What Did You Expect? An Investigation of Lexical Preactivation in Sentence Processing

Jon Burnsky
University of Massachusetts Amherst

Follow this and additional works at: https://scholarworks.umass.edu/dissertations_2

Part of the Cognitive Psychology Commons, Cognitive Science Commons, and the Psycholinguistics and Neurolinguistics Commons

Recommended Citation
https://doi.org/10.7275/30435671 https://scholarworks.umass.edu/dissertations_2/2601

This Open Access Dissertation is brought to you for free and open access by the Dissertations and Theses at ScholarWorks@UMass Amherst. It has been accepted for inclusion in Doctoral Dissertations by an authorized administrator of ScholarWorks@UMass Amherst. For more information, please contact scholarworks@library.umass.edu.
WHAT DID YOU EXPECT?

AN INVESTIGATION OF LEXICAL PREACTIVATION IN

SENTENCE PROCESSING

A Dissertation Presented
by
JON BURNSKY

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

DOCTOR OF PHILOSOPHY

September 2022

Psychological and Brain Sciences
WHAT DID YOU EXPECT?

AN INVESTIGATION OF LEXICAL PREAMPTIVATION IN

SENTENCE PROCESSING

A Dissertation Presented

by

JON BURNSKY

Approved as to style and content by:

______________________________

Adrian Staub, Chair

______________________________

Charles Clifton, Member

______________________________

Joonkoo Park, Member

______________________________

Brian Dillon, Member

______________________________

Shota Momma, Member

______________________________

Maureen Perry-Jenkins, Department Chair
Department of Psychological and Brain Sciences
ACKNOWLEDGEMENTS

I have only been able to write this dissertation, do this research and think these thoughts with the support of many extraordinary people. There's no possible way to properly thank all of them – these words will pale in comparison to the immensity of my gratitude – I hope they all know that I’m far more grateful than my writing suggests.

All of the experiments reported here were done in 2021 and 2022, though they are rooted in thoughts that had been developing for some years prior to that. While writing a dissertation and conducting the research to fill it is never an easy feat, doing so during a pandemic has come with many additional challenges. This has made me especially appreciative of all the people in my life that have offered their friendship and help through this time.

I want to first thank Adrian Staub for guiding me through my time here at UMass. When I was a prospective student, my conversations with Adrian made coming to UMass to work with him an obvious choice. His encyclopedic knowledge of eye tracking methods and careful thinking about predictability effects have been indescribably helpful throughout all 5 years of my time here, but especially in the last year. Additionally, he has a way of making sense of psycholinguistic data and clearly articulating his ideas in ways that I aspire to be able to emulate. Thank you for being an amazing mentor, teacher and role model.

Next, I want to thank Brian Dillon for being just about equally involved in my academic career. Brian made sure that I felt at home in the Linguistics department in every way humanly possible. In addition to being a truly inspirational psycholinguist, he is an
incredibly kind and caring person. Brian might have some kind of mind control powers, because his questions in our meetings always lead to a lightbulb moment. But in truth, he is simply an incredible advisor. Thank you for being the other half of the reason I came to UMass and for being a wonderful role model.

I also want to thank Shota Momma. Shota joined UMass when I began my 3rd year, but I have known him since I was in my undergrad at Maryland. Shota’s was the first dissertation I ever read, and I smile to think that in some ways that has come full circle. Shota has been an invaluable colleague and mentor for his ability to always bring psycholinguistic findings back to their theoretical foundations. Thank you for always helping me sharpen my ideas and for offering all of your expertise.

Next, I want to thank Chuck Clifton. Chuck’s importance in sentence processing is impossible to overstate. I’m very grateful to have had his vast knowledge of the field guide me. And thank you for fostering the strong sense of community among psycholinguists at UMass.

Finally, I want to thank Joonkoo Park. Joonkoo’s outside perspective was incredibly valuable for my thinking about the problems this dissertation is investigating and how they may connect more broadly to other cognitive and developmental phenomena. Thank you for your thought-provoking questions throughout this process.

I would not be a psycholinguist if not for the wonderful faculty and students in the Linguistics department at the University of Maryland. As I’ve gathered is the case with many, I had never heard of Linguistics before undergrad. After taking my first Linguistics class, I knew I wanted to enroll in as many more classes as I could fit into my schedule and I wanted to get involved. For being amazing and inspiring teachers, I want to thank Peggy Antonisse, Tonia Bleam, Omer Preminger, Bill Idsardi, Colin Phillips and Ellen Lau. Their passion for Linguistics truly was contagious; I still think back to the excitement of learning about syntax and psycholinguistics in these courses and smile. For helping me find research opportunities, and giving me extra problem sets when I asked
for them when she was my TA, I want to thank Lara Ehrenhofer. I also want to thank Chris Neufeld, who took me on as an RA collecting MEG data, when I had no research experience and just a budding interest in psycho- and neurolinguistics. I want to especially thank Colin Phillips and Ellen Lau for being incredible and caring mentors; many of the ideas in this dissertation have their roots in conversations we had years ago and continue to have whenever we cross paths.

I am also grateful to Maryland for creating my wonderful friendships with my fellow Baggetts, Nancy Clarke, Jackie Nelligan, Michaela Socolof and Hanna Muller. I can’t imagine a better group of linguists to share that office (and chaotic whiteboard) with. Throughout grad school, these have been the names I have used in all of my example sentences when explaining syntactic phenomena – sorry that I’ve implicated you in some bizarre (and ungrammatical) situations in these sentences. I also want to thank the “Beachwee” crew, Ted Levin, Paulina Lyskawa, Maxime Papillon, Sigwan Thivierge, Nancy Clarke, Phil White, Jackie Nelligan, Adam Liter, Laura Bailey and Hanna Muller for always bringing the good vibes. And also the CUNY/HSP crew who I always look forward to chatting and hanging out with at conferences: Phoebe Gaston, Tyler Knowlton, Zoe Ovans, Masato Nakamura, Rosa Lee and Aura Cruz Heredia.

Okay that was one school. But of course I wouldn’t be here if not for the wonderful people at UMass who convinced me to come and made my time here so special. First, I want to thank my cohort mate, Sandarsh Pandey, for being a great person to share an office with and a great collaborator on class projects. I also want to thank Kuan-Jung Huang, also for being a great office mate, and for offering his expertise in reading and Mandarin. I want to thank Andrea Cataldo and Sean Conway for sharing the role of faculty liaison with me and organizing the much-needed social events and white elephant holiday parties. I also want to thank the other Cognitive graduate students for being excellent colleagues to learn beside. Getting to know all of you through the cognitive lunch talks and learning about the breadth of research that occurs in this department has been inspir-
ing. I also want to thank the other Cognitive faculty who are not on my committee but who were excellent professors. I also, of course, want to thank my RAs in the eye tracking lab, in particular those who helped with the work in this dissertation, Alexa Burk, John Collier, Alexis Conover and Liam Doherty for being simply outstanding RAs, asking all of the right questions about eye tracking research and for helping out with data collection through some very uncertain times.

I also want to thank the many people in UMass Linguistics for taking me in as one of their own and for being great friends through the years. I want to especially thank, my cohort mates: Maggie Baird, Bethany Dickerson, Duygu Goksu, Kaden Holladay, Shay Hucklebridge, Seoyoung Kim, Erika Mayer, Anissa Neal, Max Nelson, Alex Nyman and Jonathan Pesetsky for creating the many thought-provoking discussions in the Linguistics classes we took together and for being wonderful people to hang out with after a colloq. I also want to thank some of the psycholinguists who are no longer at UMass but helped create the community of inclusion that persists today: Alex Göbel, Thuy Bui, Rodica Ivan, Chris Hammerly, Micheal Wilson and Sakshi Bhatia. The Psycholinguistics Workshop has been a pillar of my grad school career, and my ideas are much sharper as a result of the conversations we all shared there. I want to thank the BUMASA crew: Maayan Keshev, Jed Pizarro-Guevara, Mariam Asatryan, Özge Bakay, Breanna Pratley and Barb Hlachova for being an incredibly talented and fun group of linguists to spend time thinking about pronouns and prediction with. I also want to particularly thank the friends who got me through the final weeks of writing this dissertation: Maayan Keshev, Jed Pizarro-Guevara, Özge Bakay, Alessa Farinella, Barb Hlachova, Polina Kasyanova and Peyton Deal. The final stretch is a trying time, but you all made my last few months fly by with long talks, many laughs, delicious food, drinks and karaoke. I want to also thank the Linguistics faculty. In particular, I want to thank Lyn Frazier, Gaja Jarosz and Kyle Johnson for being great teachers but also for being truly invaluable resources in your respective fields. I loved all of my classes, but there is of course a favorite, and yours
takes the cake, Kyle. And needless to say, I sincerely want to thank my Linguist RAs, Barb Hlachova and Zander Lynch, who have truly been the smartest, nicest and most dedicated RAs one could ever hope for. Finally, I also want to thank Thuy Bui, Alex Göbel and Yosho Miyata for being such fun and amazing housemates and friends. From late night rants about research and grad school to talking about relationships, we bonded over everything under the sun and I’ll cherish those Hawley Street days forever.

I would also be remiss to not mention my non-linguist friends who have made sure I’m not always being too nerdy about language: Danny Alarcón, Nathan Wolfe, Paul Szewczyk, Zoe Sousane, Alexandra Aiken, Madison Tralka, Joey Quinn, Ellen Wengert, Seth Kahanov, Laura Bailey, Alec Stewart, Yonca Kaya, Esther Risinger, DJ Ramadani, Shannon Krehely, Phil White, Josh Samson, Sambrum Sreenivas, Matty Hauri and Nelson Boddu.

I wouldn’t be here without the support of my family. I want to thank my brothers, Paul and Ryan, and my sisters in law, Jaclyn and Sarah, for making family gatherings such a joy and for doing the older sibling thing of offering their expertise on jobs and helping me consider possible futures. I also want to thank my nieces and nephews: Colton, Chase, Reagan, Addison, Brady and Claire for being wonderful children and of course for being a nonstop supply of fascinating linguistic data. When talking to Reagan, 4 years old at the time, about her parents’ garden that she was helping them tend to, I asked “What are you growing?” Her response “Very big;” a truly amazing answer. If I were an acquisitionist, I am sure my dissertation would be a dissection of this response to probe how children learn the intricacies of wh-dependencies. I of course want to thank my parents too. They have endured many long conversations about what it is that I do and why I’ve devoted so much time to my science. Not only that, they appear to think it’s interesting. All this time they’ve been nothing but loving and supportive and I am deeply grateful for them. Finally, I want to thank Calzone (Callie). I know everybody thinks their cat is the best cat in the world, but I happen to think everyone else is very much mistaken; there
has never been as sweet and affectionate a cat as her.

And if there’s one person who deserves the most thanks, it’s Hanna Muller. I can’t believe how lucky I am to get to call such an inspiring linguist and wondrous person my best friend and partner. Thank you for being with me every step of the way.
Language users predictively preactivate lexical units that appear to the comprehen-
der to be likely to surface. Despite ample language experience and grammatical competence, it appears that language users tend to preactivate verbs in some contexts, called role-reversal contexts, that would create plausibility violations if they were to actually appear; these verbs assign thematic roles to their arguments in such a way that it leads to implausibility. These anomalous predictions provide a window into the mechanisms underlying lexical preactivation and are the case study that this dissertation focuses on. This dissertation is an exploration of what linguistic information is effectively lever-aged to modulate lexical expectations in sentence and discourse contexts (both appropriate and anomalous) and a comparison of the methodologies that are common to test for lexical preactivation.

Using a series of cloze experiments, I will argue that these anomalous predictions
are the result of two mechanisms that co-contribute to spread preactivation to role-
anomalous verbs. The first is a small but non-negligible “Bag of Words” associative prim-
ing mechanism, and the second is a more sophisticated mechanism that prioritizes pre-
verbal arguments that are unsaturated, that is they are in need of a predicate, that I am
calling a “Bag of not yet Saturated Arguments” mechanism. Though lexical predictions
derived from sentence compositional meanings are indeed the most common, these
less sophisticated mechanisms play an important role in lexical preactivation and con-
tribute to the preactivation of both role-appropriate and role-anomalous verbs in real-
time sentence processing.

By comparing cloze responses and reading times, I also argue in this dissertation
that participants make both position-specific predictions, predictions about what words
may occur in position n+1, and predictions that are expected to surface merely sooner-
or-later in the context. I additionally argue that the N400 and eye tracking measures
such as the first fixation duration are particularly reflective of this latter type of predic-
tion, while cloze responses are less so. This explains the discrepancies observed between
these common measures of lexical preactivation in role-reversed sentences.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>ACKNOWLEDGEMENTS</th>
<th>iv</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABSTRACT</td>
<td>x</td>
</tr>
</tbody>
</table>

## CHAPTER

### 1 INTRODUCTION

1.1 Background .................................. 1
   1.1.1 The Cloze Task .......................... 4
   1.1.2 The N400 ................................ 7
   1.1.3 Reading Times ............................ 10
   1.1.4 Linking Hypotheses ....................... 12
   1.1.5 Language Model Probabilities .............. 16
   1.1.6 Thematic Roles ........................... 20
   1.1.7 Thematic Role Reversals .................... 21
   1.2 A Look Ahead ............................... 27

### 2 THE ROLE OF TIME IN PREDICTION

2.1 Experiment 1: Argument Displacement by Parenthetical Phrases .......... 33
   2.1.1 Participants ............................. 33
   2.1.2 Materials ................................ 34
   2.1.3 Procedure ................................ 37
   2.1.4 Results .................................. 39
   2.1.5 Discussion ............................... 43

### 3 THE ROLE OF ARGUMENT STATUS IN PREDICTION

3.1 Experiment 2: Fully Removed Argument .......................... 49
   3.1.1 Participants ............................. 50
   3.1.2 Materials ................................ 50
   3.1.3 Procedure ................................ 52
   3.1.4 Results .................................. 53
   3.1.5 Discussion ............................... 55

3.2 Experiment 3: Argument Removed by Prepositional Phrase .......... 56
   3.2.1 Participants ............................. 57
   3.2.2 Materials ................................ 58
   3.2.3 Procedure ................................ 60
3.2.4 Results ................................................................. 61
3.2.5 Discussion .......................................................... 63
3.3 Experiment 4: Argument Removed by Relative Clause Island .... 63
  3.3.1 Participants ......................................................... 65
  3.3.2 Materials ........................................................... 66
  3.3.3 Procedure .......................................................... 67
  3.3.4 Results .............................................................. 68
  3.3.5 Discussion .......................................................... 70
3.4 Between Experiment Comparisons ..................................... 71

4 THE ROLE OF COMPETITORS IN PREDICTION ......................... 78
  4.1 Experiment 5: A 2AFC Extension of Ehrenhofer et al. (2019) .... 80
    4.1.1 Participants ....................................................... 81
    4.1.2 Materials .......................................................... 82
    4.1.3 Procedure .......................................................... 84
    4.1.4 Results .............................................................. 86
    4.1.5 Discussion .......................................................... 91

5 PREDICTABILITY EFFECTS IN NATURAL READING .................. 96
  5.1 Experiment 6: An Eye Tracking Extension of Experiment 4 ...... 98
    5.1.1 Participants ....................................................... 99
    5.1.2 Materials .......................................................... 99
    5.1.3 Procedure .......................................................... 101
    5.1.4 Results .............................................................. 102
    5.1.5 Discussion .......................................................... 118

6 LANGUAGE MODELS AS MODEL PARTICIPANTS ........................ 131
  6.1 Experiments 2, 3 and 4 Using GPT-2 Base ........................ 134
  6.2 Experiments 2, 3 and 4 Using GPT-2 Large ........................ 136
  6.3 Discussion ............................................................ 138

7 CONCLUSION .............................................................. 146
  7.1 Mechanisms Underlying Lexical Preactivation ...................... 146
  7.2 Aligning Predictability Effects Across Methodologies .......... 149
  7.3 Limitations, Future Directions and Closing Remarks .............. 151

APPENDICES ............................................................... 154
A. EXPERIMENTAL STIMULI .............................................. 155
B. SUPPLEMENTAL EYE TRACKING ANALYSES FOR EXPERIMENT 6 ...... 157
  B.1 Analyses of the Pre-Target Region ............................... 157
  B.2 Analyses of the Post-Target (Spillover) Region .................. 159
  B.3 Analyses of the Target Word with Response Type as a Predictor ... 160

BIBLIOGRAPHY ............................................................. 162
LIST OF TABLES

1  Example stimuli for Experiment 1 .................................................. 35
2  Summary of the statistical analysis of Experiment 1. ....................... 41
3  Example stimuli for Experiment 2 .................................................. 51
4  Summary of the statistical analysis of Experiment 2. ....................... 55
5  Example stimuli for Experiment 3 .................................................. 58
6  Summary of the statistical analysis of Experiment 3. ....................... 62
7  Example stimuli for Experiment 4 .................................................. 66
8  Summary of the statistical analysis of Experiment 4. ....................... 70
9  Summary of the statistical analysis Between Experiments 2, 3 and 4. . 72
10 Summary of the statistical analysis Between Experiments 3 and 4. .... 74
11 Example stimuli for Experiment 5 .................................................. 82
12 Summary of the statistical analysis of Experiment 5 selections. ......... 89
13 Summary of the statistical analysis of Experiment 5 RTs. ................. 91
14 Example stimuli for Experiment 6 .................................................. 100
15 Summary of the statistical analysis of the probability of skipping the target for Experiment 6. ................................................................. 105
16 Summary of the statistical analysis of the first fixation duration on the target for Experiment 6. ................................................................. 106
17 Summary of the statistical analysis of the first pass time of the target for Experiment 6. ................................................................. 107
18 Summary of the statistical analysis of the go past time of the target for Ex-
periment 6. ................................................................. 107
19 Summary of the Bayesian statistical analysis of the first fixation duration 
of the target for Experiment 6 with uninformed priors. ..................... 112
20 Summary of the Bayesian statistical analysis of the first fixation duration 
of the target for Experiment 6 with informed priors. ....................... 113
21 Summary of the Statistical Analysis of the Performance on the Plausibility 
Judgments for Experiment 6. ............................................. 115
22 Summary of the statistical analysis of the probability of skipping the pre-
target region for Experiment 6. .......................................... 157
23 Summary of the statistical analysis of the first fixation duration of the pre-
target region for Experiment 6. .......................................... 157
24 Summary of the statistical analysis of the first pass time of the pre-target 
region for Experiment 6. .................................................. 158
25 Summary of the statistical analysis of the go past time of the pre-target 
region for Experiment 6. .................................................. 158
26 Summary of the statistical analysis of the probability of skipping the spillover 
region for Experiment 6. .................................................. 159
27 Summary of the statistical analysis of the first fixation duration of the spillover 
region for Experiment 6. .................................................. 159
28 Summary of the statistical analysis of the first pass time of the spillover 
region for Experiment 6. .................................................. 160
29 Summary of the statistical analysis of the go past time of the spillover re-
region for Experiment 6. ................................................... 160
30 Summary of the Statistical Analysis of the Probability of Skipping the Tar-
get Word in the Reversed Condition by Response Type for Experiment 6 . 161
31 Summary of the Statistical Analysis of the First Fixation Duration of the 
Target Word in the Reversed Condition by Response Type for Experiment 6 161
# LIST OF FIGURES

1. A schematic of a modified Lemma Model ........................................ 13
2. A schematic of an augmented Lemma Model .................................... 15
3. A schematic of Rapid Serial Visual Presentation (RSVP) ...................... 37
4. Experiment 1 Cloze Results .......................................................... 40
5. Experiment 1 Item-wise Variability .................................................. 42
6. Experiment 2 Cloze Results .......................................................... 54
7. Experiment 3 Cloze Results .......................................................... 61
8. Experiment 4 Cloze Results .......................................................... 69
9. A schematic of 2 Alternative Forced Choice (2AFC) options ................. 85
10. Experiment 5 Timeout Summary .................................................... 86
11. Experiment 5 Selection Results ..................................................... 87
12. Experiment 5 Response Time Summary .......................................... 90
13. Experiment 6 Eye Tracking Measures for the Target Word Region ....... 104
14. Experiment 6 Go Past Times by Region .......................................... 109
15. Experiment 6 Plausibility Judgment Accuracy .................................. 114
16. Experiment 6 Reading Times by Response Type ............................... 117
17. A simplified schematic of the base GPT-2 model ................................ 132
18. GPT-2 Base Assigned Probability for the Target Words ....................... 135
19. GPT-2 Large Assigned Probability for the Target Words ..................... 137
20. Visualization of GPT-2 Base Attention Scores for Reversed 2 Arguments Sentence ................................................................. 140
21 Visualization of GPT-2 Base Attention Scores for Canonical 2 Arguments Sentence ................................................................. 141
22 Visualization of GPT-2 Base Attention Scores for Reversed 1 Argument Island Sentence ......................................................... 141
23 Visualization of GPT-2 large Probabilities for 2AFC Sentence .............. 144
CHAPTER 1

INTRODUCTION

1.1 Background

Native speakers of a language can often guess likely upcoming words. That is, they can predict them. Words that are predictable within their sentence context are processed differently. They receive shorter fixations in reading and elicit reduced N400 amplitudes in event-related potential (ERP) experiments. Even in these tasks that do not require comprehenders to make predictions about how the sentence they are encountering might unfold, it appears to be the case that comprehenders are still predicting upcoming material to some degree.

In order to correctly predict upcoming material, many linguistic processes must succeed. One needs to have resolved homophony/homography and polysemy correctly, to have arrived at the correct syntactic parse to know which syntactic categories one should expect, to have composed the meaning of the sentence thus far, and to use these rich representations to probe world knowledge and search for related events or objects in memory; the list goes on.

The past few decades of psycholinguistic research have focused considerably on relating measures of lexical predictability and sentence-level processing difficulty. At the same time, in the field of Natural Language Processing, the advent of transformer models ushered in the use of large Language Models that can, in many circumstances, gener-
ate lexical predictions that are coherent and human-like in some regards. This achievement too suggests that these models often successfully parse and disambiguate sentences as well as utilize world knowledge.

Words’ predictabilities can thus be seen as a tool for diagnosing which linguistic processes must (not) have concluded by the time the prediction was made. It is in the erroneous predictions that we learn the most, for if a comprehender predicts a verb, say, where there ought not to be one, then we may infer that their parse was incorrect at the time of the generation of such a prediction.

An example of this in action comes from “Garden-Path” sentences such as (1) (Frazier and Rayner, 1982; Ferreira and Henderson, 1991; Slattery et al., 2013).

1. While the chef cooked the tofu fell.

Sentences like this are syntactically ambiguous up until the verb ‘fell.’ The ambiguity is in the attachment of ‘the tofu’ either as an object of the embedded verb ‘cooked’ or as a subject of an upcoming verb. Comprehenders reliably struggle with these sorts of sentences at the disambiguating word ‘fell.’ This is thought to be because, before encountering ‘fell,’ comprehenders, in the face of uncertainty, believe the sentence structure is more likely to unfold in such a way that ‘the tofu’ is an object of ‘cook.’ And if comprehenders are asked to continue the sentence, more of them give continuations such as ‘in the pan’ or ‘with some garlic’ which are consistent only with that structure as compared to continuations such as ‘fell.’ Thus, the lexical predictions of these comprehenders offer a way to see what their incremental sentence representations look like.

This dissertation is an exploration of the algorithm(s) humans use to generate lexical predictions. Specifically, this dissertation asks to what extent lexical predictions in sentence contexts are generated using simple word-to-word lexical associations and, to the contrary, to what extent they are guided by a sentence’s compositional meaning. Using Argument Role Reversals as a case study, and by investigating specifically which circumstances lead to erroneous predictions, I will propose what I hope to be some of
the first components of a model of lexical prediction in humans. I will argue that lexical associations play a small, but non-negligible role in lexical predictions while predictions derived from sentence compositional meanings play a much larger role. I will also argue that a Bag of Unassigned Arguments mechanism, at an intermediate level of sophistication where syntax guides which lexical associations are prioritized, plays a moderately sized role.

This dissertation is also an attempt to bridge methodological divides. As will be discussed in the remainder of this introductory chapter, predictability effects are observed in many measures: cloze probabilities, the amplitude of the N400, and reading times. While these measures usually co-vary, Argument Role Reversals are a class of sentences where this relationship breaks; the N400 and reading times suggest equal predictability of anomalous and appropriate lexical items while cloze data suggest a sizable difference. I will argue that the common linking hypotheses relating these measures to lexical activations fail to fully take into account important task differences. Predictability effects observed in ERP and reading experiments (partially) reflect predictions that are made in a position-independent fashion: they are words that are predicted to surface sooner or later, but not necessarily as the next word. On the other hand, the cloze task forces the participant to give an appropriate next word.

Finally, I will contrast human data with data obtained by probing Language Models. I will demonstrate that there is a misalignment between human cloze data and Language Model probabilities which suggests that using them to generate predictability estimates is not appropriate in all contexts.

In what follows of this introduction, I will describe the most common operationalization of predictability (the cloze task) and describe in more detail some of the online measures that predictability has been shown to influence (ERPs and reading times), along with a discussion of their linking hypotheses. I will also take a brief detour to explain the common architectures and assumptions underlying modern Language Models which
are trained on the goal of next-word prediction and which are sometimes used for predictability estimates. I will then introduce the linguistic concept of thematic roles. This will set the stage for unveiling the phenomena that the experiments of this dissertation are investigating: thematic role reversals, which are a case where we observe anomalous predictions in humans. These will be all be built upon in later chapters.

1.1.1 The Cloze Task

A word’s predictability is most often operationalized as its cloze probability (Taylor, 1953). In a cloze task, participants are asked to read a sentence fragment such as (2) and provide their guess for the next word. Exact instructions and procedures vary: this can be done online or in the lab, with the sentence fragment presented all at once or word-by-word, under time pressure or unconstrained, asking for only one word or more if needed, etc. What researchers ultimately acquire is a set of sentence continuations from which they can calculate any given word’s probability of being produced (the number of its productions divided by the number of total responses). Words with cloze values close to one are almost perfectly and uniformly predictable given their context; most people expect that specific word to surface, while those with lower cloze values are less predictable.

2. For a date, I wanted to take my partner to a fancy …

There exist other means to assess a word’s predictability in its context. These include using ‘Transitional Probabilities’ and more sophisticated Language Models. These are described in more detail in subsection 1.1.5. Smith and Levy (2011) demonstrated that cloze probabilities and corpus-based Language Model estimates are only loosely correlated, often reflecting diverging biases. A more recent systematic validation of and comparison between cloze probabilities and probabilities assigned by Language Models is in order given the dramatic advances in NLP over the past decade (this is partly the topic
of chapter 6). However, there is an a priori reason to prefer cloze probabilities as the more valid measure of the comprehender’s subjective probabilities. Cloze probabilities are provided by comprehenders themselves; Language Models on the other hand provide probabilities derived from written, often published, text, which is hardly the kind of language that everyday interlocutors use. More succinctly, comprehenders’ subjective probabilities need not be influenced by the co-occurrences of words and phrases in corpora, and if what we want is a measure of subjective probabilities, we are guaranteed to get them using the cloze task, but not using Language Models. Accordingly, cloze probabilities remain the gold standard for operationalizing a word’s predictability.

Staub et al. (2015) proposed a model of the cloze task in which it is cast as a kind of race between lexical units in the participants’ minds. This will be referred to as the Race Model. According to the model, the units in the lexicon accumulate activation as the sentence fragment unfolds, and once a unit crosses an activation threshold, it is produced. This kind of model makes accurate predictions about the response times of productions as a function of their cloze value and the constraint of the sentence. This model has been widely adopted and tweaked since to incorporate lateral inhibition between certain lexical units (Ness and Meltzer-Asscher, 2021).

It is worth highlighting that the cloze task itself is both a comprehension as well as a production task. Participants in this task are ultimately asked to produce a word, and thus they must make use of (part of) the language production system. At the very least, they must go from a non-linguistic conceptual representation through lexicalizing said concept (looking up a corresponding lexical entry) and then to planning and executing a motor command to either speak or write that word. At the same time, participants are asked to use their comprehension system. In asking participants to predict the next word, researchers implicitly ask them to parse and interpret the sentence fragment so that they may search for appropriate or related concepts in semantic memory to fill in the blank. In this way, cloze responses offer a unique window into both the language
production and comprehension systems.

It is also worth highlighting that the Race Model is a model of a process that is embedded in the much deeper and complex system of sentence processing and that the Race Model is agnostic as to how exactly activation of units in the lexicon is garnered. In order for units in the lexicon to accumulate activation as a consequence of comprehending a sentence fragment, that fragment must undergo a vast series of linguistic analyses (lexical access, polysemy and ambiguity resolution, parsing, semantic composition, etc.) and then those representations must be used to guide searches through semantic memory.

An interesting question arises from making this embedding explicit: what are the representations that are racing in the Race Model? While it is often couched in terms of “words” or “lexical units” in the lexicon, it may be more appropriate to believe that a lot of the dynamic sentence-guided flow of activation is at the conceptual level. That is, if a participant produces ‘restaurant’ as a cloze response for sentence fragment (2), they must have done so by activating the restaurant concept (in addition to ‘restaurant’ the lexical entry). This is because, in a strict sense, there is nothing about the words themselves in the sentence fragment that suggest ‘restaurant’ is appropriate in this context; only the conceptual denotations of those words and of the sentence compositional meaning do. Put simply, encountering ‘date’ may very well spread some activation to ‘restaurant,’ but it is through the conceptual, or the lexico-semantic, connections between them. Putting the emphasis on conceptual representations suggests that the process of lexicalizing those concepts is straightforward, which it is demonstrably not. There are many concepts that can be lexicalized in various ways (e.g. ‘sofa’ and ‘couch’) and given the constraints of, at least, the articulatory system, participants must choose one. This invites the possibility that there could be races at many levels of representation, first to activate associated concepts and then from those concepts to lexicalizations. These issues are discussed further in subsection 1.1.4.
This is all to say that despite the prolific usage of the cloze task, there remain many open questions about the task and what it truly taps into. For the purposes of this dissertation, I will assume that in the cloze task, participants produce the lexicalization of a highly active concept.

