.
- Schlueter, Z. (2017). *Memory retrieval in parsing and interpretation* (Unpublished doctoral dissertation). University of Maryland.
- Schlueter, Z., Parker, D., & Lau, E. (2019). Error-driven retrieval in agreement attraction rarely leads to misinterpretation. *Frontiers in Psychology*, 10, 1002.
- Schotter, E. R., Angele, B., & Rayner, K. (2012). Parafoveal processing in reading. *Attention, Perception, & Psychophysics*, 74(1), 5–35.

- Sloggett, S. (2017). *When errors aren't: How comprehenders selectively violate binding theory* (Unpublished doctoral dissertation). University of Massachusetts, Amherst.
- Smith, M., & Wheeldon, L. (2001). Syntactic priming in spoken sentence production—an online study. *Cognition*, 78(2), 123–164.
- Smith, N., & Levy, R. (2011). Cloze but no cigar: The complex relationship between cloze, corpus, and subjective probabilities in language processing. In *Proceedings of the annual meeting of the Cognitive Science Society* (Vol. 33).
- Snyder, W. (2000). An experimental investigation of syntactic satiation effects. *Linguistic Inquiry*, 31(3), 575–582.
- Sprouse, J. (2009). Revisiting satiation: Evidence for an equalization response strategy. *Linguistic Inquiry*, 40(2), 329–341.
- Staub, A. (2010). Eye movements and processing difficulty in object relative clauses. *Cognition*, 116(1), 71–86.
- Staub, A., Dillon, B., & Clifton, C. (2017). The matrix verb as a source of comprehension difficulty in object relative sentences. *Cognitive Science*, 41, 1353–1376.
- Stenberg, K. (2007). *Bayesian model to enhance parameter estimation of financial assets: Proposal and evaluation of probabilistic methods* (Unpublished master's thesis). Stockholm University: Skolan för datavetenskap och kommunikation, Kungliga Tekniska högskolan.
- Sturt, P. (2003). The time-course of the application of binding constraints in reference resolution. *Journal of Memory and Language*, 48(3), 542–562.
- Sturt, P., Keller, F., & Dubey, A. (2010). Syntactic priming in comprehension: Parallelism effects with and without coordination. *Journal of Memory and Language*, 62(4), 333–351.
- Swets, B., Desmet, T., Clifton, C., & Ferreira, F. (2008). Underspecification of syntactic ambiguities: Evidence from self-paced reading. *Memory & Cognition*, 36(1), 201–216.
- Szmrecsanyi, B. (2005). Language users as creatures of habit: A corpus-based analysis of persistence in spoken English. *Corpus Linguistics and Linguistic Theory*, 1(1), 113–150.
- Szmrecsanyi, B. (2006). *Morphosyntactic persistence in spoken English: A corpus study at the intersection of variationist sociolinguistics, psycholinguistics, and discourse analysis* (Vol. 177). Walter de Gruyter.
- Tabor, W., & Hutchins, S. (2004). Evidence for self-organized sentence processing: Digging-in effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30(2), 431.
- Tanenhaus, M. K., & Carlson, G. N. (1990). Comprehension of deep and surface verbphrase anaphors. *Language and Cognitive Processes*, 5(4), 257–280.
- Taylor, W. L. (1953). "Cloze procedure": A new tool for measuring readability. *Journalism Quarterly*, 30(4), 415–433.
- STAN Development Team. (2017). *shinystan: Interactive visual and numerical diagnostics and posterior analysis for Bayesian models*. Retrieved from <http://mc-stan.org/> (R package version 2.3.0)
- Thothathiri, M., & Snedeker, J. (2008a). Give and take: Syntactic priming during spoken language comprehension. *Cognition*, 108(1), 51–68.
- Thothathiri, M., & Snedeker, J. (2008b). Syntactic priming during language comprehension in three- and four-year-old children. *Journal of Memory and Language*, 58(2), 188–213.
- Tooley, K. M., & Bock, J. K. (2014). On the parity of structural persistence in language production and comprehension. *Cognition*, 132(2), 101–136.
- Tooley, K. M., Swaab, T. Y., Boudewyn, M. A., Zirnstein, M., & Traxler, M. J. (2014). Evidence for priming across intervening sentences during on-line sentence comprehension. *Language, Cognition and Neuroscience*, 29(3), 289–311.
- Tooley, K. M., & Traxler, M. J. (2010). Syntactic priming effects in comprehension: A critical review. *Language and Linguistics Compass*, 4(10), 925–937.
- Tooley, K. M., & Traxler, M. J. (2018). Implicit learning of structure occurs in parallel with lexically-mediated syntactic priming effects in sentence comprehension. *Journal of Memory and Language*, 98, 59–76. doi: <https://doi.org/10.1016/j.jml.2017.09.004>
- Tooley, K. M., Traxler, M. J., & Swaab, T. Y. (2009). Electrophysiological and behavioral evidence of syntactic priming in sentence comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 35(1), 19.

- Townsend, D., & Saltz, E. (1972). Phrases vs meaning in the immediate recall of sentences. *Psychonomic Science*, 29(6), 381–384.
- Traxler, M. J. (2008). Lexically independent priming in online sentence comprehension. *Psychonomic Bulletin & Review*, 15(1), 149–155.
- Traxler, M. J., Tooley, K. M., & Pickering, M. J. (2014). Syntactic priming during sentence comprehension: Evidence for the lexical boost. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 40(4), 905.
- Ullman, M. T. (2016). The declarative/procedural model: a neurobiological model of language learning, knowledge and use. *The Neurobiology of Language*, 953–968.
- Van Dyke, J., & McElree, B. (2006). Retrieval interference in sentence comprehension. *Journal of Memory and Language*, 55(2), 157–166.
- Verde, M. F., Macmillan, N. A., & Rotello, C. M. (2006). Measures of sensitivity based on a single hit rate and false alarm rate: The accuracy, precision, and robustness of  $d'$ ,  $A_z$ , and  $A'$ . *Perception & Psychophysics*, 68(4), 643–654.
- Wagers, M. (2008). *The structure of memory meets memory for structure in linguistic cognition* (Unpublished doctoral dissertation). University of Maryland.
- Wagers, M., Lau, E., & Phillips, C. (2009). Agreement attraction in comprehension: Representations and processes. *Journal of Memory and Language*, 61(2), 206–237.
- Wells, J. B., Christiansen, M. H., Race, D. S., Acheson, D. J., & MacDonald, M. C. (2009). Experience and sentence processing: Statistical learning and relative clause comprehension. *Cognitive Psychology*, 58(2), 250–271.
- Zajonc, R. B. (1968). Attitudinal effects of mere exposure. *Journal of Personality and Social Psychology*, 9(2p2), 1.
- Zhang, J., & Mueller, S. T. (2005). A note on ROC analysis and non-parametric estimate of sensitivity. *Psychometrika*, 70(1), 203–212.