While the task of explicitly guessing the next word is the goal of some lines of research, word predictabilities are more often used as predictors of other neuro-psychological phenomena. Having the operationalization of a word’s predictability as the proportion of times that it wins in a cloze race allows for the correlation of predictability with other measures, such as ERPs and reading times, which I will discuss in turn below.

1.1.2 The N400

The N400 is an ERP component that surfaces as a negative-going deflection of electrical activity starting at around 250ms after the presentation of a word and peaking at around 400ms after the stimulus (Kutas and Hillyard, 1980). All content words in a sentence produce some detectable N400 and it has been demonstrated many times that a word’s N400 amplitude is inversely correlated with its semantic congruity with its preceding context and its cloze value (Kutas and Hillyard, 1984). For example, the N400 amplitude elicited by the word ‘socks’ in (3) is significantly more negative than that of ‘sugar,’ which has a considerably higher cloze probability and is a more plausible continuation. What’s more, this correlation between the N400 amplitude and cloze probabilities appears to be linear (DeLong et al., 2005; see Kutas and Federmeier, 2011 for a review).

3. I take my coffee with cream and {#socks | sugar}.

Since its discovery, the interpretation of the N400 has been disputed. Some argue that it reflects the ease or difficulty of semantically integrating the current word into

---

1 Szewczyk and Federmeier (2022) have found that this relationship is linear for the scale of around \(10^{-2}\) and up but logarithmic at lower values. They used GPT-2 (Radford et al., 2019) to estimate probabilities in the \(10^{-3}\) and below range.
the sentence context (Van Berkum et al., 1999). In other words, the N400 differences that are reported reflect a difference in plausibility between one sentence continuation and another. For example, the large N400 amplitude elicited by 'socks' (as compared to 'sugar') indexes integration difficulty; the meaning of said sentence describes an implausible event. Others, however, hypothesize that the N400 indexes the ease or difficulty of accessing lexico-semantic content from memory (Federmeier and Kutas, 1999; Lau et al., 2008; Lau et al., 2016). These researchers claim that comprehenders predictively preactivate lexical items throughout the course of incremental sentence processing and that the reported N400 differences between words reflect the amount of “work” required to fully activate a lexical item once it is encountered. Under this telling, the large N400 amplitude elicited by ‘socks’ reflects the fact that it was less pre-activated compared to ‘sugar.’ It is worth highlighting that implausible words will be less preactivated than plausible ones most of the time, if the comprehender has come to preactivate words based off of the sentence’s compositional meaning. This is discussed as preactivation because words that are implausible and unpredicted (as measured by cloze probabilities) enjoy some N400 reduction. For example, in (4) from Federmeier and Kutas (1999), the word ‘pines’ elicited a reduced N400 amplitude compared to ‘tulips’ by way of spreading activation from a more expected and plausible continuation (‘palms’). That is, since ‘pines’ and ‘palms’ share semantic features, the preactivation of those overlapping features led to less work required to access the semantic content of ‘pines’ (relative to ‘tulips’).

4. to make the hotel look more like a tropical resort, they lined the driveway with {#tulips | #pines | palms}. 

The majority of the studies reviewed in the following sections have assumed this latter lexical access account, partly because it was the trending interpretation at the time. However, it should be emphasized that the interpretation of the N400 is still not settled. One key finding that was taken as evidence for the lexical access account of the
N400 (DeLong et al., 2005) has failed to replicate in a high powered experiment (Nieuwland et al., 2018). Nieuwland et al. (2020) have in fact suggested that both integration difficulty and lexical preactivation are generators underlying the N400, as there are detectable dissociable effects of both in the classically defined N400 window of time. It has also been suggested that the N400 reflects conceptual access and thereby the preactivation of conceptual units (Federmeier, 2022). The fact, however, remains that the N400 is strongly correlated with predictability as measured by cloze probability.

Of interest regarding the reduction in the N400 amplitude has been the source of the preactivation (Lau et al., 2013). It is well established that the N400 amplitude is reduced for primed words in simple word pair studies (Bentin et al., 1985; Holcomb, 1988) as well as in sentence contexts. A longstanding question has been to what extent the N400 predictability effect in sentence contexts can be reduced to automatic associative spreading activation from single words in the context rather than being derived from compositional meanings. Nieuwland and Van Berkum (2006) investigated this by creating cartoonish contexts in which inanimate objects, such as a peanut, are animated and given human-like personalities. In such contexts, the authors found that the N400 amplitude elicited by ‘in love’ was reduced compared to ‘salty’ in the frame “… the peanut was {in love | salty}.” Despite the fact that ‘peanut’ primes ‘salty’ out of context, the broader discourse led to more preactivation of the discourse-relevant continuation ‘in love.’ This is not to say that simple word to word associations play no role in determining the level of preactivation a word receives. Metusalem et al. (2012) found that contextually implausible words did elicit reduced N400 amplitudes if they were associated with the larger context compared to if they did not. That is, ‘jacket’ produces a reduced N400 amplitude compared to ‘towel’ (though inflated compared to ‘snowman’) in the context of children staying home for a snow day and encountering a sentence such as (5).

5. They spent the whole day outside building a big {snowman | #jacket | #towel} …

Nieuwland and Van Berkum (2005) found similar results. Work on negation has also
revealed that while word to word associations lead to reduction in N400 amplitude (Fischler et al., 1983), the sentence's compositional meaning exerts a larger influence when it is sufficiently constraining (Nieuwland and Kuperberg, 2008).

1.1.3 Reading Times

In addition to ERPs, reading times have also been a valuable measure that display predictability effects. Of most interest is the effect of predictability operationalized in terms of cloze probabilities on early eye tracking measures such as the duration of the first fixation. The first fixation duration is the measure of the duration of a reader's very first fixation upon a word. The first fixation ends once the reader moves their eyes, even if it is only to refixate another part of the same word. There is ample evidence that cloze probabilities influence first fixation durations with higher cloze words leading to shorter first fixation durations (Ehrlich and Rayner, 1981; Kliegl et al., 2004; Frisson et al., 2017; a.o. see Staub, 2015 for a review). Accordingly, many models of eye movements in reading such as EZ Reader (Reichle et al., 2009) and SWIFT (Engbert et al., 2005) include a word's predictability as a predictor of fixation durations.

As is the case with the N400 literature, researchers have found that comprehenders’ predictions are not all-or-none, but rather, predictions are graded with preactivation distributed across many lexical items. Primary evidence for this comes from Luke and Christianson (2016), who demonstrated that predictability effects are observed across the full range cloze values. That is, the first fixation duration, like the N400, varies monotonically with cloze values. Moreover, they found that low cloze words with overlapping morphosyntactic or semantic features to higher cloze words benefited from such overlap. All in all, Luke and Christianson (2016) argue that the view of prediction in language processing where preactivation is graded, as opposed to all-or-none, better fits the data. Similar evidence and conclusions also come from Frisson et al. (2017) who showed that in constraining sentence contexts, where one word is highly predictable,
there is no prediction-error cost; an unpredictable but plausible word is processed similarly to when it appears in an unconstraining context.

There have also been questions as to whether the effects of predictability on the first fixation are due to preactivation per se or to plausibility. Again these are often highly correlated; implausible words are almost always unpredictable. However, researchers have demonstrated that when sentences continue in equally unpredictable, yet differentially implausible ways, there is a reliable difference in the first fixation durations as well as differences in later measures such as the first pass and go past times (Warren et al., 2008; Rayner et al., 2004). Example (7) is implausible by way of a world-knowledge violation whereas (8) is implausible by way of a selectional restriction, however, the implausibility of (7) leads to no first fixation effect as compared to (6). Plausibility effects on the first fixation duration appear to be limited to outright syntactic violations and selectional restriction violations; anomalous incremental sentence meanings do not lead to inflated times on this early measure. Relatedly, the results of Frisson et al. (2017) and Rayner and Well (1996) demonstrate that even among plausible continuations there is a predictability effect.

6. The man used a strainer to drain the thin spaghetti yesterday evening.

7. #The man used a blow-dryer to dry the thin spaghetti yesterday evening.

8. #The man used a photo to blackmail the thin spaghetti yesterday evening.

Many researchers have focused on characterizing the function that relates reading times to predictability. Smith and Levy (2013) used an N-gram language model to argue that this relationship is logarithmic for self-paced-reading (SPR), a task in which participants read a sentence by revealing words one at a time using a button press, and for natural reading. However, Brothers and Kuperberg (2021) using both SPR and eye tracking while reading to determine reading times along with cloze probabilities rather than
language model estimates demonstrated that the relationship is linear. An important difference between the two is the method for estimating a word’s predictability.

It is worth noting that there is also ample evidence that the N400 and the first fixation are not redundant measures reflecting the same underlying processes (Kretzschmar et al., 2015; Burnsky et al. under revision). These coregistration experiments collect simultaneous eye tracking and EEG data and demonstrate that within an experiment, effects visible in the first fixation duration are sometimes not detected in the N400. Of particular interest are the findings of Burnsky et al. (under revision), where the authors report that the predictability effect observed for the first fixation duration depends on valid parafoveal preview of the target word, whereas the predictability effect on the N400 is larger with invalid preview of the target word. This dissociation demonstrates that these measures cannot be reflections of an identical set of processes and suggests that there may in fact be at least two predictability effects.

1.1.4 Linking Hypotheses

The previous three sections laid out measures that are often used to investigate lexical predictability effects. This is because the common linking hypotheses between these measures and underlying processes claim that they are (partially) reflections of activations.

The Race Model of the cloze task and its variants all claim that participants in a cloze task produce the winner of a race between words accumulating activations from the preceding context (Staub et al., 2015). Accounts of the generator of the N400, except for the integration-only views as in Van Berkum et al. (1999) and Baggio and Hagoort (2011), postulate that the N400 amplitude provides a measure of the “work” required to fully activate the current word (or concept) from its pre-stimulus activation state (Federmeier, 2022). Finally, EZ-Reader assumes that preactivating lexical units leads to faster completion of the “familiarity check” stage (L1) of visual word recognition (Reichle et al., 2009).
Taking these linking hypotheses at face value, one is tempted to assume reliably co-varying results across these methodologies. That is, low cloze items should reliably yield more negative N400 amplitudes and inflated reading times because such an item is not highly preactivated.

However, one caveat is that these measures are likely reflecting the preactivation of units at different levels of representation. For example, many accounts of the N400 argue that what is preactivated is either conceptual or lexico-semantic features (Federmeier and Kutas, 1999; Federmeier, 2022), while the preactivation that leads to shorter first fixation durations in reading is thought to be due to form-based predictions (Reichle et al., 2009; Staub and Goddard, 2019; though see Staub, 2020 for conflicting evidence). Predictions of forms appear be quite restricted in ERP experiments (Ito et al., 2016).

In order to make explicit the levels of representation and their relations, I am adopting a modified version of the lemma model (Levelt et al., 1999). This is schematized in Figure 1.

Figure 1: A schematic of a modified Lemma Model (Levelt et al., 1999). Each word is represented as its form (phonological or orthographic), its lemma and its semantic features. The non-linguistic conceptual representation (outlined in red) is an addition of my own.

According to the Lemma Model, there is a hierarchy of related representations for lexical items. It is worth noting that “words” and “lexical items” refer to indivisible bits of
linguistic meaning and thus range from morphemes (e.g. the plural marker -'s’) to verb
particle constructions (e.g. ‘went off’ meaning exploded) to fully idiomatic phrases (e.g.
the non-compositional phrase ‘the cat’s out of the bag’). At the lowest level, there is the
perceptual form, which is the orthographic or phonological signal that corresponds to
the item. The next level of representation is the lemma, a more abstract representation
that houses the unit’s syntactic properties but is devoid of meaning. Finally, the lemma
links to the semantic features of the unit, where the lexico-semantic meaning of the unit
is accessed. In addition to this level of representation, I have added a non-linguistic con-
ceptual level of representation. This allows us to think more comprehensively; linguistic
expressions ultimately are used to map to non-linguistic mental objects, scenes, events
or ideas.

While the Lemma Model was originally proposed as a model for language produc-
tion, parsimony would argue that (many of) the representations used for speaking and
understanding ought to be shared (Momma, 2016). Accordingly, I will assume some-
thing analogous to the Lemma Model as a model of lexical processing in both produc-
tion and comprehension.

However, this model offers an incomplete picture, as there is no mention of the in-
tegration of words into their broader sentence or discourse contexts. Accordingly, it is
amended to include sentence-level properties in Figure 2.
Throughout the course of incremental sentence interpretation, comprehenders construct representations of the current word and integrate those representations into their representation of the sentence; the parse of the sentence is amended and the corresponding meanings of the constituents are composed. Similar to the case with single words, comprehenders ultimately use the linguistic representations of sentences and discourse to construct non-linguistic concepts: events with participants.

Having this model makes clear that there are many representations with varying levels of abstractness that may be preactivated in sentence processing (see Kuperberg and Jaeger, 2016). What’s more, this model makes it clear that there are many levels of representation and corresponding processes that use those representations, that our experimental measures may be tuned to.

With this model it is clear that the linguistic representations of semantic features, lemmas and perceptual forms can be preactivated in a few ways given some preceding context. First, and maybe simplest, is associative concept-to-concept priming. If the
comprehender encounters ‘dog’ and activates the dog concept, some activation spreads to the cat concept which can send activation down the hierarchy to activate the linguistic representations that correspond to ‘cat.’

For lexical items to be preactivated by way of the sentence context's compositional meaning, the non-linguistic event concept plays a crucial role. Throughout the course of incremental interpretation, the comprehender uses the language input to derive non-linguistic representations of the event or idea being discussed. The communicative goal of language comprehension has been argued to be just this; inferring an intended message about an event concept (Kuperberg and Jaeger, 2016). So as the comprehender builds up this event concept, the concepts that correspond to components of the event are in turn activated. These preactivated concepts then can be lexicalized which means their corresponding linguistic representations too are preactivated.

### 1.1.5 Language Model Probabilities

Before moving on from the more methodological portion of the introduction to the theoretical one, it is worth briefly introducing and explaining modern Language Models which also provide estimates for words’ predictabilities. A Language Model is a statistical model that assigns a probability to a string, that is, a sentence. This probability can be decomposed to provide individual conditional probabilities for each word in the sentence given the previously encountered words using the Chain Rule of probability. Thus, one can provide a Language Model with the beginning of a sentence and it will output a probability distribution over the lexicon; estimates for how likely each word is to surface next. Given a corpus, these probabilities are derivable.

Language Models underlie almost all Language related technologies, from Automatic Speech Recognition (ASR) to autocomplete to text classification. Their utility is based on an assumption that grammatical strings are more probable than ungrammatical ones, and otherwise natural-sounding strings are more probable than otherwise odd ones.
For example, if a model must translate a sentence about the weather and the candidates are ‘expect heavy rain’ and ‘expect overweight rain,’ the former should be selected. The model would make this decision since ‘heavy rain’, or something like it, has likely been seen before, whereas having encountered the sequence ‘overweight rain’ is highly unlikely.

The simplest language models are N-Grams (Shannon, 1948). In training an N-Gram Model, one combs through corpora counting how often specific sequences of words of length $N$ occur. For example, if one is training a trigram model ($N = 3$), the model stores every iterative sequence of three consecutive words. The first two words in the sequence are the context. All sequences sharing this same context can then be identified and one can simply calculate the proportion of times any given third word follows the context to arrive at that word’s probability (in the corpora) given the previously encountered two words.

N-Grams have obvious shortcomings. First is their limited memory. Any word occurring $N+1$ words back is forgotten. This is an obvious problem for non-local dependencies, which are common in language (e.g. agreement, licensing, etc.), as the non-local element can be arbitrarily displaced outside of any range $N$. A solution to this might be: make $N$ very large. This would still fail to capture all non-local dependencies since these spans can be arbitrarily long. It also, however, reveals another shortcoming: there are many sequences, given any $N$, that are simply unobserved in the corpora, and therefore have zero probability. This equates coincidentally unseen sequences with ungrammatical “word-salad” that never would be encountered. That is, these models will necessarily assign zero probability to sentences such as (9), which is judged to be grammatical, as well as zero probability to (10), which is ungrammatical (Chomsky, 1957), an undesirable conflation.

9. Colorless green ideas sleep furiously.

10. * Furiously sleep ideas green colorless.
To better deal with unseen sequences, Bengio et al. (2000) proposed an N-Gram that represents words as high-dimensional but dense vectors (Hinton et al., 1986). Traditional N-Grams effectively use one-hot encodings; every word is unique and discrete which can be represented as a vector that is a series of 0s with a 1 in a position that corresponds to the “dimension” of that word. Learned dense vectors, called word embeddings, have fewer dimensions than the size of the lexicon, which gives rise to the virtue that words with overlapping features tend to have similar dimensional values in this new vector space. Thus, if there is a gap in the training data such that ‘cat’ was not seen in a specific context, but ‘dog’ was, and ‘cat’ and ‘dog’ are similarly situated in vector space, then the sequence with ‘cat’ is no longer zero probability.

The use of Recurrent Neural Networks (RNNs) was intended to alleviate the memory problems of N-Grams as previously encountered elements are cumulatively stored in a expanding context (Elman, 1990). Thus, in principle, words from arbitrarily long ago can be used as the context with which the probability of the current word can be modulated. Hochreiter (1998) however, demonstrated that practically there is a limit on how far back the context is effectively leveraged (termed “the vanishing gradient problem”).

Modern Language Models such as GPT-2 (Radford et al., 2019) are built on a different kind of neural network architecture: the Transformer (Vaswani et al., 2017). The specifics of GPT-2 are described in more detail in chapter 6. Transformers are Deep Neural Networks (DNNs), meaning there is more than one hidden layer between the input and output layers. What makes Transformers unique is the “attention” mechanism. Attention allows the Transformer to prioritize certain words in the context over others; recent words are not necessarily the most influential as is the case with RNNs. By using attention, and stacking more and more layers (and more and more parameters) into the model, Modern Language Models like GPT-2 have been able to achieve state of the art status in many NLP tasks and can often generate coherent and human-like strings of text.
Language Models are not grammars, per se. They are intended to capture and distill statistical regularities of language and world-knowledge data through written text (Iyyer, p.c.). That is, in order to assign probabilities correctly, these models must mimic the rules of language (e.g. verbs in English must agree in number with the subject, which is a hierarchically defined, as opposed to a linear, position). Importantly, however, none of the models discussed here are language specific; they are general purpose sequence processors, and they are used in many other fields such as computer vision (Khan et al., 2021) and bioinformatics (Nambiar et al., 2020).

Given that modern Language Models seem to distill language statistics quite well, their probabilities are sometimes used as proxies for human comprehenders’ subjective probabilities of upcoming words when norming materials. One advantage of this method over cloze is the granularity at low values. For example, in order for a researcher to distinguish a 1% cloze value from a 2% cloze value, one needs to run 100 participants. This becomes unwieldy for finer distinctions between say, 0.1% and 0.2% or 0.01% and 0.02%. If one instead uses a Language Model, these fine distinctions are readily available.

While Language Models are intended primarily as engineering tools, psycholinguists have repurposed them as model participants in a sense. They can be used as a proof of concept to ascertain what information is extractable from raw text data with minimal assumptions (as opposed to the Chomskyan view of grammar induction guided by biological predispositions: Universal Grammar (UG)), and can thus be probed to offer insights into how one might solve certain linguistic tasks. Language Models exhibit some human-like behaviors, as indexed by their assigned probabilities, such as garden-pathing (Van Schijndel and Linzen, 2018) and the SRC/ORC processing distinction (Lakretz et al., 2020). Language Models have also been shown to demonstrate some human-like agreement patterns and, interestingly, agreement errors (Gulordava et al., 2018; Arehalli and Linzen, 2020). This has led some to probe the inner workings of the model. For example, Lakretz et al. (2021) found that the model in Gulordava et al. (2018)
achieves subject verb agreement (and produces agreement attraction errors) by distinctly representing the number of the most recently encountered DP and the number of the hierarchical subject. However, Language Models do not behave entirely as human comprehenders do; sophisticated models such as BERT (Devlin et al., 2018) fail to use context as humans do to generate expectations for upcoming words, even failing to effectively utilize sentential negation (Ettinger, 2020). Even so, their use for quick predictability estimates is warranted in some psycholinguistic circumstances.

1.1.6 Thematic Roles

Before discussing the studies in this dissertation, some theoretical groundwork is necessary. Thematic roles are proposed linguistic representations that operate at the syntax-semantics interface, linking arguments of a verb to particular types of participants in the event described (Jackendoff, 1972; Baker, 1997). Thematic roles can be thought of as a kind of entailment relationship between a verb and the argument. Knowing the thematic roles of a verb’s arguments unambiguously informs the comprehender who is “doing” the verb (the “Agent”) and who is having the verb “done to them” (the “Patient”). That is, thematic roles inform the comprehender of who is doing what to whom.

The importance of some kind of representation such as thematic roles is clear when one considers active compared to passive sentences and subject-experiencer compared to object-experiencer verbs, just to name a few. That is, there is not a 1:1 correspondence between surface syntactic position and the semantic role an argument has in an event; syntactic subjecthood is not on its own a determinant of agenthood. Agent and Patient are only a few of a handful of generalized thematic roles which are thought to characterize shared properties across sentences.

The theoretical landscape around thematic roles is not without controversy. Kratzer (2003) argues that the role Agent exists while Theme, which describes the relationship between a verb and its internal argument, doesn’t. Williams (2009) argues that poten-
tially neither truly exists. Williams (2015) further elaborates this idea by arguing that verb-specific lexicalized roles, such as ‘kicker’ and ‘kickee,’ should be considered as the representation of the relationship between verb and argument. Finally, there is not a consensus that these relations must be linguistic in the first place. Rissman and Majid (2019) argue that thematic roles may be non-linguistic representations that are a part of more conceptual knowledge of events.

Regardless of the specific characterization of thematic roles, what remains clear is that there is an unambiguous correspondence between an argument’s syntactic position and its interpretation as a participant in the event being described once a particular verb and its voicing are provided. Before the verb itself is encountered, the thematic role of an argument cannot be known for certain. However if one assumes that there are such things as generalized thematic roles such as Agent, then the comprehender may reasonably guess that the syntactic subject of a verb is likely to be an Agent, as this is a overwhelmingly common configuration.

1.1.7 Thematic Role Reversals

As discussed in subsection 1.1.2, the amplitude of the N400 elicited by a word tends to track its preactivation and its fit/plausibility in the surrounding context. This preactivation–plausibility correlation however does not hold in all contexts. Kim and Osterhout (2005) recorded ERPs and presented participants with sentences such as those in (11) through (13).

11. The hungry boy was devouring the cookies.

12. #The hearty meal was devouring the kids.

13. #The dusty tabletop was devouring the kids.

They found that the target word (underlined in the examples) elicited the same N400 amplitude in (11) and (12) despite the clear animacy restriction violation in (12). Rather,
the P600 component showed a difference, where the anomalous condition yielded a more positive P600. When the ERPs elicited by (13) were compared to (11), there was a clear N400 difference. Working under the assumption that the N400 indexed semantic integration, Kim and Osterhout took this to suggest that for sentences in which there is a tempting or attractive argument-role mapping that is not in line with the syntax (such as assigning the theme role to the active subject ‘the hearty meal,’ since meals are often devoured), comprehenders pursue that nonveridical parse and find it easy to compute a plausible meaning. When this attractiveness is removed, as is the case with ‘the dusty tabletop’ and ‘devour,’ the N400 reflects difficulty integrating or composing the two. They call this effect ‘Semantic Attraction.’ The P600 effect elicited by encountering ‘devour’ in (12) was taken to reflect the comprehender ultimately pursuing the veridical parse of the sentence; assigning the thematic roles correctly and in accordance with the syntax.

Similar findings have been observed in a number of studies (Hoeks et al., 2004; Kuperberg et al., 2006; reviewed in Kuperberg, 2007). Under the integration account of the N400, the lack of a difference in amplitudes between these conditions suggests that ‘devouring’ is surprisingly easy to integrate into certain anomalous contexts. Under the preactivation / retrieval account of the N400, these findings suggest that ‘devouring’ is anomalously preactivated. That is, despite ‘devouring’ being an unlikely form to appear, comprehenders seem to expect it more than baseline. Another possible explanation for such an anomalous prediction which falls under the preactivation / retrieval view of the N400 is that in sentence contexts with little constraint and strong lexical associations, implausible verbs are preactivated above their baseline primarily due to lexical associations between the nouns in the sentence and the verb (Chow and Phillips, 2013).

Chow et al. (2016b) followed this idea and focused on the lack of a significant N400 effect in role reversal sentences. They used sentences such as (14) and (15).

14. The restaurant owner forgot which customer the waitress had served during din-
15. #The restaurant owner forgot which waitress the customer had served during dinner.

The use of embedded wh-questions in English allowed for both of the verb's arguments to come before the verb itself. The structure of this type of clause dictates that the wh-element is the verb's internal argument (the object) while the second DP is the external argument (the subject of the verb). This is unambiguously clear by the time the comprehender enters the second DP. Thus, (15) is strange because, according to the syntax, the waitress is being served by their customer, which is the opposite of how serving events typically unfold. Again, despite this implausibility, no N400 effect was observed. Instead a P600 effect was observed, which the authors took to indicate that the participants did indeed notice the anomaly (at some level).

An N400 effect did arise, however, in another set of conditions that authors utilized, comparing sentences such as (16) and (17)

16. The exterminator inquired which neighbor the landlord had evicted last May.
17. The neighbor inquired which exterminator the landlord had evicted last May.

In (17) a landlord may very well evict an exterminator, however this scenario is not as predictable as a landlord doing so to an unspecified neighbor\(^2\). Instead, a thing that landlords are more likely to do to an exterminator is hire them for their services. Chow et al. (2016b) found that this manipulation, similar to the ‘dusty tabletops’ condition in Kim and Osterhout (2005), brought back the N400 effect on the verb. Note that this manipulation altered the identity of the verb's arguments, not just the argument-role mappings as with previous experiments. Thus, it would appear that lexical associations between recently encountered nouns and the verb in question can modulate the N400 amplitude elicited by verbs.

\(^2\)This is to say that ‘evict’ is lower cloze (0.008) when ‘exterminator’ is the object compared to when the object is ‘neighbor’ (0.22).
Chow et al. (2016b) proposed that these role reversal cases reveal that comprehenders preactivate verbs using two distinct mechanisms which utilize different features of the sentence: A Bag of Arguments mechanism and a more sophisticated mechanism that uses the full sentence compositional meaning to make more refined predictions (articulated further in Chow et al., 2016a). They propose that, first, the comprehender preactivates verbs that are simply associated with the preverbal arguments. This means that in (16), the embedded DPs (‘neighbor’ and ‘landlord’) are used as retrieval cues for a search in semantic memory for events, and corresponding verbs, that involve these entities. Importantly, these entities are used in a role-independent manner; the comprehender does not consider events in which the landlord is the agent any more than they consider events in which the landlord is the patient. It is because of this non-commitment to argument role mappings that they invoke the metaphorical grab-bag; all of the contents of the bag are in a sense jumbled and unstructured. The matrix subject ‘exterminator’ is not included in this process. They term this mechanism the Bag of Arguments to draw a contrast between a less sophisticated mechanism one could imagine being at fault here, a Bag of Words mechanism, in which the comprehender equally weights all words in the sentence to spread activation to verbs. However, in Chow et al. (2016b), they argue that this cannot be the case as keeping the words of the sentence constant but manipulating the clause of which the DP is a part, as in (16) and (17), influences the N400 amplitude. What seems to matter is the identity of those two embedded DPs, the arguments of the upcoming verb.

After quickly preactivating verbs using the Bag of Arguments mechanism, the argument role mappings are applied, making the embedded subject the agent of the upcoming verb, resulting in more refined predictions. The authors remain uncommitted as to whether this later stage should be viewed as a filter on earlier predictions or as providing a boost on top of earlier predictions (Gaston, 2020). It is worth highlighting that under the modified Lemma Model, discussed earlier, as well as the conventional Lemma Model
(Levelt et al., 1999), many semantic features of these role-anomalous verbs are actually not anomalous at all. What makes them so is the active voice they are in, which disambiguates which thematic roles should get assigned to which arguments. For example, by manipulating the immediately surrounding context of (15), we can create (18).

18. The restaurant owner forgot which waitress the customer had been served by during dinner.

By making this a passive clause, there is no anomaly; the event being described is unfolding the canonical way with a waitress doing the serving and the customer being served. So it is also worth highlighting that an implicit assumption in pondering why comprehenders preactivate ‘served’ in (15) is that comprehenders should presume that the subject is an agent; they should assume active voice. This is not an unfounded assumption; actives are considerably more frequent than passives, and children reliably presume subjects to be agents (Slobin and Bever, 1982; Huang et al., 2013). However, Trueswell et al. (1994) demonstrated that the typicality of a referent as an Agent versus a Patient is available quickly for adult comprehenders, albeit in extreme cases (by manipulating animacy). This is all to say that given the temporary but relatively long-lasting ambiguity in the voice of the embedded clause in sentences such as (15), it would not be disadvantageous for the comprehender to have some preactivation spread to passive-compatible continuations of the fragment, at least some of the time, essentially hedging their bets.

The idea that comprehenders appear to have anomalous verbs preactivated in anticipation of a passive sentence was investigated in Burnsky and Staub (2019). Participants were presented with sentences such as (19) and (20), and asked to select which alternative continuation was better.

19. The man saw which patient the doctor had been . . . {treating | #treated by}
20. The man saw which doctor the patient had been . . . {#treating | treated by}
Rather than uniformly selecting the passive option in (20) participants were lured into selecting the anomalous continuation roughly a quarter of the of the time. In conditions where ‘treated’ and ‘hugged by’ were the options given the context in (20), participants selected the anomalous continuation ‘treated’ roughly two thirds of the time. RT data revealed an overall slowdown in these conditions, indicating that participants had not stochastically landed on an argument-role mapping prior to their selection. Thus, it appears that the preactivation of ultimately role-anomalous verbs is not simply due to comprehenders expecting a passive form. Additionally, in other languages such as Mandarin, there are preverbal morphemes (‘BA’ and ‘BEI’) that may be used to disambiguate thematic role assignment before encountering the verb, and role reversals appear in these languages too (Chow and Phillips, 2013; Chow et al., 2018).

As is the case with other illusions such as Agreement Attraction (Bock and Miller, 1991; Wagers et al., 2009; Staub, 2009), NPI Illusions (Drenhaus et al., 2005; Xiang et al., 2009; Muller and Phillips, 2018) and the Moses Illusion (Erickson and Mattson, 1981; Reder and Kusbit, 1991; Muller et al., 2020), these cases where the language comprehension system appears to output surprising representations or fails to represent the input veridically are of great importance for informing models of the processes the system uses. These studies from Kim and Osterhout (2005) to Chow et al. (2018) demonstrate that the mechanism that preactivates words in sentence contexts does not immediately use presumptive thematic roles to constrain expectations. That is, the verbs in the aforementioned implausible sentences are anomalous strictly because of the implausible assignment of thematic roles that the syntax dictates, but despite this, they appear to be preactivated.

It is also worth noting that other researchers have demonstrated that difficulty with thematic assignment is not restricted to online measures. Ferreira (2003) demonstrated that roughly 25% of the time, participants will report that ‘the dog’ was the Agent in the event described in (21), despite this being a syntactically disallowed interpretation.
These offline findings however can be explained by alluding either to nonveridical en-
coding of the sentence (Ferreira, 2003), or to imperfect retrieval mechanisms operating
after interpretation has occurred (Paolazzi et al., 2019; Meng and Bader, 2021).

21. #The dog was bitten by the man.

Thematic role reversal sentences such as (15) will be the primary focus of this disser-
tation. The apparent difficulty with using thematic roles to restrict our expectations and
to report the plausibility of sentence meanings offers a window into how lexical items
come to be preactivated in incremental sentence processing and how sentence mean-
ings interface with the conceptualization of events. The mechanism(s) that are deployed
to generate plausible and grammatical predictions in sentence contexts and the mecha-
nism(s) used to compose sentences and derive meanings from them veridically are con-
strained such that they must be able to output these anomalies in addition to all of the
successes that occur in other circumstances.

1.2  A Look Ahead

A sketch of the remainder of this dissertation is provided here.

In chapter 2, building upon Momma (2016) and Chow et al. (2018), I explore the
relationship between time and lexical preactivations using a cloze experiment. While
previous findings have concluded that there is a significant role of the time between a
role-anomalous verb and the words most associated with it in the verb's preactivation,
Experiment 1 demonstrates that this relationship is weak and likely more restricted in
scope than has been argued in the literature. This is taken to suggest that the mecha-
nisms that preactivate lexical items lead to relatively long lasting preactivation.

In chapter 3, building upon Chow et al. (2016b) and Liao (2020), three cloze exper-
iments that investigated the role of argumenthood in verb preactivation while control-
ling for new factors will be discussed. Experiment 2 demonstrates that the full removal
of a target verb’s lexical associate acting as an argument to that verb significantly lowers participants’ propensity to produce the target verb. Experiment 3 finds that if this lexical associate is present but in a non-argument position, there is small but reliable increase in target verb productions. And Experiment 4 finds that if this lexical associate is an argument of some verb, but can’t be an argument of the upcoming verb, there are equally as many target productions as when this associate is allowed to be an argument of the upcoming verb. These data will be used to argue for a revised version of the Bag of Arguments mechanism that I am calling a Bag of Unassigned Arguments mechanism where all active arguments (arguments that have not been composed with a predicate) are maintained and used to generate verb predictions, even if some of them cannot be an argument of the next verb. These are argued to be the mechanisms that lead to anomalous and importantly some appropriate lexical predictions, and again these mechanisms are argued to lead to long lasting preactivation, building upon chapter 2.

In chapter 4, building upon Ehrenhofer et al. (2019), I will discuss Experiment 5: a 2AFC experiment that demonstrates that even in cases where previous research has indicated that anomalous verb predictions are reduced in role-reversal contexts, role-anomalous verbs are nevertheless preferred continuations over verbs that are plausible but unassociated with the context. This is argued to be evidence that role-anomalous verbs are still preactivated in these contexts. Thus, while the findings of Ehrenhofer et al. (2019) could be viewed as problematic for the account of role-reversals discussed in chapter 3, they are in fact not in conflict and simply offer more insight into the role of encountering new information in creating a competing prediction.

In chapter 5, I will discuss an eye tracking while reading experiment that used the same materials as Experiment 4. Experiment 6 finds that the First Fixation Duration patterns more like the N400 data than like the cloze data; there is no detectable difference in the reading times of role-appropriate and role-anomalous verbs in role-reversal contexts. The differences are only reflected in later measures (Go Past Time). This experi-
ment is used to argue for the online deployment of the Bag of Unassigned Arguments mechanism, thus it builds heavily upon chapter 2 and chapter 3 by demonstrating that these same mechanisms underlie prediction in natural reading. Experiment 6 is also used to argue that the cloze task is in a way the odd measure out, so to speak. I argue that the N400 and reading times reflect the ease of processing words that are predicted to surface sooner or later in the context, while cloze measures a more constrained set of preactivated units; only words that may surface next. Thus this chapter most directly addresses the question on what to make of the most common measures of predictability effects and I offer some explanations that claim the discrepancy may not be due to the different timecourses of linguistic information coming online but rather primarily different task demands.

In chapter 6, I will describe how GPT-2, an autoregressive Language Model, assigns probabilities to anomalous sentence continuations in role reversal environments. It is found that GPT-2 assigns substantially lower probabilities to verbs that are high cloze in the human data, and assigns proportionally higher probability to those verbs in role-reversal contexts than humans do. Given these misalignments, researchers should carefully consider the pros and cons of using modern language models to estimate predictability effects. Ultimately, I argue that cloze probabilities are a more appropriate measure for psycholinguists to employ as they directly offer humans’ subjective probabilities, whether they are plausible predictions or anomalous ones.

Finally, chapter 7 will synthesize these findings, sketch out plans for future work and offer some concluding remarks.
CHAPTER 2

THE ROLE OF TIME IN PREDICTION

Chow et al. (2018) and Momma (2016) have suggested that participants’ expectations for upcoming words are dynamic; as more time passes, more information comes online which influences predictions. This chapter furthers the investigation of the role of time in lexical preactivation patterns.

While thematic role reversals often lead to the lack of an N400 effect (Kim and Osterhout, 2005; Hoeks et al., 2004; Kuperberg et al., 2006; Chow et al., 2016b), work in both Mandarin and Japanese has suggested that with additional time and / or linguistic material between the verb and the role reversed material, the N400 effect can reemerge. Chow et al. (2018) used the ‘BA’ construction in Mandarin which allows for a nonstandard SOV sentence structure (compared to the more typical SVO structure), and which indicates that the DP preceding the ‘BA’ particle is the agent of the action. The authors used sentences such as (22) and (23).

22. Police BA suspect ZAI last week arrest and bring back to the station.

23. #Suspect BA police ZAI last week arrest and bring back to the station.

These were contrasted with items that moved the temporal PP modifier ‘ZAI last week’ to the front of the sentence. The placement of the PP allowed for more or less time to pass between encountering the arguments of the verb and the verb itself. An N400 effect for the anomalous word ‘arrest’ was observed in (23) but only when the PP intervened between the arguments and the verb and not when it was at the beginning of
the sentence. Similar results were obtained in Momma (2016), in which simple two word Japanese sentences were used. The presentation rate was manipulated so as to again alter the temporal delay between arguments and the verb but without introducing more linguistic material.

Chow et al. postulated that the early Bag of Arguments stage of prediction, in which thematic role information is not yet utilized to modulate verb predictions, is succeeded by a stage where the presumed thematic roles of the arguments are effectively used. This is to account for the fact that the N400 effect appears to reemerge when the comprehender is afforded more time (Momma, 2016; Chow et al., 2018). Importantly, the authors claim that the actual parsing of the sentence is not slow, rather it is simply the use of complex memory probes that creates this delayed effect. This is analogous to another proposal made in Parker (2019), who proposed a similar account to explain why Agreement Attraction effects dissipate with time. In Parker (2019), it is argued that agreement attraction effects arise because early attempts to retrieve and activate only relevant bits of linguistic representations are prone to error, but through iterative memory sampling (i.e. reflection), irrelevant representations decline in their influence. The author notes that this may be an explanation for linguistic illusions in general.

However, the choice of manipulations used in Chow et al. (2018) and Momma (2016) may give some pause. The BA construction in Mandarin can be followed by a relative clause, which can offset the implausible mapping of the pre-BA argument to its role (Bornkessel-Schlesewsky et al., 2011; Bornkessel-Schlesewsky and Schlesewsky, 2019). That is, the sentence could continue in such a way that the ‘police officer’ is indeed the agent of some event but it is not the ‘arresting’ event; “…suspect arrest …” is in a relative clause with another agent of ‘arrest’ yet to come (e.g. ‘detective’) (Bornkessel-Schlesewsky et al., 2011). What’s more, the presence of the PP modifier may make a relative clause parse more expected (Xiang, p.c.). That is, the inclusion and placement of the PP may change syntactic expectations in addition to affording the comprehender more
time to process the arguments.

The manipulation in Momma (2016) is more distilled; the amount of time between
the verb and the role reversed material was manipulated by altering the presentation
rate, and not by adding anything new. The concern here is with the ecological validity
of the rate used and the possible effects of slow presentation rate on processing strategies. Stimulus onset asynchronies (SOAs) of 1200ms are far from the typical durations
readers spend fixating words (≈250ms per word). This is not to say that readers have
fully processed each word in that time, but under normal conditions, readers are able to
take in new information by fixating a new word at the conclusion of that fixation. With
SOAs as inflated those in Momma (2016), it stands to reason that the participants are no
longer engaged in ‘typical’ language processing, rather this opens up the door for task-
specific strategies and more conscious reflection. This is not inherently problematic, but
it raises the question of the desired breadth of models of lexical preactivation. If these
models are to capture preactivation patterns under ‘typical’ circumstances, then these
data offer less insight.

To investigate the question of the role of time in lexical preactivation patterns, Ex-
periment 1 was conducted. Rather than changing the SOA to achieve a timing manip-
ulation, and rather than using the BA construction, English parenthetical phrases were
used. Parentheticals such as “…, as claimed by X, …” are taken to be not-at-issue con-
tent (Potts, 2005). Dillon et al. (2014) demonstrated that in the course of incremental
interpretation, parentheticals and other not-at-issue content are ‘pushed aside’ and do
not significantly interfere with the processing of dependencies in the matrix clause. That
is, while introducing material such as restrictive relative clauses significantly lowered
the overall acceptability of sentences, introducing a parenthetical did not. In short, par-
entheticals were shown to be able to lengthen a sentence at no cost to the sentence’s
acceptability; the authors argue that these phrases are processed semi-independently
from the main clause. Similarly, Dillon et al. (2017) extended this finding by demon-
demonstrating that other not-at-issue content (appositive relative clauses) induce little to no processing difficulty when intervening between a filler and gap (as in wh- dependencies), while restrictive relative clauses do. This line of research suggests that not-at-issue content such as parentheticals are the perfect construction for expanding the time and linear distance between linguistic elements without adding linguistic material that can interfere with the clause of interest or alter syntactic expectations. Accordingly, Experiment 1 utilized this manipulation.

2.1 Experiment 1: Argument Displacement by Parenthetical Phrases

This experiment tested the effect of the surface positions of arguments by modulating the amount of material that intervened between them and the location of the verb they are arguments of. This was achieved by manipulating whether a parenthetical preceded the arguments entirely or intervened. The goal of this experiment was to test the idea that allowing more time to pass between encountering the extracted wh-element and the point in time in which participants must produce a verb will lessen the influence of the wh-element and lead to fewer role-reversed productions. To foreshadow, it was found that while the order of the arguments significantly altered cloze probabilities, the location of the parenthetical phrase only marginally affected cloze probabilities, and did not interact with argument order, which is potentially at odds with conclusions from previous studies.

2.1.1 Participants

A total of 209 workers on Mechanical Turk (www.mturk.com) were recruited to participate in this experiment. All participants provided informed consent and self-reported as native speakers of American English. All were within the United States. Data from only 107 participants were used after exclusion criteria were utilized (discussed in sub-
Each participant was compensated $4 for their participation. None of the participants in this experiment, or any of the experiments in this dissertation, participated in any of the other experiments in this dissertation; each experiment utilized a unique set of participants.

### 2.1.2 Materials

A total of 48 sentence fragments were constructed and distributed across six experimental conditions as exemplified in Table 1. A subset (22 out of 48) of these experimental sentence fragments were adapted from Chow et al. (2016b). These 22 particular items were chosen because in their original norming they elicited modal responses with a cloze value of .2 or higher making them moderately constraining towards a particular verb. The remaining 26 items were created in their likeness.

In addition to these 48 experimental items, there were four practice trials and 52 filler trials. Of the fillers, 24 were used for the norming of a garden path experiment featuring NP/Z and NP/S ambiguities, 12 were simple constraining sentence fragments and 12 were simple unconstraining sentence fragments. The four remaining fillers were explicit attention checks, asking participants to enter in a specific word (e.g. 'liquor').

The six conditions were created by crossing two factors: Argument Ordering and Argument Displacement. In the Canonical conditions, the subject of the verb that the participant must provide (in the examples 'lifeguard') was a canonical agent of a highly predictable verb ('save') and the extracted object of that verb ('child') was a canonical patient. The Reversed conditions featured these same entities ('child' and 'lifeguard') but placed them in the opposite syntactic positions, making 'child' the subject of the upcoming verb and 'lifeguard' the object, thus making the plausible and predictable completion in the Canonical condition ('save') no longer plausible. Finally, the Substitution conditions were identical to their respective Canonical counterparts with the exception that the syntactic subject was no longer a canonical agent of the predictable verb, but
As stated by the parent, which child did the lifeguard ... {save | rescue }  
Which child, as stated by the parent, did the lifeguard ... {save | rescue }

As stated by the parent, which lifeguard did the child ... {#save | #rescue }
Which lifeguard, as stated by the parent, did the child ... {#save | #rescue }

As stated by the parent, which child did Isaac ... {save | rescue }
Which child, as stated by the parent, did Isaac ... {save | rescue }

Table 1: Example stimuli for Experiment 1. The words save and rescue were the target responses for this example. The target(s) was/were always plausible in the Canonical conditions and implausible in the Reversed conditions (marked with a #). A full list of the materials is provided in Appendix A.

rather a proper name (e.g. ‘Isaac’). This kind of substitution was chosen over introducing another definite entity (as was the case in Burnsky and Staub (2019)) to control for the possibility that other argument substitutions may inadvertently create a constraining sentence fragment. These names are vacuous in that they do not preactivate any particular verbs systematically. That is, there is nothing about the name ‘Isaac’ that primes or preactivates ‘save,’ let alone any verb, consistently across individuals.

As for the other factor, Argument Displacement, there were only two levels. First, the Parenthetical Beginning conditions featured a parenthetical clause at the beginning of the sentence fragment. These were always of the form {“according to X” | “as stated by X”, etc.}. This in effect left the arguments of the upcoming embedded verb close to that position. That is, of the three entities mentioned in the fragment (‘parent’, ‘child’ and ‘lifeguard’), the two linearly closest to the position where participants must produce a verb were the arguments of that verb (‘child’ and ‘lifeguard’). In the Parenthetical Middle conditions, this phrase was placed directly after the extracted wh-element, making the object of the verb the furthest entity away from the position in which participants produced a verb.
These six experimental conditions were distributed across six lists in a Latin Square design; list 1 had item 1 in condition a, item 2 in condition b, and so on, while list 2 had item 1 in condition b, item 2 in condition c, and so on. Each item was presented only once.

It is worth explicitly noting that the manipulation used here, the placement of the parenthetical, differs from the manipulation used in Chow et al. (2018) in a key respect. The additional material used in Experiment 1 to temporally displace the verb and its arguments either comes before both of the arguments at the beginning of the sentence, or between the two arguments, after the extracted wh-element, but still before the verb's subject. In Chow et al. (2018) the additional material, a PP modifier, either came before both arguments at the beginning of the sentence, or after both arguments immediately preceding the verb. A key reason for this difference is that placing the parenthetical immediately preceding the verb sounds a bit unnatural at least out of context (e.g. “Which child did the lifeguard, as stated by the parent, save . . . ”). The use of other material such as restrictive relative clauses or PPs would also come with issues: restrictive relative clauses introduce significant processing difficulty between arguments and verbs (Dillon et al., 2017) and, in English, temporal PPs such as ‘last week’ are quite infelicitous intervening between a subject and a verb, and locative PPs often introduce new entities and are at-issue.

Despite this difference, using the Parenthetical Beginning and Parenthetical Middle conditions does indeed allow us to address the key question. If the presence and proximity of ‘lifeguard’ is at all to blame for anomalous productions of ‘save,’ say, then displacing it by moving it to the beginning of the sentence and allowing more time to pass before the verb it is an argument of should lessen its influence and result in fewer anomalous target responses.
2.1.3 Procedure

The experiment was designed using UMass IBEX (www.umassibex.work; based on Drummond 2013). After an experiment information and consent page, participants were told that they would be reading sentence fragments presented to them one word at a time in the middle of the screen, in Rapid Serial Visual Presentation (RSVP; see Figure 3). Each word would be present for 400ms before being replaced by the following word. The fragment was preceded by a fixation cross + in the center of the screen present for 1200ms. The fragment would end with a 400ms display of “…” followed by an empty text form box in which they were to provide “…the word that seems like the most natural continuation of the fragment.” They had as long as they needed to enter a response. They were given four practice trials to get accustomed to the task and then presented with 100 sentence continuation trials.

![Figure 3: A schematic of Rapid Serial Visual Presentation (RSVP) used in Experiment 1. Following a 1200ms fixation cross, each word in the sentence fragment was presented in the middle of the screen for 400ms before being replaced by the next word.](image)

RSVP was chosen over a presentation method where all of the fragment would be visible at once and indefinitely (as is typically the case for cloze norming experiments)
for a few reasons. First, and foremost, with RSVP the researcher has full control over how long each word is visible to participants. This also ensures that there is no variability between participants in how long they were afforded to read each word of each sentence whereas with traditional presentation methods, participants may vary greatly in which words they (re)fixate and for how long. Finally, as Chow et al. (2018) and Momma (2016)’s claims about the role of time in lexical prediction are among the most direct inspirations for this experiment, RSVP mirrors the presentation used in those ERP experiments.

After completing the sentence continuation task, there was an exit interview asking participants to answer two questions in complete sentences. These questions were; “What is the most interesting thing about your hometown or where you live?” and “What is the most boring thing you pass by on your commute to work or school? If you are currently working from home, provide something interesting you used to pass.” The entire experiment lasted a mean time of 41 minutes (maximum: 59 minutes; minimum: 11 minutes).

The exit interview and the attention check trials were used to filter out participants believed to be bots or non-native speakers. If a participant’s responses to the exit interview questions were ungrammatical, incomplete sentences, they were excluded from further analysis (83 participants). If a participant failed more than one attention check trial, they were excluded from further analysis (12 participants); this coincided with ungrammatical exit interviews frequently. Finally, responses to experimental sentence continuation trials were viewed to assess if a participant either simply provided the same response on every trial or simply copied a word from the fragment as their response. This behavior also resulted in exclusion from further analysis (8 participants). In addition to these exclusion criteria, any participant that participated in the experiment more than once was excluded. This did not apply to any participants in this experiment. As a result of these exclusions, 107 participants were left for the analyses discussed in the next section.
2.1.4 Results

Data from the remaining 107 participants were analyzed in R (R Core Team, 2021). Preprocessing utilized the tidyverse (Wickham et al., 2019) and textstem (Rinker, 2018) R packages. Responses were preprocessed by removing leading and trailing whitespace, removing non-alphabetical characters, making all characters lower case and finally, if more than one word was given, against the instructions, removing all words after the first word provided. These responses were then lemmatized (stripped of all inflectional morphology) to correct for any errors in tense that participants may have rendered.

Target word selection followed. Responses from the Canonical Parenthetical Beginning condition were isolated and then grouped by item. The modal response was selected as a ‘target’ word for that item (in the example provided in Table 1, ‘save’); any responses that were near-synonyms to that modal response were additionally selected as “targets” following the “conceptual cloze” practice established in Ehrenhofer et al. (2019). For instance, this led to both ‘scare’ and ‘frighten’ being targets for one item as well as ‘save’ and ‘rescue’ for the example provided in Table 1, and ‘cure,’ ‘heal,’ ‘diagnose’ and ‘save’ for another. However, if a response did not have the property of being plausible for the Canonical ordering of the arguments but not for the Reversed order (i.e. it is plausible either way; e.g. ‘see’), then it was not considered a target. After this process, three items were dropped from further analysis for failing to generate any targets. The number of targets for each remaining item varied (mean = 3.2, sd = 1.8, minimum = 1, maximum = 8) A full list of the targets is provided in Appendix A.

Following target selection, the cloze probabilities of the targets in each condition were computed. If a response matched any of the targets for an item, that response was coded as a 1; all-non target responses were coded as 0s. These were averaged together to get condition-specific cloze probabilities which are presented in Figure 4. As can be
seen, cloze values are much higher in the Canonical conditions compared to the Reversed and the Substitution conditions, within both levels of parenthetical placement. Additionally, it appears that if the manipulation of argument displacement affected any of the conditions, the Canonical condition was more than the others.

Figure 4: Experiment 1 Cloze Results. Cloze Probabilities by condition are plotted on the y-axis. Error bars represent one by-participants standard error.

These data were statistically analyzed using the lme4 (Bates et al., 2015) and lmerTest (Kuznetsova et al., 2017) R packages. The data being modeled were binary in nature; whether a response was among the list of targets (1) or not (0). Accordingly, a Logistic Mixed Effects Model was used to predict the probability of a target response as a function of the two manipulations, Argument Order and Argument Displacement. The contrasts for Argument Order were calculated using treatment/dummy coding, with the Reversed condition as the reference level. This in effect made the first contrast the effect of Reversed vs. Canonical (Canonical = 1; Reversed = 0 Substitution = 0) and the second
contrast the effect of Reversed vs. Substitution (Canonical = 0; Reversed = 0; Substitution = 1). The contrasts for Argument Displacement were calculated using sum coding (Parenthetical Beginning = −.5; Parenthetical Middle = .5).

\[
\text{Target} \sim \text{Displacement} \times \text{Order} + (1 + \text{Displacement} \times \text{Order} | \text{Participant}) + (1 + \text{Displacement} \times \text{Order} | \text{Item})
\]

<table>
<thead>
<tr>
<th>Fixed Effect</th>
<th>Estimate (logits)</th>
<th>z-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>−3.205</td>
<td>−17.168</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Displacement</td>
<td>−0.444</td>
<td>−1.903</td>
<td>0.057</td>
</tr>
<tr>
<td>Reversed vs. Canonical</td>
<td>3.048</td>
<td>12.964</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Reversed vs. Substitution</td>
<td>−0.337</td>
<td>−0.855</td>
<td>0.392</td>
</tr>
<tr>
<td>Reversed vs. Canonical : Displacement</td>
<td>−0.168</td>
<td>−0.619</td>
<td>0.536</td>
</tr>
<tr>
<td>Reversed vs. Substitution : Displacement</td>
<td>−0.035</td>
<td>−0.109</td>
<td>0.913</td>
</tr>
</tbody>
</table>

Table 2: Summary of the statistical analysis of Experiment 1. Bolded rows indicate significant effects (at α = 0.05). The model specification is provided above the table.

Following the advice of Barr et al. (2013), the maximal random effect structure was used in constructing the model. In this case, this model converged; its output is summarized in Table 2. The intercept of the model was significant (p < 0.001), meaning that the proportion of target responses in the Reversed condition was significantly different from 50% (logit of 0). The effect of Reversed vs. Canonical was highly significant (p < 0.001); participants were much more likely to produce a target response in the Canonical condition compared to the Reversed Condition. The effect of Argument Displacement, however, failed to reach significance (p = 0.057). Additionally neither of the interactions reached significance (ps > 0.5).
The effect of Reversed vs. Substitution failed to reach significance (p = 0.392), despite the visually present difference in Figure 4. One might also note that the estimate of the model is negative despite the bars in the figure showing, on average, higher cloze probabilities for the Substitution condition compared to the Reversed condition. This appears to be due to item variability. Upon inspection of the items in Figure 5, it can be seen that there are a few outlier items that affect the mean cloze probability in the Substitution condition. Also, while there is a uniform decrease in target probabilities from the Canonical to the Reversed conditions (all lines trend negative), there is a mixture of positive and negative lines connecting the Reversed and Substitution conditions. This is also reflected in the random effects estimates of the model: the variance for the Reversed vs. Substitution by-items random effect slope is 4.375 which is six times larger than the
variance accounted for by the by-items slope for Reversed vs. Canonical.

Given the marginally significant effect of Argument Displacement, a supplemental Bayesian analysis using brms (Bürkner, 2017) followed. Using the same formula as was used for the frequentist analysis, and using four chains with 10000 iterations (2000 iterations for burn-in) and a relatively uninformed flat prior, $\mathcal{N}(-2,2)$, for the effect of Displacement, the Bayesian model output a Bayes Factor of 2.87 in favor of the effect of Displacement being 0. The effect itself was estimated to be $-0.31$ (comparable to that in the other analysis) with a 95% HDI of $-0.76$ to $0.15$. This suggests there is only equivocal evidence for the null; the data offer neither overwhelming support for or against there being an effect where displacing the extracted wh-element lowers the probability of producing a target response in the Reversed conditions.

### 2.1.5 Discussion

Experiment 1 failed to find a significant effect of temporal displacement of the extracted element on the frequency of productions of anomalous, role-reversed responses. The Bayesian analysis corroborated this null finding with marginal support for the null; the 2% decrease in the Parenthetical Middle Reversed condition as compared to the Parenthetical Beginning Reversed condition is effectively not a decrease. Of course, this evidence in favor of the null is weak, so conclusions should be interpreted correspondingly.

The findings of Experiment 1 suggest that the anomalous productions of ‘saved’ observed in these contexts in the cloze task are not merely due to participants having recently encountered ‘lifeguard’ which fleetingly activates ‘saved.’ Rather, it appears that the displacement of ‘lifeguard’ to a position further away from the response position yields equivalent productions of ‘saved,’ which using the Race Model of the cloze task, suggests similar preactivations (Staub et al., 2015).

This is, on the face of it, opposite the patterns observed in Momma (2016) and Chow et al. (2018). These authors found that by introducing more time and/or more linguistic
material between the arguments and the verb, an N400 effect emerged; they found an interaction between argument order and displacement. However, as noted in the beginning of the chapter, the methodological difference between these studies and Experiment 1 make making comparisons between them difficult. First, there is the difference in measures. Both Momma (2016) and Chow et al. (2018) used the amplitude of the N400 to index preactivation of anomalous, role-reversed verb once they were encountered. Here the measure is cloze responses. One important difference, already known before deploying this experiment, is that the baseline Canonical and Reversed conditions yield equivalence in the verb’s elicited N400 amplitude and a sizable difference in that verb’s cloze probability.

Another important distinction is that the participants in the cloze experiment had only self-imposed time pressure; cloze responses were provided when the participant pleased. Thus, cloze responses may reflect a combination of “knee jerk” responses as well as carefully thought out responses and responses guided by participants misremembering the fragment. That is, cloze responses are not timelocked in the same way that the N400 amplitude is.

Finally, the manipulations used to temporally displace the verb and its arguments also differ between these experiments. In Chow et al. (2018), a PP modifier intervened between both of the preverbal arguments and the verb. However, in the current manipulation, only the wh-element was displaced; the syntactic subject of the verb still immediately preceded the verb. This was done partly to address concerns about some potential unintended effects of the PP modifier (Bornkessel-Schlesewsky and Schlesewsky, 2019; Xiang p.c.). However, by only displacing one of the arguments of the upcoming verb, the time since both arguments have been encountered is equal across the Parenthetical Beginning and Parenthetical Middle conditions.

All in all, while the patterns found in Experiment 1 do indeed run opposite the corresponding findings in the literature, given the differences in the design and procedure,
there is no outright contradiction. A potentially better matched experiment would be an ERP version of Experiment 1.

Taking the observed data patterns to reflect real, meaningful effects, what should be made of the fact that displacing ‘lifeguard’ has no discernible effect on the amount of anomalous productions of ‘saved?’ First, a Bag of Words mechanism with decay, where comprehenders simply spread activation to lexical associates, makes the wrong predictions about the data. We do indeed see a difference in the Canonical and Reversed conditions, which are made up of identical words, in terms of productions of ‘saved.’ In addition, we find no effect of temporal displacement. This is, however, not to say that some cloze responses may not reflect simple Bag of Words priming, it is simply that this mechanism can’t be the only mechanism at play here. Instead, either the assumption of decay must be loosened, or these responses reflect the output of the Bag of Arguments mechanism (Chow et al., 2016b).

One way to make sense of this non-effect of temporal displacement is that the participants generated an expectation of the verb ‘save’, or otherwise spread activation to it, upon encountering ‘lifeguard,’ and they maintained this prediction; this preactivation did not decay. Stone et al. (2020) argue that there is active maintenance of predictions as evidenced by work using verb particle constrictions in German, where the verb and particle may be arbitrarily far away from each other. A similar phenomenon may be occurring here; comprehenders are aware that ‘lifeguard’ must get a theta role from some verb, so associated verbs should be kept readily accessible and active. Another way to make sense of this null effect of temporal displacement is that the comprehender may have generated a prediction for ‘save’ upon encountering ‘lifeguard’ because they believe that ‘lifeguard’ is an extracted subject out of an embedded subject question, and so a verb that has ‘lifeguard’ as its syntactic subject is immanent. They may essentially be demonstrating a weak garden path effect. This prediction then simply lingers, even after the circumstances that led to its generation are proved wrong by encountering unequiv-
ocal evidence that they are in an embedded object question. A similar idea is proposed by Rommers and Federmeier (2018), who find, what they term, a “pseudo-repetition priming effect” on the N400. When a word that was once predicted but unencountered surfaces later on, the elicited N400 amplitude is reduced compared to when it had not previously been predicted. Finally, another way this pattern could arise is if the comprehender uses the Bag of Arguments mechanism once the new preverbal argument is encountered. That is, ‘lifeguard’ is reactivated once ‘the child’ is encountered, as both of these are preverbal arguments of some verb, and thrown into a Bag of Arguments to generate an (anomalous) expectation for ‘save.’ These are all live possibilities and indeed further research will have to arbitrate between them.

One possible way to test for which hypothesis might be on the right track would be to have participants perform a lexical decision for the word ‘saved’ at random points in the sentence to see if it is still preactivated, relative to a baseline, at intermediate points in time. Or one could do the same but for the word ‘lifeguard’ to see if it is reactivated. For example one could present (24) to participants. It would be expected under any account that this decision would be faster than a baseline lexical decision of ‘save.’ It could also be tested whether there is an RT difference in a position in the parenthetical like (25) which would suggest active maintenance of the prediction. Similarly, one could present participants with (26) to assess for reactivation of the extracted element at this site.

24. Which lifeguard … **Lexical Decision: SAVE**
25. Which lifeguard, according to the … **Lexical Decision: SAVE**
26. Which lifeguard, according to the parent, did the child … **Lexical Decision: LIFE-GUARD**
CHAPTER 3

THE ROLE OF ARGUMENT STATUS IN PREDICTION

As discussed earlier, role-inappropriate verbs consistently fail to produce the N400 effect usually associated with differences in predictability and plausibility. Why, mechanistically, are these anomalous verbs preactivated?

Chow et al. (2016b) proposed a mechanism that could account for these anomalous null N400 effects in role reversed sentences which they termed a Bag of Arguments. This was further elaborated upon in Chow et al. (2016a). The core of this hypothesis is that there are stages of verb prediction in which different information about the context comes online. In the early stages of verb prediction, the identities of the verb’s arguments are used to rapidly spread activation to associated verbs. This is done without making reference to those arguments’ presumed thematic roles. The authors proposed that the verb’s arguments were of a privileged status in preactivating the verb itself due to the fact that in sentences such as (16) and (17) (presented again below), despite having the same words making up the sentence, the removal of ‘neighbor’ from the embedded clause modulated the preactivation of ‘evicted.’ That is, in (16) where ‘neighbor’ and ‘landlord’ are arguments of the verb, the N400 amplitude elicited by ‘evicted’ was reduced compared to (17). Thus, the authors claim, a Bag of Words mechanism, which would make no distinction between nouns that are arguments of the verb and share the same clause with it and those that are not clause-mates, is insufficient. This is because the sentences contain the same (unstructured) bag of words; they differ in terms of their (unstruc-
16. The exterminator inquired which neighbor the landlord had evicted last May.  
17. The neighbor inquired which exterminator the landlord had evicted last May.

Chow et al. also postulate that this Bag of Arguments stage of prediction, in which thematic role information is not yet utilized to filter out verb predictions, is succeeded by a stage where the presumed thematic roles of the arguments are effectively used. This is to account for the fact that the N400 effect appears to reemerge when the comprehender is afforded more time (Momma, 2016; Chow et al., 2018).

Some key components of the Bag of Arguments hypothesis were tested in Liao (2020). In addition to running an experiment that conceptually replicated that of Chow et al. (2016b) in Mandarin, Liao ran another pair of experiments that more directly compare the Bag of Arguments hypothesis and the Bag of Words hypothesis. In Chow et al. (2016b), the arguments of the verb were the linearly closest to it; they were the most recently encountered nouns. Liao (2020) kept the nouns’ positions relative to the verb constant and isolated the effect of argumenthood by using items such as (27) and (28).

27. Millionaire BA servant fired . . .
28. Millionaire thought servant fired . . .

in (27), both ‘millionaire’ and ‘servant’ are arguments of the verb ‘fired;’ in (28) those nouns appear in the same linear positions relative to the verb, but only ‘servant’ is an argument. ‘Millionaire’ is instead an argument of ‘thought’ which is followed by a sentential complement clause. When the stimulus onset asynchrony (SOA) was 800ms per word, there was an N400 effect between the two conditions, such that ‘fired’ elicited a more negative N400 in (28) suggesting it was less preactivated compared to (27). In another experiment, Liao reduced the SOA to 600ms per word. This timing manipulation removed the N400 effect.\footnote{It is worth noting that this was not a within-subjects design, so the critical interaction was not reported.} The author took this to indicate that there is indeed a Bag
of Words stage of prediction responsible for the preactivation of ‘fired’ in the fast presentation rate version of (28), followed by a Bag of Arguments stage, and ultimately by a thematically-informed stage. It is worth noting that this “fast” presentation rate is still rather slow compared to natural reading times (∼ 250ms/word).

However, there remain some potential concerns about the hypothesis as stated. In the experiments in which non-arguments are shown to influence verb predictions to a lesser degree than arguments, the arguments have always been the closest nouns to the verb, conflating recency and argumenthood. Relatedly, it is not clear to what extent non-arguments do modulate the N400, as there are no direct comparisons between e.g. ‘neighbor’ as the matrix clause subject in (17), or ‘millionaire’ as the matrix clause subject in (28), and a noun representing an entity less associated with evicting or firing events such as ‘doctor’ or ‘professor.’ It could very well be the case that these non arguments continue to influence verb preactivations above baseline.

In what follows, I will discuss three cloze experiments that were designed to investigate the role of arguments in the pre-activation of verbs. The Bag of Arguments hypothesis claims that arguments are of a privileged status when pre-activating verbs, though many of the experiments that have tested this have confounded argumenthood with other factors (such as linear distance). Thus part of the observed effects could simply reflect comprehenders prioritizing more recent information to guide their predictions. Here these concerns are alleviated through specific experimental designs that first establish a baseline frequency of role-incompatible continuations for role-reversed sentence fragments and then disentangle argumenthood, clausemate status and recency.

3.1 Experiment 2: Fully Removed Argument

Experiment 2 was designed to evaluate whether using embedded subject wh-questions rather than embedded object wh-questions would modulate the proportion of trials in which participants provide anomalous continuations. This manipulation removed an
argument from the sentence fragment given to the participant, which effectively iso-
lated the preactivation that is due to the remaining argument. To preview the core find-
ing, it was found that removing the canonical agent of a target verb from the sentence
fragment altogether led to significantly fewer anomalous target productions. This con-
firms that the presence of the canonical agent of the target is a significant driver of its
anomalous preactivation.

3.1.1 Participants

A total of 191 participants were recruited through MTurk. All provided informed con-
sent, were self-reported native speakers of American English and were within the United
States. Data from only 108 participants were included after the exclusion criteria were
utilized. The same exclusion criteria from Experiment 1 (described in subsection 2.1.3)
were used. Each participant was compensated $3 for their participation.

3.1.2 Materials

The same set of 48 sentence fragments as were used in Experiment 1 were repur-
posed and reused here. These 48 sentences fragments were distributed across six con-
ditions; an example item is presented in Table 3. The same practice sentence fragments
and attention checks as were used in Experiment 1 were used. There were only 24 fillers
of the 48 previously used fillers in this experiment.

These six conditions result from crossing two factors: Argument Order and Argu-
menthood. The first factor, Argument Order, is almost identical to the manipulation in
Experiment 1. In the Canonical conditions, the subject of the verb that the participant
must provide ('lifeguard') was a canonical agent of a highly predictable verb ('save') and
the extracted object of that verb ('child') was a canonical patient. The reversed condi-
tions swapped the positions of these arguments, making ‘child’ the subject and ‘life-
guard’ the extracted object, reversing the presumed argument-role mapping. The key
<table>
<thead>
<tr>
<th></th>
<th>2 Arguments</th>
<th>1 Argument</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Canonical</strong></td>
<td>The parent saw which child the lifeguard had ...</td>
<td>The parent saw which lifeguard had ...</td>
</tr>
<tr>
<td></td>
<td>{saved</td>
<td>rescued}</td>
</tr>
<tr>
<td><strong>Reversed</strong></td>
<td>The parent saw which lifeguard the child had ...</td>
<td>The parent saw which child had ...</td>
</tr>
<tr>
<td></td>
<td>{#saved</td>
<td>#rescued}</td>
</tr>
<tr>
<td><strong>Substitution</strong></td>
<td>The parent saw which child the woman had ...</td>
<td>The parent saw which woman had ...</td>
</tr>
<tr>
<td></td>
<td>{saved</td>
<td>rescued}</td>
</tr>
</tbody>
</table>

Table 3: Example stimuli for Experiment 2. The target words were selected by pooling together the Canonical 2 Argument conditions from Experiments 2, 3 and 4. A full set of the materials, including the targets is provided in Appendix A.

difference is in the Substitution conditions. Rather than using a proper name (e.g. ‘Isaac’) as the subject of the verb in the Substitution condition, as was done in Experiment 1, items alternated between ‘the man’ and ‘the woman.’ These served to estimate the isolated influence of ‘child’ as the extracted object in preactivating ‘save.’ These definites served the same function as the proper names in Experiment 1; they are not semantically rich, they offer little additional contextual information, and presumably neither ‘man’ nor ‘woman’ strongly preactivate the verbs of interest. There is a practical reason for the departure from proper names. Given the syntax of 1 Argument conditions, it would be ungrammatical or unnatural to use a proper name like Isaac in the Substitution condition (??The parent saw which Isaac had ...)². This brings us to the other factor: Argumenthood. There were two levels of Argumenthood. The first made up the 2 Arguments conditions where the embedded clause was an embedded object wh-question where the verb that the participant was asked to produce was preceded by first an extracted object and then its subject. Thus, there are two arguments for the upcoming verb.

²This is only permissible in my English if there is a set of people all of whom are named Isaac, and a parent saw one of the Isaacs do some particular action.
and they are provided to the participant. In the 1 Argument conditions, the extracted object was dropped, making the embedded clause an embedded subject wh-question. The verb is preceded only by its syntactic subject.

As with Experiment 1, Argument Order is expected to modulate the proportion of trials in which participants produce a target verb. It is expected that target productions will be significantly more frequent in the Canonical condition compared to the Reversed. The effect of Argumenthood and its interaction with Argument Order are critical in this experiment. If the productions of the target verb (e.g. ‘save’) in the 2 Argument Reversed condition are due in large part to the presence of its canonical agent (e.g. ‘lifeguard’) nearby, then the removal of ‘lifeguard’ in the 1 Argument Reversed condition should significantly reduce the number of trials in which a target is produced. This manipulation also makes it possible to evaluate how much ‘lifeguard’ preactivates ‘save’ on its own by looking at the 1 Argument Canonical condition.

### 3.1.3 Procedure

The procedure mirrored that in Experiment 1 (described in subsection 2.1.3). Sentence fragments were presented using RSVP, and participants were asked simply to continue them in the most natural way they could.

After evaluating the exit interviews, 72 participants provided ungrammatical exit responses and were excluded. Further investigation of these 72 participants revealed many (64) of them simply repeated a word from the fragment as their sentence continuation, in a clear demonstration of a failure to follow the instructions in the broader task. After evaluating the attention check trials, nine participants were excluded for poor performance (more than one error). One participant performed the experiment more than once, which resulted in the exclusion of each of their attempts. This resulted in 108 usable participants, whose data are analyzed in the next section.
3.1.4 Results

The data were analyzed in R and the preprocessing steps used in Experiment 1 were followed. Target selection differed from Experiment 1 however. Experiments 2, 3 and 4 all featured exactly identical 2 Argument Canonical conditions and used the same set of items. Accordingly, to effectively increase the number of observations for each item in the 2 Argument Canonical condition to between 51 and 61 observations (rather than only twenty observations), responses from Experiments 2, 3 and 4 were pooled together to determine the target responses for an item. As with Experiment 1, the modal response was selected as a target. Then the remaining responses were evaluated for being near-synonyms, in which case they too would be counted as a target response. This led to different items having different numbers of target responses (mean = 3.4, sd = 1.7, minimum = 1, maximum = 8). One item of the set of 48 (Item #27) differed from the others in Experiment 3; this item was removed altogether from the analyses of the experiments since target selection could not proceed in the same way as it did for the other items.

Once the target responses had been identified, each response in each condition was assessed for being in the target response set. These trials were coded with a 1; trials in which a non-target response was provided were coded with a 0. The aggregated results are plotted in Figure 6.
Figure 6: Experiment 2 Cloze Results. Cloze Probabilities by condition are plotted on the y-axis. Error bars represent one by-participants standard error.

As can be seen in the figure, Targets were produced quite frequently in the 2 Arguments Canonical condition and less so in the other 2 Arguments conditions, but crucially they were still occasionally produced. Each condition in the 1 Argument set of conditions had reduced target cloze probabilities compared to their 2 Arguments counterparts, with the Reversed and the Substitution conditions falling quite close to zero.

These data were analyzed the same way the data in Experiment 1 were; a logistic mixed effects model was fit to predict the probability of producing a target given the Argument Order and Argumenthood manipulations. The contrasts were coded the same as in Experiment 1; sum coding for the effect of Argumenthood (2 Arguments vs. 1 Argument), and treatment/dummy coding for the effect of Argument Order with the Reversed condition as the reference level. The model with the maximal random effects structure was fit first, and when this model failed to converge the correlations between
the random effects were dropped; the estimates from this model are reported in Table 4.

Target $\sim$ Argumenthood * Order + (1 + Argumenthood + Order | Participant) + (1 + Argumenthood * Order | Item)

<table>
<thead>
<tr>
<th>Fixed Effect</th>
<th>Estimate (logits)</th>
<th>z-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-4.163</td>
<td>-13.792</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Argumenthood</td>
<td>-3.131</td>
<td>-6.854</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Reversed vs. Canonical</td>
<td>2.862</td>
<td>9.455</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Reversed vs. Substitution</td>
<td>-0.361</td>
<td>-0.912</td>
<td>0.361</td>
</tr>
<tr>
<td>Reversed vs. Canonical : Argumenthood</td>
<td>1.080</td>
<td>2.158</td>
<td>0.031</td>
</tr>
<tr>
<td>Reversed vs. Substitution : Argumenthood</td>
<td>-0.381</td>
<td>-0.561</td>
<td>0.575</td>
</tr>
</tbody>
</table>

Table 4: Summary of the statistical analysis of Experiment 2. Bolded rows indicate significant effects (at $\alpha = 0.05$). The model specification is provided above the table.

As can be read from the table, the effects comparing the Canonical and Reversed condition was significant ($p < 0.001$), while the effect comparing the Substitution and Reversed condition was not ($p = 0.361$). More importantly, there was a significant effect of Argumenthood ($p < 0.001$). Since the Reversed condition was the reference level, this meant that, in the Reversed conditions, the removal of the canonical agent (e.g. ‘lifeguard’) significantly lowered the probability of the production of ‘saved’ quite close to zero. This is consistent with the bag of arguments hypothesis and suggests that a significant proportion of the anomalous target responses we observe in the 2 Arguments Reversed condition are due to the mere presence of ‘lifeguard.’ It is also, however, consistent with the Bag of Words hypothesis.

### 3.1.5 Discussion

With the Reversed condition as the reference, the main effect of Argumenthood tested for a difference between the Reversed 2 Arguments condition when ‘lifeguard was present and the Reversed 1 Argument condition where ‘lifeguard was absent. This effect was significant establishing that the removal of ‘lifeguard’ lowers the probability of producing ‘saved.’
This experiment acts as a baseline; we have now established that 3% cloze is to be expected without the presence of the canonical agent (e.g. ‘lifeguard’). What Experiments 3 and 4 do is bring back ‘lifeguard’ in various syntactic positions to test if its return makes the Reversed 2 Arguments condition and 1 Argument condition again equivalent, or otherwise increases the rate of anomalous productions.

3.2 Experiment 3: Argument Removed by Prepositional Phrase

In Experiment 2, it was shown that the removal of, e.g., ‘lifeguard’ significantly decreased the probability that a participant would anomalously produce, e.g., ‘saved’ in a sentence continuation task. Experiment 2 showed that we can still expect roughly 3% cloze for ‘saved’ without the presence of ‘lifeguard.’ However, ‘lifeguard’ is largely to blame for the inflated error rates in the 2 Arguments Reversed condition. This was consistent with the Bag of Arguments and Bag of Words hypotheses, as removing ‘lifeguard’ altogether meant it could no longer be an argument of the upcoming verb. In Experiment 3, the target verbs’ canonical subjects were reintroduced but now in non-argument positions. This decouples presence and argumenthood, which leads to the Bag of Words and Bag of Arguments hypotheses making different predictions. Under the Bag of Words hypothesis, all words in the sentence fragment may exert some influence on verb preactivations, with priority potentially given to recent words. Under the Bag of Arguments hypothesis, non-arguments of the verb are not used to preactivate verbs; only arguments influence verb preactivations.

One syntactic manipulation that allows for this is Prepositional Phrases (PPs). Recall that an argument is, loosely, an entity directly involved in the action a verb denotes. In (29), ‘the cat’ is an argument of ‘caught’ as is ‘the mouse.’ ‘The dog’ however is not an argument of ‘caught;’ it is not involved in the catching event.

29. The cat next to the dog caught the mouse.
With this manipulation, we may bury ‘lifeguard’ in a PP, making it a non-argument of the upcoming verb as in (30). We may then evaluate if this buried non-argument DP still influences cloze responses. Under the Bag of Words hypothesis, this sentence fragment is identical to its counterpart where ‘lifeguard’ is an argument, thus we should see the same number of anomalous cloze responses here as we did in the 2 Argument Reversed condition of Experiment 2. However, under the Bag of Arguments hypothesis, since ‘lifeguard’ is a non-argument, it is not used to preactivate verbs, so we should observe significantly fewer anomalous cloze responses for sentence fragments like (30) compared to the 2 Argument Reversed condition on Experiment 2. Indeed it should pattern with the 1 Argument Reversed condition of Experiment 2.

30. The parent saw which child near the lifeguard had . . .

To foreshadow, Experiment 3 shows that the drop in anomalous cloze responses in Experiment 2 from 11% to 3% was likely due to ‘lifeguard’ being an argument. In Experiment 3, a similar effect is observed; ‘lifeguard,’ present but in a PP, does not yield the same number of anomalous productions as when ‘lifeguard’ is an argument. This suggests that the mere presence or absence of ‘lifeguard’ is not fully to blame for the erroneous productions of anomalous verbs. A direct comparison between the two experiments is presented in section 3.4.

3.2.1 Participants

A total of 162 participants were recruited through MTurk. All provided informed consent, were self-reported native speakers of American English and were within the United States. Data from only 120 participants were included after the exclusion criteria were utilized. The same exclusion criteria from Experiments 1 and 2 (described in subsection 2.1.3) were used. Each participant was compensated $4 for their participation.
3.2.2 Materials

The same set of 48 sentence fragments as were used in Experiments 1 and 2 were repurposed and reused here with one exception. One item was dropped from the analysis as it was not comparable across experiments 2, 3 and 4. These 48 sentence fragments were distributed across six conditions; an example item is presented in Table 5. The same practice sentence fragments, filler sentence fragments and attention checks as were used in Experiment 1 were used here.

<table>
<thead>
<tr>
<th></th>
<th><strong>2 Arguments</strong></th>
<th><strong>1 Argument (PP)</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Canonical</strong></td>
<td>The parent saw which child the lifeguard had ...</td>
<td>The parent saw which lifeguard beside the child had ...</td>
</tr>
<tr>
<td></td>
<td>{saved</td>
<td>rescued }</td>
</tr>
<tr>
<td><strong>Revers</strong>d</td>
<td>The parent saw which lifeguard the child had ...</td>
<td>The parent saw which child beside the lifeguard had ...</td>
</tr>
<tr>
<td></td>
<td>{#saved</td>
<td>#rescued }</td>
</tr>
<tr>
<td><strong>Substitution</strong></td>
<td>The parent saw which child Isaac had ...</td>
<td>The parent saw which child beside Isaac had ...</td>
</tr>
<tr>
<td></td>
<td>{saved</td>
<td>rescued }</td>
</tr>
</tbody>
</table>

Table 5: Example stimuli for Experiment 3. The target words were selected by pooling together the Canonical 2 Argument conditions from Experiments 2, 3 and 4. A full set of the materials, including the targets is provided in Appendix A.

These six conditions result from crossing two factors: Argument Order and Argumenthood. The first factor, Argument Order, is again almost identical to the manipulation in Experiment 2. The 2 Arguments Canonical and 2 Arguments Reversed conditions are identical to Experiment 2; the Canonical condition features a canonical agent of the target verb as the verb's subject (e.g. 'lifeguard' as the agent of the 'saved') and a canonical patient as the extracted object of the verb (e.g. 'child'). The Reversed condi-

---

3This item (Item 27) failed to produce a modal response in this experiment (which was temporally run first). Since all responses were unique, this item was deemed unconstraining and replaced in the other experiments for the purposes of counterbalancing.
tion switches that mapping. The 2 Argument Substitution condition in this experiment more closely matches that in Experiment 1 rather than Experiment 2. Here, the canonical agent ‘lifeguard’ is replaced with a proper name (e.g. ‘Isaac’) and this entity acts as the agent. The rationale for this substitution is that proper names do not strongly pre-activate any verbs and they allow for more diversity than recycling ‘the man’ and ‘the woman.’ The other factor was Argumenthood. These conditions are in the right column of Table 5. Here the sentence fragment contained an embedded subject wh-question that has a prepositional phrase (PP) modifier attached. The DP inside the PP is not an argument of the upcoming embedded verb, rather it is simply an argument of the preposition. In the 1 Argument Canonical condition, the canonical agent of the targets (‘lifeguard” is the subject of the embedded verb and attached to this subject is the phrase ‘beside the child.’ For the 1 Argument Reversed condition, these two DPs are switched, ‘child’ is the subject of the verb and ‘lifeguard’ is part of the DP in the PP modifier. The 1 Argument substitution condition in this experiment was different from that of Experiment 2. Here this condition took the 1 Argument Reversed condition and substituted a proper name like ‘Isaac’ for ‘the lifeguard.’ This manipulation allowed for the investigation of the difference we might see between the 1 Argument Reversed and 1 Argument Substitution conditions which would implicate the presence of ‘the lifeguard’ in the PP. That is, if entities in a PP, and moreover non-arguments, are irrelevant in the pre-activation of verbs, then these two conditions should look identical. This simply offered another way to test different predictions from the Bag of Words and Bag of Arguments hypotheses.

As with Experiments 1 and 2, Argument Order is expected to modulate the proportion of trials in which participants produce a target verb; target productions are expected to be significantly more frequent in the Canonical condition compared to the Reversed. The effect of Argumenthood is the critical question in this experiment. If the productions of the target verb (e.g. ‘save’) in the 2 Argument Reversed condition are due in
large part to the mere presence of its canonical agent (e.g. ‘lifeguard’) nearby (a la Bag of Words), then burying ‘lifeguard’ in a PP in the 1 Argument Reversed condition should have no effect on the number of trials in which a target is produced. At the same time, and for the same reason, we should also find that target productions are higher in the 1 Argument Reversed condition compared to the 1 Argument Substitution condition. If, however, the effect in Experiment 2 was due to the fact that ‘lifeguard’ was not present as an argument, then we should observe fewer target responses in the 1 Argument Reversed condition compared to the 2 Arguments Reversed condition and equivalent target productions in 1 Argument Reversed and 1 Argument Substitution conditions.

The actual comparisons that are needed to evaluate the question of what caused the decrease in target productions across the 1 Argument and 2 Arguments Reversed conditions in Experiment 2 are across experiment comparisons. That is, the results of Experiments 2 and 3 should be compared side by side in the same statistical model to look for interactions between them. This is done in section 3.4.

### 3.2.3 Procedure

The procedure mirrored that in Experiments 1 and 2 (described in subsection 2.1.3). Sentence fragments were presented using RSVP, and participants were asked simply to continue them in the most natural way they could.

After evaluating the exit interviews, 29 participants provided ungrammatical exit responses and were excluded. Further investigation of these 29 participants revealed that many (18) of them simply repeated a word from the fragment as their sentence continuation, in a clear demonstration of a failure to follow the instructions in the broader task. An additional two participants demonstrating this behavior in experiment, but with coherent and grammatical exit interviews, were removed from the analysis. After evaluating the attention check trials, ten participants were excluded for poor performance (more than one error). One participant performed the experiment more than
once, which resulted in the exclusion of each of their attempts. This resulted in 120 usable participants, whose data are analyzed in the next section.

3.2.4 Results

Data were analyzed in R following the same steps as Experiment 2. The targets were the same as in Experiment 2 since they were selected by pooling data from Experiments 2, 3 and 4 together (discussed in subsection 3.1.4).

Each response in each condition was assessed for being in the target response set. These trials were coded with a 1; trials in which a non-target response was provided were coded with a 0. The aggregated results are plotted in Figure 7.

![Figure 7: Experiment 3 Cloze Results. Cloze Probabilities by condition are plotted on the y-axis. Error bars represent one by-participants standard error.](image)

As can be seen in the figure, target responses were more common in the 2 Argument Canonical condition as compared to the other 2 Arguments conditions. Target responses
were also less common in the 1 Argument conditions across the board.

These data were statistically analyzed the same way the data in Experiments 1 and 2 were; a logistic mixed effects model was fit to predict the probability of producing a target given the Argument Order and Argumenthood manipulations. The contrasts were coded the same as in Experiment 1; sum coding for the effect of Argumenthood (2 Arguments vs. 1 Argument), and treatment/dummy coding for the effect of Argument Order with the Reversed condition as the reference level. The model with the maximal random effects structure was fit first, and this model converged; the estimates are reported in Table 6.

\[
\text{Target} \sim \text{Argumenthood} \times \text{Order} + (1 + \text{Argumenthood} \times \text{Order} | \text{Participant}) + (1 + \text{Argumenthood} \times \text{Order} | \text{Item})
\]

<table>
<thead>
<tr>
<th>Fixed Effect</th>
<th>Estimate (logits)</th>
<th>z-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>−3.934</td>
<td>−14.608</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Argumenthood</td>
<td>−2.494</td>
<td>−6.242</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Reversed vs. Canonical</td>
<td>3.045</td>
<td>11.897</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Reversed vs. Substitution</td>
<td>−0.418</td>
<td>−1.311</td>
<td>0.190</td>
</tr>
<tr>
<td>Reversed vs. Canonical : Argumenthood</td>
<td>0.449</td>
<td>1.069</td>
<td>0.285</td>
</tr>
<tr>
<td>Reversed vs. Substitution : Argumenthood</td>
<td>−0.745</td>
<td>−1.600</td>
<td>0.110</td>
</tr>
</tbody>
</table>

Table 6: Summary of the statistical analysis of Experiment 3. Bolded rows indicate significant effects (at \( \alpha = 0.05 \)). The model specification is provided above the table.

As can be read from Table 6, there was a significant effect of Argumenthood (\( p < 0.001 \)). That is, the 1 Argument Reversed condition yielded significantly fewer target productions compared the 2 Arguments Reversed conditions. There was also a significant main effect of Argument Order for the Reversed vs. Canonical conditions (\( p < 0.001 \)), but not for the Reversed vs. Substitution condition (\( p = 0.190 \)). None of the interactions were significant.
3.2.5 Discussion

Since the Reversed condition was the reference, the main effect of Argumenthood tested for the significance of the difference between the 10% cloze in the Reversed 2 Arguments condition and the 5% cloze in the Reversed 1 Argument condition. This was significant which is most consistent with the Bag of Arguments hypothesis and less consistent with the Bag of Words explanation. That is because even though the canonical agent of the target verb (e.g. 'lifeguard') was present, and not to mention the linearly closest DP to the verb location, its non-argument status led to fewer target responses compared to when it was linearly further away, but crucially an argument.

However, the actual statistical test required to assess if the findings of this experiment are truly the same kind of effect observed in Experiment 2 requires comparing the data from these two experiments together and looking for (the lack of) an interaction. This is done in section 3.4.

3.3 Experiment 4: Argument Removed by Relative Clause Island

Recall, in Experiment 2, it was shown that removing, e.g. 'lifeguard,' significantly decreased the probability that a participant would anomalously produced e.g. 'save.' In Experiment 3 'lifeguard' was reintroduced into the sentence but in a non-argument position. This also led to a significant reduction in anomalous productions of 'saved,' relative to the Reversed 2 Arguments condition, lending support for the Bag of Arguments hypothesis and casting doubt on the Bag of Words hypothesis. One concern, however, is that in sentences such as (31), the DP inside the PP could be coreferent with an argument of the upcoming verb.

31. The parent saw which child near [the lifeguard], had . . . thanked [him|her].

Note that any productions of 'saved' here would still be anomalous and implausible.
However, since this construction allows for the coreference of this non-argument and an upcoming argument, one could argue it is not the cleanest test case; one might want the non-argument entirely blocked from being coreferential with a later argument. This is achieved in Experiment 4.

In Experiment 4, the Bag of Arguments and Bag of Words hypotheses are again pitted against each other, as they were in Experiment 3, but with another syntactic manipulation: syntactic islands. Ross (1967) noted that while wh-elements may ‘move’ to the beginning of the sentence from the position where they are semantically composed (denoted by a $t$) during wh-question formation, as in (32) and (33), there exist syntactic positions where this is not permitted. For example, wh-movement from within a relative clause is not allowed, resulting in the ungrammaticality of (34).

32. Q: What$_1$ did Jackie eat $t_1$ at the restaurant?
   A: Jackie ate tofu$_1$ at the restaurant.

33. Q: What$_1$ did Jackie think the chef overcooked $t_1$ at the restaurant?
   A: Jackie thought the chef overcooked [the tofu]$_1$ at the restaurant.

34. Q: * What did Jackie think the chef who cooked $t_1$ was nice?
   A: Jackie thought the chef who cooked [the tofu]$_1$ was nice.

Moreover, while some have reported that the unacceptability of movement out of islands is partially ameliorated with a ‘resumptive pronoun’ in the place of the wh-phrase’s origin, as exemplified in (35) (Ackerman et al., 2018), the use of such coreferring pronouns is nonetheless still ungrammatical in English despite inflated acceptability (Hammerly, 2021; Meltzer-Asscher, 2021).

35. Q: * What did Jackie think the chef who cooked it$_1$ was nice?
   A: Jackie thought the chef who cooked [the tofu]$_1$ was nice.

This phenomenon is not restricted to matrix wh-questions; islands affect embedded wh-questions equally. Islands thus grant a way to block wh-elements from being argu-
ments of a predicate within the island. Experiment 4 uses relative clause islands for this property, which allows us to have the canonical agent of a verb (e.g. ‘lifeguard’ for ‘save’) present in the sentence but barred from being an extracted argument of such a verb when the verb is in a relative clause. The critical test uses sentences like (36) to see how often participants provide ‘saved’ as their cloze response.

36. The parent saw which lifeguard the child [who \_ \_ \_ island \_ \_ \_] ...

Under the Bag of Words hypothesis, this sentence fragment is identical to its counterpart where ‘lifeguard’ is an argument, thus we should see the same proportion of anomalous cloze responses here as we did in the 2 Argument Reversed condition of Experiments 2 or 3. However, under the Bag of Arguments hypothesis, since ‘lifeguard’ is a non-argument, and cannot corefer with a pronominal argument, it is not used to preactivate verbs, so we should observe significantly fewer anomalous cloze responses for sentence fragments like (36) compared to the 2 Argument Reversed condition in Experiments 2 or 3. Rather, it should pattern with the 1 Argument Reversed condition of Experiment 2.

To foreshadow, Experiment 4 finds that ‘lifeguard’ as a non-argument does in fact influence cloze responses, casting doubt on the Bag of Arguments hypothesis and lending support instead to the Bag of Words. This on the surface contradicts the findings of Experiment 3.

### 3.3.1 Participants

A total of 236 participants were recruited through MTurk. All provided informed consent, were self-reported native speakers of American English and were within the United States. Data from only 106 participants were included after the exclusion criteria were utilized. The same exclusion criteria from Experiments 1, 2 and 3 (described in subsection 2.1.3) were used. Each participant was compensated $4 for their participation.
### 3.3.2 Materials

The same set of 48 sentence fragments as were used in Experiments 1, 2 and 3 were adapted for this experiment. These 48 sentences fragments were distributed across six conditions; an example item is presented in Table 7. The same practice sentence fragments, filler sentence fragments and attention checks as were used in Experiment 1 were used here.

<table>
<thead>
<tr>
<th>2 Arguments</th>
<th>1 Argument (Island)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Canonical</strong></td>
<td>The parent saw which child the lifeguard had ...</td>
</tr>
<tr>
<td></td>
<td>{\textit{saved}</td>
</tr>
<tr>
<td><strong>Reversed</strong></td>
<td>The parent saw which lifeguard the child had ...</td>
</tr>
<tr>
<td></td>
<td>{#\textit{saved}</td>
</tr>
<tr>
<td><strong>Substitution</strong></td>
<td>The parent saw which child the woman had ...</td>
</tr>
<tr>
<td></td>
<td>{\textit{saved}</td>
</tr>
</tbody>
</table>

Table 7: Example stimuli for Experiment 4. The target words were selected by pooling together the Canonical 2 Argument conditions from Experiments 2, 3 and 4. A full set of the materials, including the targets is provided in Appendix A.

These six conditions resulted from crossing two factors: Argument Order and Argumenthood. The first factor, Argument Order, was identical to the manipulation in Experiment 2 and similar to the manipulation in Experiment 3. The only difference between this manipulation and that in Experiment 3 was that in the 2 Arguments Substitution condition, a definite DP (e.g. ‘the woman’) was used rather than a proper name, which was the same manipulation in Experiment 2. The second factor of Argumenthood indicated whether the extracted wh-element could be an argument of the upcoming verb or not by manipulating if the verb was in a relative clause island. In the 2 Arguments conditions, both the extracted wh-element (e.g. ‘which child’ in the Canonical condition)
and the embedded subject (e.g. ‘the lifeguard’) may be arguments of the next upcoming verb after ‘had.’ However, in the 1 Argument conditions, by replacing ‘had’ with ‘who’ a relative clause is introduced which indicates the beginning of an island environment. As discussed earlier, extraction out of this environment is barred, meaning that the wh-element cannot be an argument of a verb within the relative clause. Thus, the embedded subject of the upcoming verb is the sole given argument; the wh-phrase must be an argument of a verb that occurs outside of the relative clause. Thus, this sentence must continue with at least two more verbs, one in the relative clause and one later, outside of the relative clause with which the extracted wh-element can be predicated (e.g. “The parent saw [which child] the lifeguard who . . . saved a toddler helped t1”). That is, the wh-element must be an argument of some verb, however it simply cannot be an argument of the next verb, which is the verb that participants are asked to produce (Chomsky, 1981).

The effects of Argument Order are expected to be the same as they were in Experiments 1, 2 and 3, with significantly more target responses in the Canonical compared to the Reversed conditions. The critical question is the effect of Argumenthood; whether the 1 Argument Reversed condition yields significantly fewer target responses than the 2 Arguments Reversed condition, as was the case for Experiments 2 and 3. This would offer the strongest support for the Bag of Arguments hypothesis.

As with Experiment 3, the actual comparisons that are needed to evaluate the question of what caused the decrease in target productions across the 1 Argument and 2 Arguments Reversed conditions in Experiment 2 are across experiment comparisons. This is discussed in section 3.4.

3.3.3 Procedure

The procedure was the same as in Experiments 1, 2 and 3 (see subsection 2.1.3). Sentence fragments were presented using RSVP, and participants were asked simply to
continue them in the most natural way they could.

After evaluating the exit interviews, 56 participants provided ungrammatical exit responses and were excluded. Further investigation of these participants revealed that 31 of them simply repeated a word from the fragment as their sentence continuation, in a clear demonstration of a failure to follow the instructions in the broader task. After evaluating the attention check trials, 69 participants were excluded for poor performance (more than one error). Five participants performed the experiment more than once, which resulted in the exclusion of each of their attempts. This amount of data loss was considerably higher than previous iterations. This resulted in 106 usable participants.

### 3.3.4 Results

Data were analyzed in R following the same steps as was done for Experiments 2 and 3 (see subsection 3.1.4).
Each response in each condition was assessed for being in the target response set. These trials were coded with a 1; trials in which a non-target response was provided were coded with a 0. The aggregated results are plotted in Figure 8.

As is clear in the figure, the total number of target responses appears to be slightly lower across the board, as compared to Experiments 2 and 3. There was still a clear effect of Argument Order as was the case in the previous experiments. However, the effect of Argumenthood in the Reversed conditions is clearly small.

These data were modeled using a logistic mixed effects model as was done for Experiments 2 and 3. The maximal model converged; the estimates for this model are provided in Table 8.

As can be read from Table 8, the critical effect of Argumenthood did not reach significance ($p = 0.948$); the null that the 2 Argument and 1 Argument Reversed conditions do
Target ∼ Argumenthood * Order + (1 + Argumenthood * Order | Participant) + (1 + Argumenthood * Order | Item)

<table>
<thead>
<tr>
<th>Fixed Effect</th>
<th>Estimate (logits)</th>
<th>z-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>−3.038</td>
<td>−14.905</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Argumenthood</td>
<td>−0.013</td>
<td>−0.065</td>
<td>0.948</td>
</tr>
<tr>
<td>Reversed vs. Canonical</td>
<td>1.632</td>
<td>7.570</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Reversed vs. Substitution</td>
<td>−1.028</td>
<td>−3.517</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Reversed vs. Canonical : Argumenthood</td>
<td>−1.011</td>
<td>−3.529</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Reversed vs. Substitution : Argumenthood</td>
<td>−0.274</td>
<td>−0.825</td>
<td>0.409</td>
</tr>
</tbody>
</table>

Table 8: Summary of the statistical analysis of Experiment 4. Bolded rows indicate significant effects (at $\alpha = 0.05$). The model specification is provided above the table.

not differ with respect to the number of target responses cannot be rejected. This is in contrast to the findings of Experiments 2 and 3, where this factor did lead to a significant decrease in target responses.

### 3.3.5 Discussion

The lack of a significant effect of Argumenthood in Experiment 4 calls into question the Bag of Arguments hypothesis. In the 1 Argument Reversed condition, ‘lifeguard’ is not, and importantly cannot be coreferent with, an argument of the embedded verb. Despite this, anomalous productions of ‘saved’ are roughly equivalent to the 2 Arguments Reversed condition, where ‘lifeguard’ is indeed an argument of the upcoming verb; ‘lifeguard’ as a non-argument of the verb is nevertheless exerting considerable influence over verb preactivations. These findings are indeed more predicted by the Bag of Words hypothesis.

This would appear to be at odds with Experiment 3, where ‘lifeguard’ as a non-argument led to significantly fewer anomalous target productions compared to when ‘lifeguard’ was an argument. As was the case with Experiment 3, the crucial test is a test of across experiment differences which are discussed in the next section.
3.4 Between Experiment Comparisons

As noted in previous sections, the true test of whether the effects of Argumenthood or lack thereof in Experiments 3 and 4 are different from the effect of removing an argument altogether in Experiment 2 involves modeling the data across experiments together and assessing for significant interactions. This analysis is pursued here.

The data from Experiment 2, 3 and 4 were pooled together and a logistic mixed effects model was used to assess the probability of producing a target response as a function of Experiment (2, 3 or 4), Argument Order and Argumenthood. For this analysis, the Substitution conditions were dropped, as they were not identical across experiments (recall Experiment 3 used proper names and had a different configuration for the 1 Argument condition). Thus, there were two levels of Argument Order: Canonical and Reversed. The effect of Argument Order was dummy coded to match the previous analyses with the Reversed condition serving as the reference level. The effect of Argumenthood was sum coded, as was done in the previous analyses. The effect of Experiment was dummy coded with Experiment 2 serving as the reference level. Thus, the first contrast compared Experiment 2 to Experiment 3 and the second contrast compared Experiment 2 to Experiment 4. The crucial statistical test is the two-way interaction between experiment and Argumenthood; if these interactions effects are significant (and positive), it would provide evidence that the presence of a lexical associate in a non-argument position does modulate the preactivation of verbs, counter the Bag of Arguments hypothesis.

The random effects of this model had to be altered in comparison to the previous models that analyzed each experiment individually. Again, no participants participated in more than one experiment. Because of this, by-subjects slopes for the effect of Experiment were not possible. This is not the case for the items, which were used across all three experiments. Thus, the maximal random effects structure is did not feature any random by-subject slopes for the experiment factor. This maximal model was fit first
and converged; the estimates of this model are provided in Table 9.

\[
\text{Target} \sim \text{Experiment} * \text{Argumenthood} * \text{Order} + (1 * \text{Argumenthood} * \text{Order} | \text{Participant}) + (1 + \text{Experiment} * \text{Argumenthood} * \text{Order} | \text{Item})
\]

<table>
<thead>
<tr>
<th>Fixed Effect</th>
<th>Estimate (logits)</th>
<th>z-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-4.059</td>
<td>-13.985</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Experiment 3</td>
<td>0.393</td>
<td>1.617</td>
<td>0.106</td>
</tr>
<tr>
<td>Experiment 4</td>
<td>0.895</td>
<td>4.222</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Argumenthood</td>
<td>-2.971</td>
<td>-7.900</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Canonical</td>
<td>2.764</td>
<td>9.617</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td><strong>Experiment 3 : Argumenthood</strong></td>
<td>1.126</td>
<td>2.597</td>
<td>0.009</td>
</tr>
<tr>
<td><strong>Experiment 4 : Argumenthood</strong></td>
<td>2.995</td>
<td>6.663</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td><strong>Experiment 3 : Canonical</strong></td>
<td>-0.054</td>
<td>-0.187</td>
<td>0.851</td>
</tr>
<tr>
<td><strong>Experiment 4 : Canonical</strong></td>
<td>-0.892</td>
<td>-3.244</td>
<td><strong>0.001</strong></td>
</tr>
<tr>
<td>Argumenthood : Canonical</td>
<td>0.969</td>
<td>2.159</td>
<td>0.031</td>
</tr>
<tr>
<td><strong>Experiment 3 : Argumenthood : Canonical</strong></td>
<td>-1.290</td>
<td>-2.571</td>
<td><strong>0.010</strong></td>
</tr>
<tr>
<td><strong>Experiment 4 : Argumenthood : Canonical</strong></td>
<td>-1.955</td>
<td>-3.261</td>
<td><strong>0.001</strong></td>
</tr>
</tbody>
</table>

Table 9: Summary of the statistical analysis Between Experiments 2, 3 and 4. Bolded rows indicate significant effects (at \( \alpha = 0.025 \)). Highlighted interaction effects are the effects of interest. The model specification is provided above the table.

As can be read from the table, there were significant and positive interactions between the Experiment and the effect of Argumenthood (\( p = 0.009 \), and \( p < 0.001 \) for Experiments 3 and 4, respectively). This suggests that there is evidence that ‘lifeguard’ as a non-argument, either embedded in a PP or blocked by requiring island extraction, yields more anomalous target responses compared to Experiment 2 where it simply was not present. Put plainly, the mere presence of ‘lifeguard’ appears to increase the probability of participants providing ‘saved.’ Both of these suggest that there is some role for a Bag of Words mechanism leading to the anomalous cloze responses we observe across all of the experiments.

Some may worry that too much is being made of the differences from one experiment to the other and that these “errors” are simply too few to be reliable. It is worth remembering that because of the high number of participants, these proportions, while small, reflect a sizable number of actual responses. In Experiment 2, where we observed a 3% target response rate in the 1 Argument Reversed condition, that corresponded to 22
out of 847 responses. The 5% in Experiment 3 corresponded to 43 out of 939 responses; nearly a doubling of the number of anomalous responses. And the 8% in Experiment 4 corresponded to 63 out of 831 responses. These are highly powered experiments with many observations.

There is still a puzzle. In Experiment 4, where 'lifeguard' was blocked from being an argument of the upcoming verb because doing so would require extraction from an island, the effect of argumenthood was not significant. All that can be concluded is that we do not have sufficient evidence to reject the null hypothesis that the 2 Arguments Reversed condition and the 1 Argument Reversed condition in that experiment led to the same number of target responses. But this null finding was not what the Bag of Arguments hypothesis predicted. Thus we have some evidence for the Bag of Arguments and some against it. One more inter-experimental statistical test was carried out too assess if Experiment 3 and 4 differed with respect to the effect of Argumenthood.

The previous model compared Experiment 2 to 3 and 2 to 4. In order to compare Experiment 3 to 4, a separate model was fit which modeled the same dataset but excluding Experiment 2. All of the contrasts remained the same but now the effect of experiment was dummy coded with Experiment 3 as the reference level. The estimates of this model are provided in Table 10. Since this model and the model Table 9 describes look at overlapping data, the alpha level was decreased from the traditional ($\alpha = 0.05$) to ($\alpha = 0.025$) using a Bonferroni correction.

As can be read from this table, the interaction between Experiment and Argumenthood was significant ($p < 0.001$) suggesting that the effect of argumenthood was indeed different between Experiments 3 and 4. Thus, a Bag of Words mechanism can be appealed to to explain the significant jump from 3% to 5% from Experiment 2 to 3. But the significance of the jump from 5% to 8% between Experiment 3 and 4 suggests there is another mechanism at play.

An explanation that allows for this is a mechanism I am calling a Bag of Unassigned
Arguments. It may be viewed as a refinement of the Bag of Arguments mechanism.

The Bag of Unassigned Arguments hypothesis posits that comprehenders preactivate verbs using all active arguments, that is, arguments that have not been composed with a predicate, in a role-independent fashion. What distinguishes this from Chow et al.’s Bag of Arguments hypothesis is that their mechanism does not specify how active arguments that are not or cannot be arguments of the immediately upcoming verb are handled. There appears to be an implicit assumption that the DPs that are prioritized in verb preactivation are the arguments of that verb; other linguistic material (including active arguments of other predicates) is devalued.

One might wonder why it might be that unassigned arguments are held in this privileged status. In particular, one might wonder why a DP that cannot be an argument of the upcoming verb is privileged. Part of the answer can be found in recalling what the island versions of the items in Experiment 4 need in order to continue grammatically. There needs to be a verb in the relative clause island and there needs to be a verb outside of the island environment in order for the wh-phrase to receive a thematic role. The wh-phrase, in the 1 Argument island condition and in the 2 Argument condition, is in an open dependency with a verb that has not yet surfaced, and so it must be held in
memory until it can be composed with its predicate. Only then can it may be discharged from memory. Indeed the other argument in the 1 Argument Island condition and in the 2 Arguments condition (the syntactic subject of the immediately upcoming verb) is also in an open dependency. It too is unassigned to a predicate prior to the verb. Many established principles of sentence processing posit similar things; the comprehender wanting to minimize the load on memory: active gap filling (Omaki et al., 2015) and minimal attachment (Frazier, 1979), to name a few. The Bag of Unassigned Arguments thus fits nicely in this broader picture.

One might also wonder how the Bag of Unassigned Arguments could work in the framework of the model I outlined in chapter 1. At the point in time that the comprehender has encountered arguments that are not yet assigned to a predicate (they are in need of a verb), their incrementally built parse of the incomplete sentence can be used to guide expectations about the event that will ultimately be described. That is, if the comprehender has encountered both ‘lifeguard’ and ‘child,’ and both are in need of predication, the comprehender may begin to conjure up ideas for what event(s) these may ultimately be participants in. In terms of the model presented in Figure 2, information about the sentence fragment is fed up from the parse to the non-linguistic event concept. Importantly, the thematic roles that will are ultimately be assigned to these preverbal arguments is temporarily unknown and there may be multiple parses entertained for the sentence fragment at this time. For each of these parses, there are many candidate events that could be the intended message conveyed by the sentence, and so the comprehender can entertain many different (and sometimes competing) events. With these events in mind, the system works as it does for the lexical predictions derived from sentence compositional meanings. These events contain concepts that have not yet been uttered; these concepts are preactivated in parallel. Among these preactivated concepts is the non-linguistic concept for the action itself which can be lexicalized as a verb. Finally the preactivation of these concepts leads to the preactivation of linguistic
representations further down the hierarchy.

How does the Bag of Unassigned Arguments explain these findings and those established in the literature? In the 2 Arguments conditions (e.g. (37)), which used the same basic structure as the sentences of Chow et al. (2016b), ‘parent’ is a saturated argument, it receives its thematic role from ‘saw,’ and is thus no longer prioritized. ‘Lifeguard’ and ‘child’ however are unsaturated; they have not been composed with a predicate. Accordingly they are used of have a privileged status and are used in a role-unspecified way to preactivate ‘saved’ some of the time. Now, in Experiment 3, with sentences such as (38), ‘parent’ is again largely irrelevant since it is an argument that has already been composed with its predicate. ‘Child’ is an unsaturated argument since it has not been associated with a predicate yet, but ‘lifeguard’ is saturated as it has received its thematic role and been composed with the preposition ‘beside.’ Thus only ‘child’ is in the Bag of Unassigned Arguments. The 2% increase in anomalous ‘saved’ responses may simply reflect the fleeting influence of a Bag of Words style preactivation from ‘lifeguard’ to ‘saved’ (Liao, 2020). Finally, in Experiment 4, with items like (39), ‘parent’ is again already saturated, but now ‘lifeguard’ and ‘child’ are unsaturated, even though only ‘child’ may be an argument of the immediately upcoming verb. Thus they are both privileged.

37. The parent saw which lifeguard the child had saved ...
38. The parent saw which child beside the lifeguard had saved ...
39. The parent saw which lifeguard the child who saved ...

The Bag of Arguments and Bag of Unassigned Arguments make identical predictions in most contexts, with the exception of cases like (39) where being an unsaturated argument of any predicate at all is decoupled from being an unsaturated argument of the very next predicate. Other such exceptions exist, such as sentences with multiple embeddings like (40). Here a transitive verb may come up next but that verb cannot take ‘lifeguard’ as an argument because doing so would create crossing dependencies, which is not permissible in English.
40. The lifeguard that the parent that the child ...

Under the Bag of Unassigned Arguments hypothesis, sentences like this should generate a non trivial amount of preactivation for verbs associated with ‘lifeguards.’ This is an untested prediction to my knowledge, however it is notable that these configurations are notoriously difficult to process (Miller and Isard, 1964; Frazier, 1985) and often ungrammatical versions of them are more readily accepted than truly grammatical ones (Gibson and Thomas, 1999; Häussler and Bader, 2015).

Turning to experiments in the broader literature, most investigations of thematic role reversals have confounded argument saturation with being an argument of the next verb, that is all of the preverbal arguments that had not yet been predicated were, or at least could be, arguments of the next verb (Kim and Osterhout, 2005; Kuperberg et al., 2006; Chow et al., 2016b; Chow et al., 2018; Liao, 2020). There are some findings that demonstrate this mechanism is not the sole reason that there are anomalous cloze responses. One is the interaction between Experiments 2 and 3 in this dissertation. A related finding is that in Liao (2020), where at an SOA of 600ms (24) and (25) led to similar N400 effects, which the authors take to indicate similar levels of preactivation for the anomalous verb. The authors suggest that this is due to the fleeting influence of a Bag of Words mechanism through which ‘Millionaire’ spreads activation to ‘fired’ despite ‘Millionaire’ already being predicated.

24. Millionaire BA servant **fired** ...

25. Millionaire thought servant **fired** ...

This sort of explanation can be invoked for the pair of interactions observed between Experiments 2 and 3 and 3 and 4 in this dissertation. There is a small effect of a Bag of Words mechanism and another effect of a Bag of Unassigned Arguments mechanism and these two conspire to yield anomalous role-inappropriate cloze responses on some trials.
CHAPTER 4

THE ROLE OF COMPETITORS IN PREDICTION

Ehrenhofer et al. (2019) suggested that the Bag of Arguments hypothesis fails to capture the importance of competitors in modulating N400 amplitudes. They found that in sentence contexts such as (41) and (42), the N400 amplitude elicited by anomalous verbs is significantly more negative than the plausible counterpart, resembling a typical N400 effect despite these being simply “role-reversed” sentences.

41. The rancher remembered which cowboy the bull had [#ridden | gored] out on the range.
42. The rancher remembered which bull the cowboy had [ridden | #gored] out on the range.

The presence of an N400 effect here runs counter the findings of Chow et al. (2016b) and others in the literature. Ehrenhofer et al. suggested that this may be due to the fact that their sentences are constraining in both directions; there is a clear modal response in the cloze norming data for both (41) and (42), whereas in the sentences from Chow et al. (2016b), and other similar studies, there was not a strong prediction in the Reversed conditions. Indeed, Experiment 2 of Ehrenhofer et al. (2019) was a within-participants manipulation using half doubly constraining sentences (Ehrenhofer-style sentences) and half sentences that were constraining in one direction only (Chow-style sentences), and they found an interaction between predictability and sentence type; an N400 effect for doubly constraining sentences and not for the other set. The authors leverage this
finding to suggest that there is an important role of lexical competitors that the staged Bag of Arguments hypothesis doesn’t account for.

Ettinger (2018) investigated how the materials in Ehrenhofer et al. (2019) and Chow et al. (2016b) may have differed in a number of respects. After finding that the sentences were comparable along many dimensions, one way in which they were substantially different was the “Subject-Verb Cosine Similarity” measure. This measurement was simply the cosine between the word embeddings (using GloVe; Pennington et al., 2014) of the verb (e.g. ‘ride’) and its subject (e.g. ‘cowboy’ or ‘bull’). The cosine between word embeddings (high dimensional vector representations of words) has been demonstrated to reliably correlate with priming effects (Günther et al., 2016; Auguste et al., 2017). The basic idea behind the creation of this measure was to quantify how strongly each argument in isolation preactivated the verb. Ettinger (2018) found that for the subset of items used in Ehrenhofer et al. (2019)’s within-participant manipulation, the Ehrenhofer-style sentences featured a significant discrepancy between their Canonical and Reversed conditions’ Subject-Verb Cosine Similarity measure; ‘cowboy’ and ‘ride’ were significantly closer together in this high dimensional vector space than were ‘bull’ and ‘ride.’ This however, was not true for Chow et al. (2016b)’s items; ‘waitress’ and ‘serve’ were as close together as were ‘customer’ and ‘serve.’ If this is taken to translate to spreading activation rather plainly, then this discrepancy in the sets of items suggests that when ‘bull’ is encountered in subject position, it will not activate ‘ride’ as much as ‘cowboy’ will. However, after encountering ‘customer’ in subject position, ‘served’ is as activated as when ‘waitress’ is the subject.

A mechanistic explanation is however in order; it is not sufficient to know that Ehrenhofer et al. (2019)’s items differed from Chow et al. (2016b)’s in some respect and led to different N400 patterns without speculating why that detail matters. Ehrenhofer et al. suggested that one possibility is that as comprehenders are progressing through a sentence, at least at each argument, a set of verbs is preactivated in a Bag of Arguments (or
possibly Bag of Words) manner. These activations then decay as time moves on. For example, taking Ettinger (2018)'s cosine metric as a proxy for priming/preactivation and using example (41), encountering ‘cowboy’ may preactivate ‘ride’ but then once ‘bull’ is encountered, ‘ride’ receives little to no boost in preactivation while ‘gore’ does. Since this is more recent, ‘gore’ is more preactivated compared to ‘ride’ at the point in time when the verb is actually encountered. That is, in Ehrenhofer et al.’s items, the anomalous verb may have indeed been preactivated, but a competitor overtook it as the most preactivated verb. This can be contrasted with Chow et al. (2016b)'s items, where an anomalous verb receives initial preactivation, but no verb overtakes it.

One complication is that there is no baseline context or baseline word to compare the anomalous verbs in Ehrenhofer et al. (2019) to. It is entirely plausible that while there is a difference between ‘gored’ and ‘ridden,’ ‘ridden’ is still preactivated relative to its baseline activation. If this is true, its N400 amplitude would be reduced compared to another more neutral word (cf. Federmeier and Kutas, 1999). Under the account just laid out, one would expect to see that the anomalous verbs in (41) and (42) are indeed preactivated, they just aren't the most preactivated verb. This is what Experiment 5 attempts to address.

### 4.1 Experiment 5: A 2AFC Extension of Ehrenhofer et al. (2019)

In many ways, Ehrenhofer et al. (2019)'s results bring up an issue in inferring activations from cloze probabilities. There is not a clear way to transform a word's cloze probability back into that word's activation level, at least without knowing more about the distribution of responses, RTs or a more specified theory of activations. That is, a word may be 20% cloze, say, and have five unit of activation, or it may be 20% cloze and have fifty units of activation. In the first case, if the context is broadly unconstraining, five units of activation may be comparably high leading to some productions of said word, and some productions of other slightly preactivated words. In the latter case, a
very highly active word may only be produced 20% of the time because there exists another word that is dwarfing its activation level.

Experiment 5 does away with this problem by using a “2 Alternative Forced Choice” paradigm (2AFC). Here, we are able to collect responses, analogous to productions, while controlling which words the participant is considering. With this paradigm, we are able to assess the relative preactivation of one alternative over another. Thus this paradigm allows for potentially seeing differences in activations at the very low end of cloze values. If, for instance, two words are zero cloze, but they are differentially preactivated, this task should enable us to see which is truly more active, as it should be selected more than the other.

In this experiment, we investigate whether verbs like ‘ridden’ are nevertheless still preactivated in sentence like (41) as compared to a “neutral” verb like ‘seen.’ To foreshadow the findings, “lures” (e.g. ‘ridden’) were selected more frequently than “neutral” verbs, and additionally when pitted against “targets” (e.g. ‘gored’) they also had a detectable and significant effect, pulling some selection probability away from the target suggesting that these lures are indeed somewhat attractive. The discussion section discusses whether this truly implicates them as being preactivated.

4.1.1 Participants

A total of 95 participants were recruited through MTurk. All provided informed consent, were self-reported native speakers of American English and were within the United States. Ultimately, data from only sixty participants was used after exclusion criteria were utilized (discussed in subsection 4.1.3). Each participant was compensated $3 for their participation.
4.1.2 Materials

A total of 36 sentence fragments and continuations were constructed and distributed across six conditions, as exemplified in Table 11. All 36 items were taken from Ehrenhofer et al. (2019). In addition to the 36 experimental items, there were eight attention check trials. These were trials in which participants were explicitly instructed to press either the ‘F’ or ‘J’ key, and were used for excluding participants.

The 36 experimental items were only a subset of the sixty total items used by the original authors. These were selected based on two criteria. First, if an item’s reported modal response in one of the two directions was a verb particle construction (e.g. ‘run from,’ ‘complain about,’ etc.), that item was dropped. Thus only items which were constraining towards verbs that could stand on their own were left after the first round of filtering. The second criterion was whether a neutral verb could be determined for the item. Neutral verbs were the baseline conditions that were plausible continuations for the sentence fragment regardless of the argument-role mapping/argument order (e.g. ‘see,’ ‘ignore,’ etc.).

<table>
<thead>
<tr>
<th>Order 1</th>
<th>Sentence Fragment</th>
<th>Target vs. Lure</th>
<th>Target vs. Neutral</th>
<th>Lure vs. Neutral</th>
</tr>
</thead>
<tbody>
<tr>
<td>The rancher remembered which cowboy the bull had ...</td>
<td>ridden</td>
<td>gored</td>
<td>seen</td>
<td>gored</td>
</tr>
<tr>
<td>Order 2</td>
<td>The rancher remembered which bull the cowboy had ...</td>
<td>ridden</td>
<td>#gored</td>
<td>ridden</td>
</tr>
</tbody>
</table>

Table 11: Example stimuli for Experiment 5. Participants were presented with the sentence fragment followed by 2 alternative sentence continuations. Sentence continuations that would result in an implausible sentence are marked with a #. A full list of the materials are provided in Appendix A.

These items were distributed across six lists in a Latin Square design. The six conditions were created by crossing two factors: Argument Order (two levels) and Continua-
tion Alternatives (three levels). The factor of Argument Order was somewhat analogous to the Canonical and Reversed conditions of Experiments 1 through 4, however both arguments (‘cowboy’ and ‘bull’ in the example) are canonical agents of their respective target verbs (‘ridden’ and ‘gored’). Accordingly these levels are simply termed Order 1 and Order 2, with no significance placed on which is which.

For Order 1, the targets had an average cloze probability of 35% (standard deviation: 0.14; maximum: 0.69; minimum: 0.13) and the average cloze probability of the lures was 2% (sd: 0.03; max: 0.1; min: 0). For Order 2, these numbers were comparable; the targets had an average cloze of 32% (sd: 0.1; max: 0.61; min: 0.17) and the lures had an average cloze of 0.01 (sd: 0.02, max: 0.09; min: 0). The neutral verbs were not reported to have been produced in any of the conditions in Ehrenhofer et al. (2019); they were all zero cloze.

The other factor of Continuation Alternative was made up of the levels: “Target vs. Lure,” “Target vs. Neutral” and “Lure vs. Neutral.” In each condition, the sentence fragment was followed by two alternatives. In the “Target vs. Lure” condition, one of these alternatives was the “target”, which was the predictable verb (as operationalized by cloze) for that sentence fragment in Ehrenhofer et al. (2019), while the other was the “lure,” the most predictable verb for the other ordering of the arguments in that sentence fragment, which always rendered the sentence implausible if selected. This condition thus pitted the two preactivated options against each other head on. In the next conditions, “Target vs. Neutral,” the alternatives were the “target” and a “neutral” verb that was unpredictable but that nevertheless would create a plausible sentence if selected (e.g. ‘seen’ in the example). This condition acted as baseline to see how much of a propensity there was for participants to select a predictable continuation when a plausible alternative was provided. Finally, the “Lure vs. Neutral” condition featured the “lure” and that same “neutral” verb alternative. Thus, in this condition, participants must select between a word that is predictable under the wrong argument-role mapping and a verb that is un-
predictable, but plausible under the correct argument-role mapping.

If the “Lure” is indeed preactivated, as a kind of lingering effect from encountering its canonical agent as an extracted wh-element, then we would expect to see that it is selected to a somewhat comparable extent in the “Lure vs. Neutral” condition as the “Target” is in the “Target vs. Neutral” condition, though perhaps to a lesser extent. We would also expect to see that in the “Target vs. Lure” condition, the “Target” is selected more frequently than the “Lure” as a conceptual replication of Ehrenhofer et al. (2019)’s findings and demonstrating that the “Target” is indeed more preactivated than the “Lure.”

4.1.3 Procedure

The experiment was again designed using UMass IBEX (Drummond, 2013) and utilized a 2 Alternative Forced Choice (2AFC) design (similar to Staub, 2009, and Sikos et al., 2016). After an experiment information and consent page, participants were told that they would be reading sentence fragments presented to them one word at a time in the middle of the screen (RSVP). As was the case for the previous experiments, each word was present for 400ms before it would be replaced by the following word. The fragment was preceded by a fixation cross + in the center of the screen present for 1200ms. The fragment would end with a 400ms display of “…” followed by a screen that would offer them two alternatives to choose from to continue the sentence. These two alternatives depended on the condition. This screen is schematized in Figure 9. Participants made their selection by pressing either the ‘F’ or ‘J’ key corresponding to the left or right alternative, respectively. The allocation of the ‘target,’ ‘lure,’ and ‘neutral’ verb to the left and right sides was counterbalanced. Participants had 3 seconds to make a selection before the page would timeout and they would be moved on to the next item.
In addition to collecting participants' selections, this procedure also enabled the collection of response times (RTs), which in Experiments 1 through 4 was not an interpretable measure.

Following the experiment, the same exit questions were asked to filter out bots and non-native speakers. The experiment lasted, on average, 27 minutes (maximum: 55 minutes; minimum: 10 minutes).

Before analyzing the data, data quality was assessed using many of the same criteria used in Experiments 1 through 4. Participants' responses to the exit interview questions were read and assessed for grammaticality. Of the 95 participants run, 13 provided ungrammatical responses to the exit interviews and were thus dropped before any further analyses. 1 participant was removed for completing the task twice. The attention check trials were also used for data quality purposes; if a participant failed more than one of these trials (they responded incorrectly ≥25% of the time), they were removed before further analyses took place (16 participants). Finally, if a participant failed to respond on more than 20% of the trials (>8 total trials), they were removed before further analyses took place (five participants). Thus, after these exclusions, data from only sixty participants were ultimately analyzed.
4.1.4 Results

The data were analyzed in R using the tidyverse, lme4 and lmerTest packages. There were two main analyses: an analysis of the selections and an analysis of the RTs.

Trials in which the participant failed to make a selection within the three second window were removed from the analyses of selection accuracy and RTs. The proportion of trials in which this occurred, by condition, is presented in Figure 10. As can be seen, participants were most likely to fail to make a selection in the “Lure vs. Neutral” condition. The “Target vs. Neutral” condition led to the least timeout trials.

![Figure 10: Experiment 5 Timeout Summary by condition. The mean proportion of trials in which participants failed to make a selection within the allotted three seconds is plotted on the y-axis. Error bars represent one by-participants standard error.](image)

The proportion of “correct” selections by condition is plotted in Figure 11. In the Target vs. Lure and Lure vs. Neutral conditions, selecting the Lure was deemed incorrect as it would create an implausible sentence; the other alternative was the correct option. In
the Target vs. Neutral conditions, selecting either resulted in a plausible sentence, but the Target was deemed the correct alternative for the purposes of analyzing the data. It is clear that participants were more likely to select the Target when it was present. It also appears that this propensity is possibly diminished when the Lure is present (the yellow bars are shorter than the green bars). And finally, it appears that in the Lure vs. Neutral condition, participants were actually more prone to incorrect responses (selecting the Lure) than they were to selecting the plausible but unpredictable Neutral alternative. There did not appear to be substantial differences between the Orders. Statistical analyses followed to support these observations.

![Figure 11: Experiment 5 Selection Results by condition. Accuracy is plotted on the y-axis. Error bars represent one by-participants standard error.](image)

The selection data were coded as either 1 for correct alternative selection or 0 for incorrect alternative selection. Accordingly, a Logistic Mixed Effects Model was used to statistically analyze differences between conditions. The contrasts for Argument Order
(Order 1 and Order 2) were sum coded (Order 1 = −0.5; Order 2 = 0.5). The effect of Alternatives was treatment coded with Target vs. Lure as the reference condition. This meant that the first level compared the ‘Lure vs. Neutral’ condition to the Target vs. Lure condition, effectively testing if having the correct response be predictable (i.e. the Target) significantly increased the probability that a participant would select it. This is referred to as the Target effect on Lure Selections. It corresponds to the red vs the yellow bars in Figure 11. The other level compared the ‘Target vs. Neutral’ condition to the ‘Target vs. Lure’ condition; effectively testing if the presence of the ‘Lure’ significantly altered participants’ propensity to select the Target. This effect is referred to as the Lure effect on Target Selections.

Again, following the advice of Barr et al. (2013), the maximal random effect structure was used in constructing the model. This model converged; no random effects had to be dropped. The estimates of this model are presented in Table 12. This model confirmed what was suggested by Figure 11. There was no significant effect of Order (p = 0.102), as was to be expected; these items were constructed to be equally constraining in both directions. The intercept was significant (p < 0.001), meaning that in the Target vs. Lure condition, participants were significantly more likely to select the Target alternative. Both of the effects of Alternatives were significant (ps < 0.001). The Target effect on Lure Selections being significant demonstrated that participants were significantly more likely to make a correct selection when doing so meant choosing the Target over the Lure as compared to the Neutral option over the Lure. To reframe this in terms of the Lure, the Lure was a tempting alternative; participants selected it ~60% of the time in the Lure vs. Neutral condition. However, the presence of the Target significantly increased the likelihood that participants would overcome the temptation to select the Lure. The Lure effect on Target Selections being significant demonstrated that the presence of the Lure did indeed lead to less accurate behavior; participants were significantly less likely to select the Target when the competitor was the Lure compared the Neutral verb.
ther of these effects significantly interacted with Order.

$$\text{Correct} \sim \text{Order} \times \text{Alternatives} + (1 + \text{Order} \times \text{Alternatives} | \text{Participant}) + (1 + \text{Order} \times \text{Alternatives} | \text{Item})$$

<table>
<thead>
<tr>
<th>Fixed Effect</th>
<th>Estimate (logits)</th>
<th>z-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>0.636</td>
<td>4.507</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Order</td>
<td>−0.457</td>
<td>−1.633</td>
<td>0.102</td>
</tr>
<tr>
<td>Target effect on Lure Selections</td>
<td>−1.1490</td>
<td>−7.125</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Lure effect on Target Selections</td>
<td>0.659</td>
<td>3.979</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Target effect on Lure Selections : Order</td>
<td>0.381</td>
<td>1.210</td>
<td>0.226</td>
</tr>
<tr>
<td>Lure effect on Target Selections : Order</td>
<td>0.307</td>
<td>0.932</td>
<td>0.351</td>
</tr>
</tbody>
</table>

Table 12: Summary of the statistical analysis of Experiment 5 selections. Bolded rows indicate significant effects (at $\alpha = 0.05$). The model specification is provided above the table.

RTs were also analyzed. RTs were split by condition and also by selection type (correct and incorrect). The mean RTs for each selection type within each condition are plotted in Figure 12. The plot suggests that the Lure vs. Neutral conditions possibly had inflated RTs compared to the other two. Another visually salient pattern in Figure 12 is that the more common selection tended to have lower RTs compared to the less common selection within each condition. That is, in the Target vs. Lure and Target vs. Neutral conditions, when the Target was selected (a correct response), it was generally done more quickly than the competitor. However, in the Lure vs. Neutral conditions this pattern means that when the Lure (the incorrect alternative) was selected it was done quickly.
Figure 12: Experiment 5 Response Time (RT) Summary by condition. Time (ms) to make a selection is plotted on the y-axis. Red-outlined bars represent incorrect selections; black-outlined bars represent correct selections. Error bars represent one by-participants standard error.

As these RT values were continuous (though values greater than 3000 were not allowed; those trials would have resulted in a timeout), RTs were analyzed with a Linear Mixed Effects Model. In addition to the factors of Order and Alternatives, the effect of Selection Accuracy was now included as a predictor of RT; it was sum coded (Correct = −.5; Incorrect = .5). The model with the maximal random effects structure did not converge. The correlations between random effects were however not removed as there is reason to think that the accuracy of an item may depend on the attractiveness of the Neutral alternative which may influence RTs. Instead, the random effects that accounted for the least variance were removed successively until the model converged. The estimates from this model are provided in Table 13.

The analysis revealed that the interaction between Order and the Target effect on Lure
RT ∼ Order * Alternatives * Correct + (1 + Order + Alternatives * Correct | Participant)

<table>
<thead>
<tr>
<th>Fixed Effect</th>
<th>Estimate</th>
<th>t-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1596.27</td>
<td>31.617</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Order</td>
<td>56.30</td>
<td>1.688</td>
<td>0.092</td>
</tr>
<tr>
<td>Target effect on Lure Selections</td>
<td>59.34</td>
<td>1.972</td>
<td>0.054</td>
</tr>
<tr>
<td>Lure effect on Target Selections</td>
<td>−29.56</td>
<td>−1.024</td>
<td>0.311</td>
</tr>
<tr>
<td>Correct</td>
<td>−48.04</td>
<td>−1.256</td>
<td>0.212</td>
</tr>
<tr>
<td>Target effect on Lure Selections : Order</td>
<td>−31.23</td>
<td>−0.687</td>
<td>0.4923</td>
</tr>
<tr>
<td>Lure effect on Target Selections : Order</td>
<td>−35.36</td>
<td>−0.744</td>
<td>0.4572</td>
</tr>
<tr>
<td>Correct : Order</td>
<td>47.70</td>
<td>0.724</td>
<td>0.469</td>
</tr>
<tr>
<td>Target effect on Lure Selections : Correct</td>
<td>142.13</td>
<td>2.488</td>
<td>0.016</td>
</tr>
<tr>
<td>Lure effect on Target Selections : Correct</td>
<td>−34.80</td>
<td>−0.602</td>
<td>0.549</td>
</tr>
<tr>
<td>Target effect on Lure Selections : Correct : Order</td>
<td>35.59</td>
<td>0.385</td>
<td>0.700</td>
</tr>
<tr>
<td>Lure effect on Target Selections : Correct : Order</td>
<td>−16.20</td>
<td>−0.168</td>
<td>0.866</td>
</tr>
</tbody>
</table>

Table 13: Summary of the statistical analysis of Experiment 5 RTs. Bolded rows indicate significant effects (at α = 0.05). The model specification is provided above the table.

Selections was significant (p = 0.016). That is, while the correct responses had shorter RTs in the Target vs. Lure condition compared to the incorrect responses, this relationship did indeed switch in the Lure vs. Neutral condition; the incorrect responses had shorter RTs. None of the other interactions reached significance. Additionally, none of the main effects were significant. While Figure 12 suggested that the Lure vs. Neutral condition had inflated RTs compared to the Target vs. Lure condition, the Target effect on Lure Selections just failed to reach statistical significance (p = 0.054). It is also worth noting that the effect of Order also neared statistical significance (p = 0.092), though there was no a priori reason to expect such a difference.

### 4.1.5 Discussion

Experiment 5 provided a number of takeaways. First, it was found that Lures, role-anomalous verbs, are indeed tempting continuations in the sentences used in Ehrenhofer et al. (2019). Lures were selected more frequently than neutral verbs, and the presence of the Lure as an alternative significantly decreased the probability a participant selected the “correct” Target response. Additionally the RT data showed that when par-
Participants selected the Lure, they did so faster than when they selected the neutral verb, but slower than when they selected the Target. The main Target effect on Lure Selections effect was only marginally significant.

These data can be taken to suggest that the Lures were in fact preactivated. However an alternative linking hypothesis between 2AFC selections and activations makes this less clear. Two linking hypotheses are spelled out in turn below.

One way to think about the 2AFC task is as a “constrained cloze task.” That is, participants treat their task as essentially a cloze task, but rather than running a “race” in their minds to produce the winner (the lexical item that accrues the most activation the fastest), they evaluate the relative activations of the two alternatives they are presented with and select the more active one. A way to formalize this decision is Luce’s Choice Rule, or a softmax function, provided in Equation 4.1, where the probability of a selection is that selection’s activation over the sum of the activations of the two alternatives.

\[ P(a_i) = \frac{a_i}{\sum_j a_j} \]

(4.1)

With this linking hypothesis in mind, the results fall out as follows. First the Target is about about three times as active as the neutral alternative (∼75% vs ∼25%) and 1.5 times as active as the Lure (∼60% vs ∼40%). And crucially the Lure is about 1.5 times as active as the neutral verb (60% vs 40%). Since there were only alternatives, the ratio of these proportions is analogous to the ratio of the activations between the alternatives. Using this linking hypothesis there is clearly more activation of the Lure than the neutral alternative, which was originally obscured by them both sharing effectively zero cloze, and which was impossible to discern from the ERPs in Ehrenhofer et al. (2019) as there was no comparable baseline to compare the Lure to (compare to Federmeier and Kutas, 1999).

However, another linking hypothesis has been proposed; one where preactivation plays a much less central role in 2AFC selections. The 2AFC task can be thought of as a
one-shot Maze Task (Forster et al., 2009). The Maze Task is a psycholinguistic task that has participants iteratively construct a sentence by selecting either of two options, one of which is grammatical and the other not. Forster et al. (2009) argue that the Maze Task is a task that taps into integration difficulty, i.e., that the RTs and selections reflect the time that it takes to integrate the current word into the prior context rather than reflecting lexical access per se. It is worth noting that frequency effects, which are not usually taken to be effects on integration (Reichle et al., 2009), have been detected using the Maze Task (Forster et al., 2009). This suggests that there may be some impact of lexical access time in Maze Task performance. Indeed, lexical access ought to precede integration, so even if the Maze Task primarily reflects integration difficulty, some of the measure should reflect the ease of earlier subprocesses such as lexical access, which is generally thought to be modulated by preactivation (Reichle et al., 2009).

There are clear parallels with the 2AFC task used here; the key differences are that the sentence was provided to the participant up until the embedded verb; they did not have to make any selections prior to the target region, and the alternatives did not differ in grammaticality, but rather plausibility. Even so, it is plausible that this integration linking hypothesis is a fitting description of the 2AFC task used in Experiment 5. Under the most extreme form of this linking hypothesis, the results of Experiment 5 are cast in a new light. The Target is easier to integrate than the neutral verb and the Lure, since it was selected more frequently and with shorter RTs. However, the Lure (which again is implausible), would appear to be easier to integrate than the neutral verb as it was selected more frequently and with shorter RTs.

It is however hard to reconcile this last point with the fact that the Lure is actually the implausible alternative in the Lure vs. Neutral condition. Essentially, this amounts to saying that participants are either largely unaware of the implausibility of the Lure, or feel as though the Lure is truly a better fit in the context than neutral verb, after perhaps corrupting the representation of the context Gibson et al. (2013). This last interpretation
of the data is dispreferred however, as it would most straightforwardly predict longer RTs for the anomalous selections in the Lure vs. Neutral condition. That is, if that condition features the participant making some rational inferences about the intended sentence to accommodate the Lure, this process presumably takes some time. The first possibility, that participants often simply fail to notice the implausibility of the Lure, is a live interpretation of the data, and further investigation of the 2AFC task would be needed to rule it out.

We may now turn to comparing the results more directly with the N400 patterns observed in Ehrenhofer et al. (2019). Recall in Ehrenhofer et al. (2019) there was a detectable N400 effect observed when participants were presented with the Lure; the Lure elicited a more negative N400 than the Target did. This is quite analogous to the current findings. Specifically, the significance of the intercept in Table 12 indicates that the Target was more likely to be selected than the Lure was. However, the Lure was apparently still tempting, as indicated by the significance of the Lure effect on Target Selections. This second finding, combined with the finding that the Lure was more likely to be selected than the neutral alternative (Target effect on Lure Selections), adds to the findings of Ehrenhofer et al. (2019) by offering a baseline to compare the Lure to, and suggests that it is indeed a tempting alternative as a sentence continuation, which, depending on the linking hypothesis of the task, may implicate it in being preactivated.

The main theoretical takeaway of Experiment 5 is that even in the exceptional cases of role-reversal contexts in which there is the typical predictability N400 effect, reported in Ehrenhofer et al. (2019), there is evidence that the role-anomalous verb is still preactivated. It is not as though these sentence contexts fully de-activated these Lures; they remained more active than neutral verbs that were plausible but not strongly associated with the context. However, strong conclusions are difficult to draw given the ambiguity of the measure in this task. Preactivation may play a large role or a much smaller one, and without knowing what the correct linking hypothesis between 2AFC selections and
activations is, these findings are only suggestive.
CHAPTER 5

PREDICTABILITY EFFECTS IN NATURAL READING

In chapter 3, I proposed that the Bag of Arguments mechanism proposed by Chow et al. (2016b) should be amended to be a Bag of Unassigned Arguments. This was to account for the fact that we observed the same degree of argument role reversal cloze responses for sentences where the embedded wh-element was barred from being an argument of the verb by means of an island violation.

As noted earlier, there are known discrepancies between measures of predictability. The cloze experiments in chapter 3 were in principle untimed, and may allow for the participants’ subjective probability distributions to look different than they would in other tasks such as natural reading. Accordingly, to test if the Bag of Unassigned Arguments is indeed deployed rapidly enough to influence language comprehension in a natural setting, an eye tracking while reading experiment was performed.

This methodology was chosen over the ERP paradigm for a few reasons. First is the degree of certainty around the linking hypothesis of early reading measures such as the first fixation duration and lexical activations. The first fixation is believed to index the speed with which a word is identified (Reichle et al., 2009; Staub, 2015; Burnsky et al. under review). In the framing of Staub (2011), predictable words are more quickly recognized due to a “head start” on processing; there is simply less activation required to reach threshold. The first fixation does not appear to index later processes like integration of a word into context (Warren et al., 2008). The interpretation of the N400 is simply
less clear.

Secondly, there have been very few eye tracking while reading investigations of these sorts of illusions; this literature has been dominated by ERP experiments. Weiss et al. (2018) used materials similar to those in Kim and Osterhout (2005), and found that the region containing the role reversed verbs (‘is picking’ in (43)) received shorter first pass fixations than they the region ‘is drawing’ does in (44) suggesting anomalous ease with which ‘picking’ is recognized. However, they also found an effect of plausibility such that the first pass reading times for ‘picking’ in (43) was indeed slower than in (45) where there is no anomaly. Unfortunately, the authors do not report skipping behavior nor the first fixation durations, which are the two measures least susceptible to plausibility effects.

43. #On the sunny afternoon, the flower is picking the girl ...
44. #On the sunny afternoon, the flower is drawing the girl ...
45. On the sunny afternoon, the girl is picking the flower ...

The authors also found a significant interaction effect for the first pass times in the spillover region (‘the girl’ | ‘the flower’), such that the association between the verb and its subject lead to speed ups on the verb, but slowdowns on the object. That is, in (43), while ‘is picking’ elicited short first pass times (as compared to ‘is drawing’ in (44)), this measure was inflated in the spillover region ‘the girl,’ potentially as the effect of noticing the anomaly in this condition comes online.

Chow (2013) used materials similar to those in Chow et al. (2018). They are presented again below.

22. Police BA suspect ZAI last week arrest and bring back to the station.
23. #Suspect BA police ZAI last week arrest and bring back to the station.

Across three reading experiments, they find mixed results. In one of these experiments, there is a significant effect of the reversal on the first fixation duration: ‘arrest’
received a shorter first fixation in (22) compared to (23). This contradicts the N400 pattern of equivalence between the canonical and reversed conditions. In the other two experiments, there was no effect of the reversal on the first fixation duration. One reason for the mixed findings is the relatively small sample sizes. The number of participants in these experiments were 24, 24 and 36, with 15 items per condition. With only one out of three experiments finding the effect, it is worth further investigating the possible eye tracking-N400 discrepancy with a more powerful eye tracking experiment. Additionally, the eye tracking experiments in Chow (2013) used Mandarin sentences; testing this effect in English test the generalizability of these effects.

With these issues in mind, a higher-powered eye tracking while reading experiment is a crucially missing piece of the puzzle for establishing the true magnitude of the online preactivation of role-anomalous and role-appropriate verbs in role reversal contexts. This also serves as an online test of the Bag of Unassigned Arguments proposal from chapter 3.

5.1 Experiment 6: An Eye Tracking Extension of Experiment 4

The goals of Experiment 6 were twofold. First, this experiment could assess if the Bag of Unassigned Arguments mechanism postulated in chapter 3 is deployed quickly enough to impact reading times. Second, this experiment is aimed at assessing if reading times, and in particular the First Fixation Duration, pattern more like the cloze data or more like the N400 data in role-reversal contexts. To preview the results, Experiment 6 shows that the Bag of Unassigned Arguments makes the correct predictions about the first fixation duration; there is no significant interaction between argumenthood and argument reversal. And moreover, the eye tracking data pattern strikingly mirrors that of the ERPs with no detectable effect of the reversal on early measures (skipping and the First Fixation Duration), but a later emerging effect in Go Past Time, analogous to the N400 neutralization and the P600 effect.
5.1.1 Participants

A total of 95 participants participated in this experiment. All participants were undergraduate students at the University of Massachusetts, Amherst, participating in exchange for Psychology course credit. All provided informed consent, were native speakers of American English and had no history of language impairments or disorders.

Data from only 84 participants were included after the exclusion criteria were utilized. A participant was removed from the analysis if they performed poorly (< 75%) on the comprehension questions and plausibility judgments on the filler trials. There were 32 such filler trials; a participant had to answer 24 or more of them correctly. This resulted in the removal of 11 participants. Additionally, if nine or more trials overall, or four or more trials within a single condition, had sufficient track loss, the participant was removed. This criterion was not met by any participants; none were removed due to excessive track loss.

5.1.2 Materials

The experimental materials were adapted from Experiment 4. There were 48 experimental items distributed across six conditions in a 2 x 3 design crossing Argument Order and Argumenthood. An example item is provided in Table 14.

These sentences were produced by appending the modal target response to the sentence fragments from Experiment 4. In order to have spillover regions and in order for the 1 Argument Island conditions to unfold grammatically, continuations were constructed to follow the target verb. In the 2 Arguments conditions, these took the form of two PPs denoting a location and time at which the described event took place. For the 1 Argument Island conditions, there was an object provided for the verb in the RC Island (e.g. ‘the swimmer’) and then a verb phrase relating the extracted wh-element to the argument heading the relative clause (e.g. ‘helped’). Thus, all of the experimental
2 Arguments

The parent saw which child the lifeguard had saved in the pool on Saturday.

Reversed

#The parent saw which lifeguard the child had saved in the pool on Saturday.

Substitution

The parent saw which child the woman had saved in the pool on Saturday.

Table 14: Example stimuli for Experiment 6. The underlined word is the target verb. Stimuli were presented as a single line of text. A full set of the materials is provided in Appendix A.

sentences were grammatical and all featured spillover regions that could be analyzed. Within an item, the spillover regions were the same within each level of Argumenthood, as is shown in Table 14.

The modal response from the Canonical 2 Arguments condition from Experiments 2, 3 and 4 was the target verb. The mean cloze values for the target verb were thus lower than for the target verb set used in Experiment 3. This is because in Experiments 2, 3 and 4, there were multiple target responses that were counted toward the target cloze values. The mean cloze values in each condition for the target verb presented in this experiment were: Canonical 2 Arguments = 23%; Reversed 2 Arguments = 4%; Substitution 2 Arguments = 6%; Canonical 1 Argument = 13%; Reversed 1 Argument = 4%; Substitution 1 Argument = 2%.

Following the presentation of each experimental item, participants were asked to provide a plausibility judgment. For the Canonical and Substitution conditions, the correct answer was ‘YES’ indicating that the event described in the sentence is indeed something that is likely to happen in the real world. For the Reversed conditions, the correct answer was defined as ‘NO.’ It is worth noting that, especially in the 1 Argument Island condition, the Reversed sentences did vary in their plausibility. That is, a child very well
could save a swimmer and help a lifeguard. This is a necessary consequence of keeping
the spillover regions the same across conditions. However, in the 2 Arguments Reversed
conditions, these are classic role reversal sentences and describe unlikely events.

There were also 66 filler sentences for a total of 114 trials per participant. These took
a number of different forms. Sixteen were part of a subexperiment, not described in this
dissertation, investigating noun-noun compounds. These were all followed by compre-
hension questions. Eighteen were part of a subexperiment, also not a part of this disser-
tation, investigating active gap filling. The remaining 32 sentences were fillers designed
to partially balance out other experimental factors and to filter out participants for poor
performance on non-experimental questions. Four of these fillers were simple plausi-
ble sentences followed by a plausibility judgment. 12 of them were simple implausible
sentences followed by a plausibility judgment. These are imbalanced because the exper-
imental conditions feature more plausible than implausible sentences, thus, by adding
these the overall distribution of correct responses in plausibility judgments is closer to
50% (the actual distribution is 54% of the questions have plausible as the correct an-
swer and 46% have implausible as the correct answer). Another four filler sentences
were plausible sentences that featured an embedded wh-question (as our experimental
items do) but followed by a comprehension question. And finally, the last 12 filler sen-
tences were implausible sentences with an embedded wh-question that were followed
by a comprehension question. These were put in so that participants could not know a
priori whether they would receive a comprehension question or a plausibility judgment
following sentences with an embedded wh-question (the experimental items).

5.1.3 Procedure

Participants had the movements of their right eye recorded using an Eyelink 1000 (SR
Research, Toronto, ON, Canada) eyetracker. The sentences were presented on a monitor
placed 55 cm in front of the eyetracker. The experiment was deployed using the UMass
Prior to the experiment, the eyetracker was calibrated using three-point calibration in a horizontal line across the middle of the screen. Each participant had a mean calibration error < .5°.

Participants were instructed to read the sentences silently to themselves for comprehension and to answer any questions that followed the sentences using a game controller. The experiment lasted approximately 45 minutes.

5.1.4 Results

The data were preprocessed using RoboDoc and EyeDry. These read in the ASC files produced during the recording session and parsed out fixations and their locations in predefined regions of interest (ROIs). These ROIs are exemplified in (46) and (47) where the “|” symbol marks the boundaries of each region. Leading whitespace was included in the ROI.

46. The parent saw| which child| the lifeguard| had| saved| in the pool| on Saturday.|

47. The parent saw| which child| the lifeguard| who had| saved| the swimmer| helped to the shore.|

The data were then analyzed in R. The measures of interest for this experiment were the first fixation duration, the probability of skipping, the first pass time and the go past time. The first fixation duration is the span of time between when the reader’s eye lands in a region for the very time and when they launch a saccade, even if this saccade is to simply refixate the same word or region. The first pass time is a similar measure: it is the span of sum of the fixation times between when the reader’s eye lands in a region for

1EyeTrack, along with the preprocessing programs: EyeDry and RoboDoc, is available at https://blogs.umass.edu/eyelab/software/
the very first time and when they launch a saccade but to a new region. Time spent in a saccade is not included; only fixation time is counted. Refixations upon the same region thus appear in first pass times, but not the first fixation duration. Finally, the go past time is the sum of the fixation times between when the reader’s eye landing in a region for the very time and when they move on in the sentence by fixating a region to the right of the current region. That is, if the reader makes a regressive eye movement by returning to a previously read portion of the sentence, the go past time takes into account all of the time spent in those regions as well as time spent on the current region. Visualization of the means of these measures by condition, for the target verb, are provided in Figure 13.
As can be seen in the plot, there is little difference between the conditions within each level of Argumenthood in the early measures of Skipping and the First Fixation Duration. Differences between the Canonical and the Reversed conditions appear to only begin surfacing in later measures, especially Go Past Time.

The primary analyses dealt with reading measures on the target verb region (underlined in the examples). First, the data were analyzed under the Null Hypothesis Sig-
Significance Testing Framework using linear (and logistic) mixed effects models as implemented by \texttt{lme4} (Bates et al., 2015) and \texttt{lmerTest} for p-values (Kuznetsova et al., 2017). A Bayesian analysis followed using \texttt{brms} (Bürkner, 2017). The frequentist analysis is reported first.

Since four eye tracking measures are being assessed, the alpha level for the frequentist analyses was reduced using a Bonferroni correction to control against an inflated Type I error rate (Von der Malsburg and Angele, 2017). Rather than using the typical $\alpha = 0.05$, an $\alpha$ of $0.0125$ ($0.05 / 4$) was used.

The probability of skipping was modeled using a logistic mixed effects model since the dependant variable is binary. The effect of Argument Order was dummy coded with the Canonical condition as the reference. The first level of the Order effect corresponded to the difference between the Canonical and the Reversed conditions; the second corresponded to the difference between the Canonical and the Substitution conditions. The effect of Argumenthood/Island was sum coded with the 2 Arguments condition as ($-0.5$) and the 1 Argument Island condition as ($0.5$). The maximal model was fit first, however after failing to converge, the random effects that accounted for the least variance were successively dropped until the model converged. The estimates from the model are summarized in Table 15.

\begin{table}[h]
\centering
\begin{tabular}{l|ccc}
\hline
Fixed Effect                                      & Estimate (ms) & z-value  & p-value  \\
\textit{Intercept}                                & -1.912       & -13.584  & < 0.001  \\
\textit{Argumenthood}                             & 0.369        & 2.052    & 0.040    \\
\textit{Reversed}                                 & -0.084       & -0.768   & 0.443    \\
\textit{Substitution}                             & -0.148       & -1.338   & 0.181    \\
\textit{Reversed : Argumenthood}                  & -0.225       & -1.027   & 0.305    \\
\textit{Substitution : Argumenthood}              & -0.067       & -0.304   & 0.761    \\
\hline
\end{tabular}
\caption{Summary of the statistical analysis of the probability of skipping the target for Experiment 6. Bolded rows indicate significant effects (at $\alpha = 0.0125$). The model specification is provided above the table.}
\end{table}

As can be gleaned from the table, there were no significant effects on skipping be-
behavior (all p's > 0.03). Under the typical $\alpha$ of 0.05, the effect of Argumenthood would be significant ($p = 0.040$), however this effect is not significant given the adjusted $\alpha$ of 0.0125 following Von der Malsburg and Angele (2017). This effect is the direction of more skips of the target verb in the 1 Argument Island condition as compared to the 2 Arguments condition.

The first fixation duration was modeled using a linear mixed effects model. The contrasts were the same as for the model of the skipping data. The estimates from the model, and its final random effects structure, are summarized in Table 16.

First Fixation Duration $\sim$ Argumenthood * Order + (1 | Participant) + (1 | Item)

<table>
<thead>
<tr>
<th>Fixed Effect</th>
<th>Estimate (ms)</th>
<th>t-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>235.284</td>
<td>54.729</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Argumenthood</td>
<td>−12.942</td>
<td>−2.767</td>
<td>0.006</td>
</tr>
<tr>
<td>Reversed</td>
<td>2.940</td>
<td>0.893</td>
<td>0.372</td>
</tr>
<tr>
<td>Substitution</td>
<td>4.104</td>
<td>1.249</td>
<td>0.211</td>
</tr>
<tr>
<td>Reversed : Argumenthood</td>
<td>−0.028</td>
<td>−0.004</td>
<td>0.997</td>
</tr>
<tr>
<td>Substitution : Argumenthood</td>
<td>7.368</td>
<td>1.121</td>
<td>0.263</td>
</tr>
</tbody>
</table>

Table 16: Summary of the statistical analysis of the first fixation duration on the target for Experiment 6. Bolded rows indicate significant effects (at $\alpha = 0.0125$). The model specification is provided above the table.

The model summarized in Table 16 found a significant main effect of Argumenthood ($p = 0.006$). The 1 Argument Island condition was in general read faster as compared to the 2 Arguments condition. No other main effects were significant and none of the interactions were significant (all p's > 0.2).

The first pass time was also modeled using a linear mixed effects model. Contrasts were the same as the previous models. The model’s estimates and the model specification are summarized in Table 17.

The first pass data resemble the first fixation data considerably. There was a single significant main effect of Argumenthood, with the 1 Argument Island condition leading to shorter reading times compared to the 2 Arguments conditions ($p < 0.001$). No other effects were significant (all p's > 0.1).
The statistical analysis of the Go Past Times is summarized in Table 18.

Go Past Time $\sim$ Argumenthood $\times$ Order + (1 + Argumenthood | Participant) + (1 | Item)

<table>
<thead>
<tr>
<th>Fixed Effect</th>
<th>Estimate (ms)</th>
<th>t-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>397.24</td>
<td>15.106</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Argumenthood</td>
<td>35.89</td>
<td>1.103</td>
<td>0.270</td>
</tr>
<tr>
<td>Reversed</td>
<td><strong>59.00</strong></td>
<td><strong>2.576</strong></td>
<td><strong>0.010</strong></td>
</tr>
<tr>
<td>Substitution</td>
<td>30.69</td>
<td>1.343</td>
<td>0.179</td>
</tr>
<tr>
<td>Reversed : Argumenthood</td>
<td>6.49</td>
<td>0.142</td>
<td>0.887</td>
</tr>
<tr>
<td>Substitution : Argumenthood</td>
<td>−2.96</td>
<td>−0.065</td>
<td>0.948</td>
</tr>
</tbody>
</table>

Table 18: Summary of the statistical analysis of the go past time of the target for Experiment 6. Bolded rows indicate significant effects (at $\alpha = 0.0125$). The model specification is provided above the table.

For the go past time, a new pattern emerged. The effect of Argumenthood was no longer significant ($p = 0.270$). The effect of Canonical vs Reversed was now significant ($p = 0.010$), with the Reversed conditions leading to almost 60ms longer go past times as compared to the Canonical conditions. The effect of Substitution was not significant ($p = 0.179$) and neither of the interactions were significant either (both $p$'s $> 0.8$).

As an interim summary, the results from the frequentist analyses indicate that the early measures (skipping and the first fixation) yielded no “predictability effect.” That is, neither of the effects of Reversed nor Substitution were significant, despite the cloze differences obtained in Experiment 4. The effect of Reversed only reached significance in the later measure of go past time. Recall that in previous experiments from Kim and
Osterhout (2005) to Liao (2020), the N400 amplitude elicited by the role reversed verbs was not significantly different in the Canonical and Reversed conditions. Here, we see a fundamentally similar pattern for the first fixation which is the measure of the most interest; the reading times of these verbs in these two conditions do not significantly differ suggesting they are similarly preactivated in both the Canonical and Reversed contexts. What’s more, this is a relatively high-powered experiment with 84 participants; if there were an effect the size of typical predictability effects, it would very likely have been detected (this experiment had 86% power).\(^2\) It is possible that previous estimates of the magnitude of the predictability effect are larger than what one should have expected to surface here since the difference in cloze probabilities is less in the current experiment, which features a roughly 20% difference in cloze values between the Canonical and Reversed conditions. That is, the cloze values are not as different in this experiment as in some others, making the effect harder to detect in general.

There was however a significant effect of Argumenthood on these early measures, with the 1 Argument Island conditions featuring marginally more skips of the target verb and shorter reading times of the target verb when it was fixated. While this may seem mysterious, an explanation presents itself through Figure 14. Analyses of the pre-target and spillover regions are reported in Appendix B.

\(^2\)Power was calculated using G*Power (Faul et al., 2007), assuming the use of a by-subjects t-test (which is admittedly not the planned statistical test). The estimates of the predictability effect on the first fixation from Rayner et al. (2011) were used since both means and variances were provided.
Figure 14: Experiment 6 Go Past Times by Region. Error bars represent 1 by-participants standard error.

Likely because of the regioning of the sentences, the 1 Argument Island conditions have longer go past times for the precritical region (‘who had’ vs ‘had’). What’s more, these are short and frequent function words, which are presumably relatively easy to access. If the reader spends longer fixating the pretarget region, and is more likely to fixate it to begin with, they may garner some information about the target verb while it is in their parafoveal vision. In the 2 Arguments conditions, readers were likely to skip the single pretarget word (‘had’) quite often and thus were less capable of previewing the target word, as they would be doing so from at least two words back while fixating a content word, either ‘lifeguard,’ ‘child,’ or ‘woman,’ which likely require more attention and increases “foveal load” (Henderson and Ferreira, 1990; Hohenstein and Kliegl, 2014). Of course, since the regions are defined differently, this is simply speculation. Moreover, the main effect of Argumenthood was not of direct interest, and this finding and the corresponding explanation does not have bearing on the questions this experiment was
aimed at addressing.

Additionally, in none of the measures did the interaction between Argumenthood and Order reach significance. This null finding is consistent with the Bag of Unassigned Arguments; in both the 2 Arguments Reversed condition and the 1 Argument Island condition where 'lifeguard' was not an argument of the verb 'saved,' the target verb was as preactivated as in the Canonical counterparts, save for the main effect of the Argumenthood factor.

Additionally, if attention is guided again to Figure 14, it can be seen that in the pretarget region of the 2 Arguments conditions ('had') there is an apparent difference between the conditions such that the Reversed condition is leading to inflated reading times. This effect is not significant, however the current experiment was not designed to detect such an effect if does truly exist. This would be an interesting finding, as it would demonstrate an effect of the reversal that is occurring before the target word that ultimately makes the sentence anomalous, which could in principle arise due to the fact that the Canonical and Reversed conditions are different in the two regions prior; e.g. 'lifeguard' and 'child' appear in opposite positions and non-canonical positions. What's more, Figure 14 also shows that in the spillover region, the Reversed and Substitution 2 Arguments conditions yield quite inflated reading times. This is borne out as significant main effects of Reversed and Substitution; the interactions did not reach significance. Readers appear to move past the target region ('saved') but encounter cognitive difficulty moments later, presumably when integration of 'saved' into these contexts produces world knowledge violations or otherwise unlikely events. It is also worth noting that this pattern mirrors the results from the ERP literature of a P600 effect between Canonical and Reversed conditions, despite the N400 equivalence. As has been argued to explain the P600 effect, the significant effect of Reversed on Go Past Times, both on the target region itself and in the spillover region, suggests that the participants do experience online difficulty with the implausibility of the Reversed sentences. The spillover effect on Go Past Times in
the Substitution condition is perhaps a different kind of effect. In the Substitution conditions, the syntactic subject of the embedded verb (‘woman’ in the example) does not signal to the participant that they are reading about a situation involving a pool or any body of water. The reader only learns this information upon reading the spillover region. This is informationally rich for the reader to construct a more fleshed out picture of the event they are reading about and it is possible that encountering this is what leads to a slowdown in this region for the Substitution conditions. It is also possible that this is a predictability effect of sorts in the Canonical and Reversed conditions; cloze norms were not gathered for this region and it is plausible that “in the pool” is simply less expected in the Substitution condition.

Finally, one might worry that the failure to find any predictability effect is indicative of a problematic design or overall poor data quality. A post-hoc subset analysis of the items for which the cloze of the target verb in the Substitution conditions was zero (twenty out of the 48 items), revealed a marginally significant effect of Substitution (p = 0.093). While this is closer to a significant effect compared to when all of the items were included, it does not cross the desired α level. This may simply reflect the fact that these conditions differ with respect to cloze probabilities by only 20 to 30%, which is a more modest difference compared to many of the predictability manipulations in the literature, which typically feature differences closer to 50 to 60% (Rayner et al., 2011; Frisson et al., 2017; Staub and Goddard, 2019).

Since the early eye tracking measures failed to find a significant effect of Reversed, instead of simply failing to reject the null, Bayesian analyses of the first fixation duration were also performed using the brms package in R (Bürkner, 2017). The first Bayesian analysis used flat, uninformed priors. The estimates of this model are provided in Table 19. The model used four chains with 10000 iterations (2000 iterations for burn-in). A Region of Practical Equivalence (ROPE) was defined around 0 ranging from −3 to 3. An effect of 3ms is considered small enough to be theoretically equivalent to no effect.
The amount (density) of the posterior distribution falling within the ROPE was used to determine if there practically is no effect (Vasishth et al., 2018; Kruschke and Liddell, 2018).

First Fixation Duration $\sim$ Argumenthood * Order + (1 + Argumenthood * Order | Participant) + (1 + Argumenthood * Order | Item)

<table>
<thead>
<tr>
<th>Fixed Effect</th>
<th>Estimate (ms)</th>
<th>$-$95% HDI</th>
<th>+95% HDI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>235.16</td>
<td>226.23</td>
<td>243.85</td>
</tr>
<tr>
<td>Argumenthood</td>
<td>$-$12.91</td>
<td>$-$22.53</td>
<td>$-$3.22</td>
</tr>
<tr>
<td>Reversed</td>
<td>3.01</td>
<td>$-$4.13</td>
<td>10.12</td>
</tr>
<tr>
<td>Substitution</td>
<td>4.02</td>
<td>$-$3.49</td>
<td>11.48</td>
</tr>
<tr>
<td>Reversed : Argumenthood</td>
<td>$-$0.05</td>
<td>$-$13.80</td>
<td>13.85</td>
</tr>
<tr>
<td>Substitution : Argumenthood</td>
<td>7.19</td>
<td>$-$6.42</td>
<td>20.76</td>
</tr>
</tbody>
</table>

Table 19: Summary of the Bayesian statistical analysis of the first fixation duration of the target for Experiment 6 with uninformed priors. Bolded rows indicate effects where the 95% HDI falls outside the ROPE. The model specification is provided above the table.

As can be read from the table, the estimate for the effect of Reversed is just outside the ROPE at 3.01ms. 50% of the posterior distribution either falls within the ROPE or is on the other side of $-$3; an anti-predictability effect. This is taken to be further support for the non-effect and essential equivalence of the Canonical and Reversed conditions with respect to First Fixation Durations of the target word.

The second analysis used priors taken from Staub (2020), who set the prior on the effect of predictability on the first fixation duration to be normally distributed around 20ms with a standard deviation of 10ms (i.e. the priors for Reversed and Substitution were $\mathcal{N}(20, 10)$); the other predictors had normal priors centered at 0ms with a standard deviation of 10ms (i.e. $\mathcal{N}(0, 10)$).

As can be gathered from Table 20, even with informed priors that considerably pull the parameter estimates higher towards the expected predictability effect, some of the 95% HDI for the effect of Reversed still falls within the ROPE and indeed includes zero itself. In particular, 22% of the HDI is either included in the ROPE or is less than $-$3. In sum, the Bayesian analyses supplement the frequentist analyses in demonstrating...
First Fixation Duration ~ Argumenthood * Order + (1 + Argumenthood * Order | Participant) + (1 + Argumenthood * Order | Item)

<table>
<thead>
<tr>
<th>Fixed Effect</th>
<th>Estimate (ms)</th>
<th>−95% HDI</th>
<th>+95% HDI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>233.42</td>
<td>224.77</td>
<td>242.01</td>
</tr>
<tr>
<td>Argumenthood</td>
<td>−10.10</td>
<td>−17.78</td>
<td>−2.29</td>
</tr>
<tr>
<td>Reversed</td>
<td>5.65</td>
<td>−1.08</td>
<td>12.36</td>
</tr>
<tr>
<td>Substitution</td>
<td>6.72</td>
<td>−0.25</td>
<td>13.71</td>
</tr>
<tr>
<td>Reversed : Argumenthood</td>
<td>−2.04</td>
<td>−12.74</td>
<td>8.67</td>
</tr>
<tr>
<td>Substitution : Argumenthood</td>
<td>3.71</td>
<td>−6.80</td>
<td>14.18</td>
</tr>
</tbody>
</table>

Table 20: Summary of the Bayesian statistical analysis of the first fixation duration of the target for Experiment 6 with informed priors. Bolded rows indicate effects where the 95% HDI falls outside the ROPE. The model specification is provided above the table.

Moderate support for a null effect of the Reversed conditions, mirroring the N400 results of Chow et al. (2016b) and others.

Additionally, the plausibility judgments themselves were analyzed. Mean accuracy by condition is plotted in Figure 15. To reiterate, the correct response for the Canonical and Substitution conditions was ‘YES’ (plausible), and for the Reversed conditions, the correct answer was ‘NO.’ All trials are included in the plot and analysis below, even those for which the participant did not fixate the target word. When the analysis is restricted to only trials where the participant did fixate the target, which removes approximately 16% of the total trials, the qualitative and statistical patterns did not change.
As can be seen in the figure, there is a marked decline in accuracy for the Reversed conditions. Roughly 50% of the time, participants responded correctly (that these described implausible events), and 50% of the time they responded inaccurately. This may be contrasted with their generally better performance on the plausibility judgments following the Canonical sentences.

To analyze participants’ performance statistically, a logistic mixed effects model was fit to the data. When the maximal model failed to converge, the random slope that accounted for the least variance was successively dropped until convergence. The estimates from the resulting model are provided in Table 21.

The model corroborated the visual pattern in Figure 15. Participants responded significantly less accurately in the Reversed conditions compared to the Canonical conditions (p < 0.001). They were also less accurate in the Substitution condition (p = 0.003)
Correct ~ Argumenthood * Order + (1 + Argumenthood | Participant) + (1 + Argumenthood + Order | Item)

<table>
<thead>
<tr>
<th>Fixed Effect</th>
<th>Estimate (ms)</th>
<th>z-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1.632</td>
<td>10.780</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Argumenthood</td>
<td>-1.294</td>
<td>-6.337</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Reversed</td>
<td>-1.852</td>
<td>-11.134</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Substitution</td>
<td>-0.398</td>
<td>-2.932</td>
<td>0.003</td>
</tr>
<tr>
<td>Reversed : Argumenthood</td>
<td>1.111</td>
<td>5.349</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Substitution : Argumenthood</td>
<td>0.148</td>
<td>0.691</td>
<td>0.490</td>
</tr>
</tbody>
</table>

Table 21: Summary of the Statistical Analysis of the Performance on the Plausibility Judgments for Experiment 6. Bolded rows indicate significant effects (at $\alpha = 0.05$). The model specification is provided above the table.

and less accurate in the 1 Argument Island conditions ($p < 0.001$).

Of the most interest here is the degradation in the Reversed condition. The magnitude of this difference diverges considerably from previous work (Ferreira, 2003; Chow et al., 2016b; Chow et al., 2018). These studies have reported only roughly 25%, 25% and 10% errors, respectively, as compared to the slightly more than 50% errors obtained here. One possibility is that the diversity of question types (comprehension questions and plausibility judgments) led to a different, and possibly more ecologically valid, comprehension strategy. In the experiments listed above, participants had only one task and one type of question asked of them: either naming the agent or patient of the sentence, or giving plausibility judgments. On the other hand, Meng and Bader (2021) had participants provide both a plausibility judgment and name the agent/patient in the sentence and they obtained plausibility judgment accuracies much closer to these: roughly 40% in the comparable condition. Thus, it stands to reason that since the participants in the current experiment do not know ahead of time what kind of question they may receive after reading the sentence, they attend less solely to features that are more helpful in one task over another, in particular, role-argument mappings which are the feature that make many implausible sentences implausible.

Finally, by collecting plausibility judgments after every trial, we may analyze the eye
tracking data conditioned on responses. That is, the eye tracking data can be split into trials where the participant answered correctly and incorrectly, and their response may be used as a predictor (in the statistical sense) of their preceding reading behavior. This data can shed light on the nature of the relatively poor accuracy observed for the plausibility judgments in the Reversed conditions. One could imagine, for instance, that incorrect responses were preceded by faster reading times on the target word and perhaps faster reading of the sentence as a whole. This pattern would occur if readers are being particularly careless on these trials. On the other hand, one could imagine that incorrect responses were preceded by markedly slow reading of the target word. This data pattern would arise if the reader is engaged in a costly and conscious revision or correction process akin to that described within the “noisy channel” framework (Levy et al., 2009; Gibson et al., 2013).

The eye tracking measures for the Reversed conditions conditioned on response type are summarized visually in Figure 16. The Canonical and Substitution conditions were not included as the error rates were not as high, and they are of less theoretical importance.

For a post-hoc analysis of the reading measures conditioned on response type, linear mixed effects models for the First Fixation Duration, First Pass and Go Past Times for the target word were fitted using response as a predictor. A logistic mixed effects model was used to analyze the probability of skipping the target word. Response was sum coded with the correct as (−.5) and incorrect as (.5). Again, only data from the Reversed conditions were included in these models. These analyses are summarized in the tables in Appendix B.
Despite the numerical trends present in the majority of the panels in the figure, where incorrect responses are preceded by longer reading times of the target word, for none of the measures was there a significant main effect of Response on the reading times of the target word (all p’s > 0.25). That is, reading behavior was actually quite similar between trials where the participant correctly indicated that these sentences were implausible and trials where they mistakenly responded that they were plausible. One
issue is that the current experiment was not designed explicitly to test this effect and it is possible that this is simply underpowered, and the numerical trends visible in the plot are a small but real effect. Future research can follow up on this.

5.1.5 Discussion

There are several key takeaways from Experiment 6. First, and perhaps most importantly, is the lack of a difference in early reading measures, like the first fixation duration, between the Canonical and Reversed conditions. This pattern mirrors that of the ERP experiments discussed throughout this dissertation which demonstrate equivalent N400 amplitudes elicited by the target verb in these kinds of sentences. Recall, this was not a given since the N400 and first fixation do not always covary (Ledoux et al., 2007; Kretzschmar et al., 2015; Burnsky et al. under review). This finding corroborates the claim that 'saved' is similarly preactivated in both the Canonical and the Reversed contexts.

A second major takeaway is the support for the Bag of Unassigned Arguments mechanism in the form of the null interaction between argument order and argumenthood. In no measure did this interaction approach significance, though the primary data predictions were for the first fixation duration. Additionally, the Bayesian analyses estimated this effect to be 0ms and -2ms which are within the ROPE (-3ms to 3ms). Thus it appears that there is an influence of non arguments (e.g. ‘lifeguard’ in the Reversed 1 Argument Island condition) leading to anomalous preactivation of ‘saved’ to a comparable degree as when it is an argument of the verb.

A third finding of interest is the effect of the Reversed condition on the later measure, Go Past Time. This indicates some awareness of the anomaly online and is reminiscent of the P600 effect observed in ERP experiments. But a related finding goes against the literature. Participants’ accuracy in the plausibility judgments following Reversed sentences in particular was rather low. While previous experiments have reported errors for closer to 25% of trials, the data here put the error rate at just over 50%.
There is a clear discrepancy between the findings of the cloze experiments in chapter 3 and the current findings. On the one hand, the first fixation data suggest that role-appropriate ‘saved’ in the Canonical condition and role-anomalous ‘saved’ in the Reversed condition are equally preactivated; on the other, there is a significant difference in their cloze probabilities with role-appropriate ‘saved’ being produced roughly 40% of the time and role-anomalous ‘saved’ being produced only about 10% of the time. Adding to this is the N400 data, which also suggests equal preactivation. Recall the linking hypotheses of these measures spelled out in chapter 1: cloze responses are productions of the most active lexical unit, the first fixation is modulated by the preactivation of word forms and the N400 is modulated by the preactivation of conceptual representations. It stands to reason that if the N400 data and the first fixation data are covarying, there ought to be a corresponding pattern in the cloze data, yet there is not. Thus some aspect of these linking hypotheses should be reevaluated. This invites thinking more deeply about the second main question that this dissertation is attempting to answer: how can we bridge methodological divides between tasks that are used to study lexical predictability in sentence processing?

There are a few possible solutions, which I will discuss in turn. First is the idea that cloze responses do indeed straightforwardly reflect preactivations, but the other measures index not only preactivation, but also (at least some) ease of integration of the current word into its sentence context. This will be termed the “Integration Hypothesis.”

This logic of this explanation is however unintuitive if not at times contradictory. If cloze responses reflect preactivation, while the other first fixation duration and the N400 reflect ease of integration, then the discrepancy between the Canonical and Reversed conditions in the cloze experiments suggests that role-appropriate ‘saved’ is preactivated much more than role-anomalous ‘saved.’ In short, argument-role mappings would appear to be used effectively for the generation of lexical predictions. There is
also no denying that predictability influences reading times and the N400. Thus, for the first fixation and the N400 to show no effect whatsoever, despite the cloze difference, then there would need to be an equal and opposite effect of integration, such that role-anomalous ‘saved’ is easier to integrate into the sentence than role-appropriate ‘saved;’ enough to perfectly counter the predictability effect. This is already an unintuitive prediction. To make matters worse, if role-anomalous ‘saved’ were indeed easier to integrate into the sentence context, that would amount to saying that despite the effective use of argument-role mappings to preactivate lexical items, argument-role mappings are not effectively used in composing sentence meanings.

All in all, an explanation of the current set of findings that postulates the first fixation equivalence in Experiment 6 is due to anomalously easy integration of role-anomalous ‘saved’ into its context runs up against extant prior research and theoretical considerations, making this explanation undesirable.

Another way to make sense of the current data patterns is to instead hold onto the idea that the N400 and the first fixation are more or less direct reflections of preactivation, and question the linking hypothesis of the cloze task. Another solution proposed by Lee et al. (2022) is that participants in the cloze task do run a sort of “race” to produce an active lexical item, but there is a filtering or monitoring process that blocks a number of anomalous cloze responses. That is, ‘saved’ is highly active in Reversed contexts, but in the cloze task, participants check whether this is a good continuation before producing it, and often it is flagged for being anomalous and thus another word is produced in its place. This will be referred to as the “Cloze Filter” hypothesis.

Lee et al. (2022) disprefer this explanation in light of RT data they obtained. They collected vocal cloze data which allowed them to analyze the speech onset latency (SOA) of the responses and found that when participants did produce an anomalous cloze response (roughly 5% of the time with their materials), they did so with slower SOAs compared to role-appropriate responses. They interpreted this to mean that the filter-
ing or monitoring hypothesis is no longer live, as it could be cached out to predict faster
SOAs for anomalous responses, as role-appropriate responses would be an amalgama-
tion of responses that win the race and pass the check and responses that were runner
up winners of the the race after an anomalous word first won the race (a process which
presumably takes time). However, this is not the only way in which this RT data pat-
tern could arise if there is a filtering/monitoring process at play. Presumably, monitor-
ing one's possible cloze responses requires attention. If a participant's attention lapses
for a given trial, the filter would be less effective at flagging anomalous responses. Addi-
tionally, if a participant is not allocating a lot of attention to the task at hand, they may
be overall slower on those very same trials. In fact this pattern of inflated RTs for errors,
and fewer error than correct responses, was observed in Staub (2009). Here, this pat-
tern was taken to suggest that participants were experiencing genuine confusion about
how a sentence should continue (with respect to subject-verb agreement with a non-
intervening attractor). This reasoning may be imported over to Lee et al. (2022)'s data;
if the participant occasionally experiences lapses of attention which causes confusion
about how the sentence should in fact continue, one would expect to see slow errors.
Additionally, the linking hypotheses of the filter and RT data are not fully spelled out;
it is not clear which subprocesses are expected to take considerable time: the flagging
of anomalous productions or the generation of runner up productions. Thus, the filter-
ing/monitoring hypothesis of the cloze task need not be discarded because of these RT
results.

An interesting facet of this proposal is that it essentially casts the cloze task as having
a plausibility judgment sub-task. While the “Integration” hypothesis postulated that the
first fixation and N400 (partially) reflect integrating the current word into its context, and
failing to notice the anomaly, the “Cloze Filter” hypothesis postulates that integration is
essential for the cloze task and here the anomaly is regularly detected. Thus another
difference between these two proposals is that integration is either quite error-prone
(the “Integration” hypothesis), possibly relying on a Bag of Unassigned Arguments-type mechanism for evaluating plausibility, or integration is quite reliably accurate (the “Cloze Filter” hypothesis). On the face of it, the plausibility judgment data from Experiment 6 suggest that integration, or at least interpretation, may be quite error-prone, as nearly 50% of the time, participants responded inaccurately in the Reversed conditions. However, this should be taken with a grain of salt; it reflects end-of-sentence plausibility judgments rather than judgments of plausibility immediately after encountering the target word. It could very well be that the anomaly is in fact readily detected, but post-interpretive processes lead to erroneous end-of-sentence responses (Paolazzi et al., 2019; Meng and Bader, 2021).

Finally, another possible solution is that all three measures (the N400, the first fixation and cloze responses) do reflect preactivations, but they are reflections of different underlying distributions of upcoming words. In particular, cloze responses reflect words that are preactivated and are fitting as the next word, where fitting may be construed as perhaps being of the right syntactic category (Dell et al., 2008; Momma et al., 2020), but this need not be a complete plausibility filter. In fact it need not be a filter at all; it could be a boost in activation to words of the correct syntactic category (Dell et al., 2008; Gaston, 2020). It is possible that the predictability effect on the N400 and first fixation reflect the preactivation of words that the participant thinks are likely to appear at some point in the current context. This will be called the “Sooner or Later” hypothesis.

Trials in reading experiments, ERP experiments and cloze experiments all begin in very similar ways; the participant is engaging their comprehension system to analyze and interpret the sentence they are presented with. Throughout the course of the comprehension process, they generate expectations about the situations and events that are being described (Kuperberg and Jaeger, 2016). These predicted events lead to the expectations of their component conceptual parts: the actors, the action, etc. which may in turn be lexicalized as a corresponding lexical item. Importantly, events are unordered...
objects; there is no inherent linear order with which the actors in an event, or other components of events, are available. For instance, in a circumstance where a sentence begins as in (48), the event in the speaker’s mind surely involves a vessel for this coffee (e.g. a ‘mug’), it may very well invoke a morning setting, they may mention that they like to sweeten their coffee with sugar, and finally this event surely culminates in the drinking of coffee. If the comprehender has this similar situation in mind, there are a number of lexicalizations that may be preactivated as a consequence of constructing this event and scene (e.g. ‘mug,’ ‘morning,’ ‘sugar,’ ‘drink,’ etc.). Indeed, encountering any of these words would be unsurprising in this context.

48. I take my coffee with cream and …

However, now is when the tasks diverge. In the cloze task, the participant must consider the position in the sentence where there is a blank. Thus, there is a forced linearization of these predicted concepts and words, such that saying any preactivated word could result in obvious ungrammaticality, perhaps by being of the wrong syntactic category. Participants in the cloze task must restrict themselves to good candidates for being the next word. In reading and ERP experiments, the participant may rely on these “sooner or later” predictions for upcoming material, rather than the restricted subset of good next words. There is in fact some independent evidence for this. Nieuwland and Van Berkum (2005) for instance found that when listeners are given a context about a tourist on an airplane with an oversized suitcase, and an airline agent, ‘tourist’ and ‘suitcase’ elicited equivalent N400 amplitudes in (49).

49. Next, the woman told the {tourist | #suitcase} …

Importantly, out of context and in isolation, ‘suitcase’ elicited a more negative N400 compared to ‘tourist.’ One might worry that the contexts featured these words and so the finding could be a kind of repetition priming (which happens to exactly counter the
normal N400 effect). However, in Metusalem et al. (2012) sentences such as (5) were used and the contexts featured no such repetitions.

5. They spent the whole day outside building a big \{snowman | #jacket | #towel\} …

Here, while the N400 elicited by ‘jacket’ was more negative than that of ‘snowman,’ ‘jacket’ elicited a significantly reduced N400 compared to ‘towel.’ This is of course despite the fact that ‘jacket’ is zero cloze. What ‘jacket’ and ‘suitcase’ in the previous example have in common is that they are both event/situation-compatible. They are likely activated because they are natural conceptual components of the scenario being described by the speaker and they very well may be encountered at some point in the broader context. However, since they are not plausible next words, they are rarely, or actually never, produced in the cloze task.

One way to think about this more explicitly is to invoke Levy (2008)’s “Surprisal Theory.” Under the standard surprisal account, most sentence processing effects can be thought of as a reflection of the comprehender’s distribution of likely words. Words have differing activations which may be transformed into probabilities, via a softmax function. Importantly, the surprisal account postulates that this distribution shifts for each incoming word. However, these distributions need not shift entirely from word to word; the activations need not be completely flushed at each new time step. It is conceivable, and studies suggest, that comprehenders keep some concepts and words active over spans of text exactly because they expect to encounter them not as the very next word, but eventually. Again, Stone et al. (2020) suggest that there is actively maintained activation of verbs in German verb-particle constructions. On the point of activations not resetting to baseline instantaneously, Rommers and Federmeier (2018) demonstrated that there is a “pseudo-repetition priming” effect, a significantly reduced N400, elicited by the once preactivated, but not actually encountered, word ‘hot’ in the neutral second sentence in (50).
50. Be careful, the top of the stove is very dirty ['hot is high cloze']. The proofreader asked her to replace the word hot.

In fact this pseudo-repetition priming effect on the N400 was elicited with two intervening sentences between the first sentence, where the prediction for ‘hot’ is generated, and the target sentence where it appears. All in all it appears that there is not a radical resetting of lexical activations for each incoming word. Taken together, it stands to reason that the N400 and first fixation may be more tuned to these sooner-or-later predictions and lingering preactivations from predictions made earlier in a way that the cloze task is not, because the cloze task forces the participant to consider position-specific predictions. This shares a number of characteristics with the “Cloze Filter” hypothesis, however rather than there being a filter that checks the winner of a cloze race, there is instead a broad boost in activation spread to words that are appropriate as the next word, which leaves many already preactivated items active, just relatively less so. This kind of mechanism has been independently argued for to explain why many speech errors, like substitutions and transpositions, are grammatical, even if they result in implausible sentences (Dell et al., 2008).

There is also a practicality to predicting upcoming linguistic material without initially making commitments to when exactly said prediction should be expected to surface. This is because there are an infinite number of ways for most dependencies to be separated and comprehenders cannot know whether a verb predicted in light of a preverbal argument will be preceded by a relative clause modifying that argument, or a series of adverbs, etc. That is, the predicted verb could easily surface in n+2 words or n+12 words. Local syntactic context can inform position specific predictions, but this need not negate a prediction for a verb that is predicted to surface eventually.

To connect the “Sooner-or-Later” hypothesis to chapter 3 and the lemma-inspired model from chapter 1, one can think of predictions from the Bag of Words, Bag of Unassigned Arguments and sentence compositional meanings as “sooner-or-later” predic-
tions. Of course, as demonstrated in chapter 3, these mechanisms generate predictions to varying degrees. Nevertheless, since these mechanisms work by simple association or via the event concept space, these mechanisms preactivate concepts in an unordered fashion. In comprehension-only tasks there is not the same pressure to linearize these unordered, “sooner-or-later” predictions as in the cloze task or other production tasks.

How do the data patterns in Experiments 2, 3 and 4 work under these various hypotheses? As a reminder, these three experiments demonstrated that roughly 10% of the time, participants produce a role-anomalous verb in the Reversed conditions when the canonical agent of said target verb (‘lifeguard’ in all of the examples) is in need of a predicate to be associated with. These anomalous productions drop to 3% when lifeguard is not present at all, and drop to 5% when lifeguard is present but not in need of a predicate.

The “Integration” hypothesis states that ‘saved’ in the Reversed conditions is not as preactivated as the N400 or eye tracking results suggest. This is because the N400 and eye tracking results are reflecting easy integration of ‘saved.’ There are then two issues this hypothesis must address: why are there significant differences between the experimental manipulations in Experiments 2 through 4, and why is ‘saved’ easy to integrate (mechanistically)? On the first question, the solution is actually the one already proposed in chapter 3: cloze responses reflect the usage of various mechanisms, there is a small role for a Bag of Words (simple associative priming) mechanism, a more sizable role for a Bag of Unassigned Arguments mechanism and a much larger role for a mechanism that uses sentence compositional meanings to generate continuations. To address the second question, one possibility is that integration too is guided by these less sophisticated Bag of Words and Bag of Unassigned Arguments mechanisms; a kind of quick and dirty as opposed to careful and veridical integration. In order there to be equivalence between the Canonical and Reversed conditions, as there is for the N400 and First Fixation, these would need to be the only (detectable) mechanisms that guide integra-
tion. This obviously can't be the end of the story; the Reversed sentences are ultimately
detectable as implausible, which necessitates veridical integration and composition in
accordance with the syntax and semantics of the language.

The “Cloze Filter” hypothesis states that 'saved' is indeed preactivated equally in the
Canonical and Reversed conditions (as shown by the N400 and eye tracking data), it
is only provided less in the Reversed condition in the cloze task because a monitoring
process often catches the error. There exists a tension then regarding why this process
fails more in the Reversed 2 Arguments and Reversed Island conditions as compared
to the other Reversed conditions; why are there differences between experimental ma-
nipulations? Again one could say that this filter is simply imperfect; it perhaps catches
anomalous productions that are active from irrelevant syntactic positions more often
than those have been preactivated using more relevant pieces of the sentence (e.g. pre-
verbal arguments preactivating a verb). This is to say, the filter too has a gap driven by a
Bag of Unassigned Arguments type mechanism.

The “Sooner or Later” hypothesis states that ‘saved’ is preactivated equally in the
Canonical and Reversed conditions (the N400 and eye tracking data). However, the role-
anomalous version of ‘saved’ becomes comparably less active when participants are
asked to consider position-specific predictions, which leads to the reduction in cloze re-
sponses. This hypothesis can make sense of the differences between experimental ma-
nipulations by again adopting some of the Bag of Unassigned Arguments logic. When
a participant encounters the blank position in Experiments 2 through 4, they must re-
strict their responses to verbs. What's more, they have have many verbs already preacti-
vated, some because they need to predicate 'the child' and some because 'the lifeguard'
too needs to be associated with a predicate, and in the 2 Arguments conditions a sin-
gle predicate may satisfy both of these needs. However since there is an immediate and
unavoidable need to provide a verb to predicate the syntactic subject ('the child' in the
Reversed conditions), verbs fitting this are boosted. All the while 'saved' has been pre-
activated because it is fitting within a broader discourse context and is a fine word to 
show up sooner-or-later. ‘Saved’ additionally meets the requirement of being a verb that 
can predicate ‘the child.’ Thus it receives enough of a boost to sometimes be uttered. 
This doesn’t happen as much in the 1 Argument conditions of Experiments 2 or 3 be- 
cause ‘saved’ simply was not preactivated to the same extent to begin with as the main 
generating associate (‘lifeguard’) was either not present at all or present only in a syn-
tactically irrelevant position in the sentence (in a PP) with no need for involvement in 
the situation or events being discussed.

This explanation is preferred, though only slightly. Partly this is on the grounds of 
parsimony. Under this account, the Bag of Words and Bag of Unassigned Arguments 
are simply mechanisms to preactivate items, and not shortcuts used for integration. 
Additionally, this explanation allows us to mostly maintain the linking hypotheses laid 
out in chapter 1; the only additional machinery is the boost in activation for position-
appropriate items (Dell et al., 2008), and the idea of predictions not being tied to specific 
positions, which again upon reflection, seems likely independently.

The data from the experiments reported so far in this dissertation leave each of these 
hypotheses open. Though it is worth reemphasizing that there is perhaps more grounds 
to doubt the “Integration” hypothesis in light of the eye tracking data in Experiment 6. 
Future work can be designed to arbitrate between the “Cloze Filter” and the “Sooner-or-
Later” hypotheses. One possible way to tease them apart would be to come at this from 
the N400 or eye tracking side. A possible experiment could use items similar to those in 
Metusalem et al. (2012), where there is a broader discourse context about a snow day 
and children playing in the snow, and then they receive (51).

51. They spent the whole day outside building a big {snowman | #shoveling | #squint-
ing} . . .

The critical comparison would be between ‘shoveling,’ a verb that is associated with 
the general situation of a snow day, and ‘squinting,’ a verb that isn’t associated with the
situation. Both ‘shoveling’ and ‘squinting’ would presumably be zero cloze, and would be lead to integration difficulty, as the parse breaks upon encountering them.\(^3\) Thus, the “Integration” hypothesis would predict the N400 elicited by them to be equivalent and both more negative than the N400 elicited by ‘snowman.’ A strict interpretation of the “Cloze Filter” hypothesis would also predict this pattern, as neither ‘shoveling’ nor ‘squinting’ should be active at this position as no verb can grammatically be here. However, under the “Sooner or Later” hypothesis, ‘shoveling’ should elicit a reduced N400 compared the ‘squinting’ since it is a valid prediction for a word that may eventually surface in this context.

Another possible way to tease apart these hypotheses would be to modify the cloze task. First, one could quite explicitly test the “Sooner or Later” hypothesis by instructing participants to simply provide a word that they believe may eventually show up in the sentence they are reading. That is, rather than asking them to provide the word they believe will be next, in the place of the blank, they can provide any word they believe the sentence may contain even if it is a poor fit for the place of the blank; essentially freely associate. If participants can do this task (it would likely feel unnatural), their responses may align more with the N400 and eye tracking data. Similarly, one could attack the “Cloze Filter” hypothesis by diverting attention away from the filter using a distractor task. One could perhaps sandwich cloze task trials between simple math problems, as in (52), or have participants monitor for and have to report mispellings in the sentence fragment, as in (53), both of which detract attention from being able to monitor one’s productions and make cloze responses look more like the N400 or eye tracking results. It is not clear that any of the other hypotheses predict that these manipulations would influence cloze responses.

52. 5 + 4. The parent saw which lifeguard the child had … Cloze: ___ Math: ___

\(^3\) These examples may not be the best cases as ‘shoveling’ and ‘squinting’ could be read as adjectives in which case the parse does not fail. They are still, however, different with regards to their coherence in the discourse.
53. The parent saw which lifeguard the child had . . . Cloze: ___ Typo: ___

These experiments could service the field by finding a way to align our measures of lexical preactivation and in doing so inform our understanding of the correct linking hypotheses between these measures and the representations and processes of interest.

The idea of the sooner-or-later predictions is quite novel and invites some exploration of more specified models. Language Models, which necessarily provide position specific predictions, have no attested mechanism for interference from sooner-or-later predictions. In the next chapter, a large modern Language Model, GPT-2 (Radford et al., 2019), is probed for its sensitivity to these reversals as a test case of the extreme opposite hypothesis of the “Sooner or Later” hypothesis.
As mentioned in chapter 1, Language Models provide probabilities for next words. In recent years, Language Models have significantly improved, enough to break through to popular media (Markoff, 2020; Johnson, 2022). These breakthroughs are attributable to the size of the models (the number of parameters), the amount of web corpus training data that can be combed through, and also the Transformer architecture (Vaswani et al., 2017). One such model is GPT-2 (Radford et al., 2019).

GPT-2 is an autoregressive (as opposed to a Masked or Bidirectional) Transformer Language Model: it only uses the preceding context to generate probabilities for the next word. It was trained on forty gigabytes of text from the web. There are different sizes of the model that Open AI released; the smallest is a 12 layer neural network with 12 attention heads and 117 million free parameters; the largest is a 48 layer network with 25 attention heads and 1.5 billion parameters. A schematic of the inner workings of GPT-2 is provided in Figure 17.
In the example in the figure, the model has been given three words as context: $w_1$, $w_2$ and $w_3$. These words are turned into word embeddings, vectors with 768 dimensions. GPT-2 additionally uses position embeddings which encode the position of a word in its context. The embeddings are then transformed into three different vector representations by being multiplied against weight matrices: a query vector (e.g. $q_1$), a key vector (e.g. $k_1$) and a value vector (e.g. $v_1$). These are used for the “self-attention” mechanism. Attention is the main architectural advancement of Transformers as a class of neural networks. Attention works by taking the dot product of the query vector of the current word and the key vectors for all of the previously encountered words in the context, and then normalizing them (using a softmax function) to return attention scores. These attention scores roughly translate to the “importance” of each word in predicting the current word; a high attention score means higher importance. The attention scores are then multiplied by the value vectors and concatenated to create a vector of the entire
context. This vector is then run through a non-linear (ReLU) transformation to create a contextual embedding. So in the example in Figure 17, dot products of the query vector for the unknown word \( q_4 \) and the key vectors of the previous words \( k_1, k_2 \) and \( k_3 \) are used to generate the attention scores. These scores are used as multipliers for the value vectors \( v_1, v_2 \) and \( v_3 \) to create the contextual embedding for word four.

The base GPT-2 model has 12 “attention heads” meaning that the computations just described actually occur in parallel for 12 different learned representations of the query, key and value vectors for each word. That is, there are 12 distinct weight matrices (all filled with learned parameters) to transform an embedding into a query vector, one for each attention head, as is the case for the key and value vectors. The composite vectors created using each attention head are concatenated and fed forward as the input contextual embedding for the next layer. The base GPT-2 model has 12 layers of these computations. The final layer is run through a softmax function that transforms activations of words in the lexicon into probabilities. Thus, like all Language Models, GPT-2’s output is a probability distribution over next words. This is merely a sketch of how the model works and is not intended to be a full tutorial (see Radford et al., 2019 for the full model specification and explanation).

GPT-2 necessarily makes position-specific predictions. This is enforced largely by the position embeddings, but it is also a consequence of its training regimen. Thus in a way, it is a model (in the non-technical sense) of a participant in the cloze task; it must provide a good next word and should avoid putting any probability on words that are likely to surface sooner-or-later. Given its training regimen, the activations that feed forward into the softmax layer should be concentrated solely on good next words with little to no activation on inappropriate next words.

GPT-2 can be used to test for the rate of anomalous productions in role-reversal contexts within this strict framework to test if the consideration of position information is crucial for creating the difference observed in the cloze probabilities for ‘saved’ in the
Reversed and Canonical conditions. It is worth noting up front that GPT-2 is not a cognitive model of the human language processing system; any pattern observed in GPT-2's productions would be merely a proof-of-concept.

6.1 Experiments 2, 3 and 4 Using GPT-2 Base

Using Google Colab\textsuperscript{1}, the transformers (Wolf et al., 2019) and pytorch (Paszke et al., 2019) libraries were loaded. First, the GPT-2 tokenizer and base model were loaded. The sentence fragments from Experiments 2, 3 and 4 (available in Appendix A) were fed to the tokenizer which associated each word in the fragment with a numerical ID from GPT-2's lexicon. This tokenized sentence fragment was then given to the model which fed it up through the 12 layers of the network and returned probabilities assigned to each word in the lexicon. The probabilities assigned to the targets were summed and recorded, giving a comparable measure to the human data.

These data are summarized visually in Figure 18. Since all of the 2 Arguments conditions were the same in Experiments 2 and 4, they are equivalent, however, Experiment 3 (PP in the figure) utilized proper names in the Substitution conditions, so the 2 Arguments conditions could not all be collapsed together.

First and foremost, the scale of this figure should be emphasized. Whereas the human participants in Experiments 2, 3 and 4 provided a target response in the Canonical conditions with between 30% and 50% probability, GPT-2 assigns only roughly 5% probability to the targets in these conditions. GPT-2 does appear to assign probabilities in a more or less correct ordering within each facet. That is, the probability assigned to the targets in the Canonical conditions is higher than that in the Reversed and Substitution conditions, which mirrors the behavior of the human participants. However, the magnitude of these differences across conditions is rather different compared to those in the human data. In the human data, the Reversed 2 Arguments conditions yielded $\frac{1}{3}$ to $\frac{1}{5}$ the

\textsuperscript{1}https://colab.research.google.com
amount of target responses observed in the Canonical condition. Here, GPT-2 assigns more than $\frac{1}{2}$ the target probability in the Canonical 2 Arguments condition to the Reversed 2 Arguments condition, making them much closer together than the actual cloze data. The model also consistently assigned higher probability to the targets in the Reversed conditions as compared to the Substitution conditions. This is not consistently observed in the human data. Another pattern present in the data that is not seen in the human data is increase in target probability in both the Canonical and Reversed conditions in the 1 Argument Island conditions. The probability increases from 0.051 in the Canonical 2 Arguments condition to 0.061 in the Canonical 1 Argument Island condition, and the probability inches up from 0.037 in the Reversed 2 Arguments condition to 0.039 in the Reversed 1 Argument Island conditions, which run counter to the human data. This is indeed strange behavior for the model to exhibit.

Figure 18: GPT-2 Base Assigned Probability for the Target Words.

Additionally, across all of the conditions, it was found that the GPT-2 base probabili-
ties were correlated with cloze probabilities with \( r = 0.274 \). When fit with a simple linear regression model, it was found that cloze only marginally predicted GPT-2 probabilities (\( p = 0.063 \)).

All in all, while the the GPT-2 base model correctly determines that more probability mass should be assigned to the targets in the Canonical conditions than in the others, the model outputs qualitatively different patterns than humans in the cloze task.

### 6.2 Experiments 2, 3 and 4 Using GPT-2 Large

The same procedure was carried out for GPT-2 large, which has 36 layers, 20 attention heads and 774M free parameters. GPT-2 large's assigned target probabilities are presented visually in Figure 19. Across all of the conditions, GPT-2 large's probabilities were more highly correlated with the cloze data overall (\( r = 0.428 \)), and cloze was found to be a significant predictor of GPT-2's probabilities (\( p = 0.003 \)).
Figure 19: GPT-2 Large Assigned Probability for the Target Words.

Again, the scale is distinctly off as compared to the human data, with the maximum probability (in the Canonical 2 Arguments conditions) being only 11.8% as compared to roughly 50% for the human data. However, it is worth noting that GPT-2 large does assign higher probabilities to the targets in the Canonical conditions, especially in the 2 Arguments condition. This is a desirable improvement. At the same time, GPT-2 large assigned less probability to the targets in the Reversed and Substitution 1 Argument conditions across all of the 1 Argument experimental manipulations, which is also desirable.

Again the magnitude of the difference between the Canonical and Reversed conditions is not comparable to the human data, with the Reversed conditions leading to relatively higher target probability as a proportion of the Canonical conditions. And the model additionally assigned higher probability to the targets in the Reversed conditions as compared to the Substitution conditions across all of the panels, which was not true for the human data.
In sum, despite GPT-2 large producing more human-like patterns, there are still shortcomings that make its utility as a model of a human cloze participant questionable. Speculation about why GPT-2 assigns far less probability to the targets in the Canonical 2 Arguments conditions, and relatively high probability in the Canonical 1 Argument Island condition is discussed in the next section. The next section also explores why GPT-2 assigns so much relative probability mass to role-anomalous targets in the Reversed 2 Arguments conditions.

6.3 Discussion

There are a few key findings from looking at GPT-2’s behavior with role-reversals. First is the fact that GPT-2 assigns less probability to the targets than humans do in every condition, but this is particularly pronounced in the Canonical conditions. Second, GPT-2 does order the Canonical and Reversed conditions correctly, with more target probability assigned in the Canonical conditions, but the magnitude of the difference is quite reduced. And finally, in GPT-2 base, but not in GPT-2 large, the 1 Argument Island conditions lead to more target probability even than the 2 Arguments conditions. Each of these findings are discussed below, starting with the difficulty with islands.

GPT-2 and other large neural language models transform data into intermediate representations that often cannot be interpreted. Nevertheless, there are ways to investigate what it is that these models have learned (or failed to learn), which can inform researchers of why they behave certain ways in specific contexts. The main method of inquiry is to use probe tasks. These are carefully constructed sentences that test the model's knowledge of some some linguistic phenomenon; essentially linguistic experiments. Probe tasks have been used to investigate if language models learn island constraints, which is of particular importance for the current study as GPT-2 base in particular demonstrated peculiar behavior in the 1 Argument Island conditions.

While it has been argued that some modern language models do learn some of the
restrictions on wh-dependencies (Wilcox et al., 2018), Warstadt et al. (2020) found that GPT-2 does not appear to be sensitive to islands. Warstadt et al. (2020) probed GPT-2 large, along with other language models, by providing the model with a sentence featuring licit wh-movement (e.g. “What does Kathleen reveal she hates?”) and illicit, island-violating movement (e.g. “What does Kathleen reveal who hates?”), and testing if the model assigned higher probability to the grammatical sentence than the ungrammatical one. Thus chance performance was 50%. GPT-2 correctly ordered the sentences in 70% of the of the 1000 examples. This is significantly better than chance but also significantly lower than human performance, suggesting that GPT-2 has not achieved human-like competence with island effects. Despite this, GPT-2 did correlate the most highly with human judgments across a battery of other probes ($r = 0.54$) as compared to other language models.

Thus, it is possible that the pattern where GPT-2 base outputs larger target probabilities in the 1 Argument Island conditions compared even to the 2 Argument conditions is a reflection of the base model having acquired even less knowledge about islands. Additionally, the results from GPT-2 large, where the 1 Argument Island conditions do not feature as sharp a decline in target probabilities as they did with humans, may also be due to the model’s failure to acquire island constraints. This discrepancy noted by Warstadt et al. (2020) undermines confidence in using language model predictabilities as proxies for human predictions.

A qualitative way to assess the inner workings of GPT-2 is to visualize where the attention heads are “looking;” to see which words receive high attention scores for the purposes of predicting the current word. This was done using the “exBERT” tool available through huggingface. A visualization of the attention heads in layer eight are presented in Figure 20.

---

2https://huggingface.co/exbert/
As can be seen, there are relatively high attention scores for the which phrase in this layer. This is also true, and perhaps more so, for the Canonical version of the sentence, visualized in Figure 21. In none of the other layers is there as much attention for this example. Layer eight may thus be tentatively taken to attend to objects.

Interestingly, when the attention scores are visualized for the Reversed 1 Argument Island condition, as in Figure 22, there is some attention still being paid to to ‘lifeguard,’ but it appears to be decreased. This supplements the findings of Warstadt et al. (2020) by suggesting that GPT-2 does show some weak sensitivity to islands, as the attention scores to the illicitly extracted object appear to be less than those to the grammatical object, but at the same time is not able to block the usage of the illicit object entirely.
Figure 21: Visualization of GPT-2 Base Attention Scores for Canonical 2 Arguments Sentence. Thicker and darker lines indicate higher attention scores with the target word ‘saved.’

Figure 22: Visualization of GPT-2 Base Attention Scores for Reversed 1 Argument Island Sentence. Thicker and darker lines indicate higher attention scores with the target word ‘saved.’
Given that this investigation with exBERT was qualitative in nature, no hard conclusions should be drawn. A better investigation of GPT-2’s inner workings would make use of more probe tasks. However, in conjunction with the work of Warstadt et al. (2020), it appears that a plausible reason GPT-2 appears particularly non-human-like in the 1 Argument Island conditions is because GPT-2 simply does not have proper knowledge of island constraints.

Turning to the other patterns, recall that GPT-2 assigned substantially less probability to the target words than humans did in the Canonical conditions across the board. To restate this, GPT-2 is not likely to go “all-in” on its predictions even in cases such as the Canonical 2 Arguments conditions which are quite constraining toward the target verbs. This appears to potentially be a more general feature of GPT-2. For instance, the GPT-2 large assigned probability to ‘sugar’ as the next word in ‘I like my coffee with cream and …’ is 59% which is high, but considerably less than the nearly ubiquitous agreement among humans that ‘sugar’ will be the next word. Thus, the overall low target probabilities are likely a reflection of GPT-2’s general tendency to avoid putting too much probability mass on a single item, which minimizes the cross entropy loss (avoids penalization) in training.

It is also worth noting that while GPT-2’s goal is similar to the goal of cloze participants, it is qualitatively different. That is, researchers do not ask cloze participants to provide an estimate of the probability of sentence continuations, per se. GPT-2 essentially provides key-value pairs for words and their probabilities in the next position, whereas humans provide simply the word and no associated probability. Impressionistically, this would be an incredibly hard task for humans to perform. However, this discrepancy sheds light on why GPT-2 might underestimate the target cloze probabilities. It is possible and indeed almost necessary that if cloze participants were asked to provide a cloze response and provide a guess of its probability of actually occurring next, this probability would not be 100%. Even if a participant said their cloze response
was likely to be the next word 80% of the time, using these values to create a weighted-average cloze probability (summing the participants’ target probabilities and dividing by the number of responses) would pull the human numbers down. Nevertheless, this sort of variant of the cloze task could be run in future studies to more directly compare human and language model probabilities.

Additionally, it is known that humans do not produce cloze responses by probabilistically sampling from a distribution of words. Staub et al. (2015) argue against this by demonstrating that there is an effect of cloze probability and item constraint on cloze RTs. If participants were simply sampling from a distribution, there is no clear reason why high cloze responses would be in general produced faster, and why high item constraint would also lead to faster RTs. This is indeed the data that they leverage to argue for the Race Model of the cloze task.

Given these differences, it is in many ways unsurprising that GPT-2 probabilities and cloze probabilities are not of the same magnitude. Indeed probe tasks often don’t focus on absolute probability values, but rather they focus on comparisons between alternatives (Ettinger, 2020). Thus the fact that GPT-2 generally ordered the target probabilities correctly across conditions is perhaps the takeaway that many psycholinguistic researchers are interested in. However, the magnitude between differences does matter, especially when considering the power needed to detect an effect. For instance, if the 12% versus 7% difference from GPT-2 large were taken as estimates for the predictability of the targets in the Canonical 2 Arguments and Reversed 2 Arguments conditions, respectively, then one would not expect to find any detectable difference in cognitive measures; that difference is simply too small to translate to any N400 or eye tracking difference (if these were taken as cloze probabilities). One might be tempted by the fact that since we don’t actually observe any effects in the N400 or first fixation between the Canonical and Revered conditions, that the predictability estimates from GPT-2 are more fitting for the data. However, recall that Ehrenhofer et al. (2019) did observe an N400 ef-
fect for their stimuli. The probabilities from GPT-2 large for the subset of their stimuli that were used in Experiment 5 are summarized in Figure 23.

Again, it is plainly visible that GPT-2 assigns lower probabilities to the Target continuations (3%) compared to humans (35%). Additionally, the difference between the Targets and Lures is minimal, and certainly less than the differences obtained when using the sentences from Experiments 2, 3 and 4. Yet, these sentences do yield an N400 difference (Ehrenhofer et al., 2019). Even on a log scale, which some have argued is more appropriate at the lower end of the predictability scale (Smith and Levy, 2013; Szewczyk and Federmeier, 2022), the difference obtained for the 2AFC Ehrenhofer-style sentences is less than the differences in the 2 Arguments condition sentences in Experiments 2, 3 and 4 (0.41 versus 0.54, respectively).

Figure 23: Visualization of GPT-2 large Probabilities for 2AFC Sentences from Ehrenhofer et al. (2019).

All in all, the GPT-2 results offer reason for concern if one is to use language model
probabilities as predictability estimates for humans. While certainly there are benefits, as Szewczyk and Federmeier (2022) have noted, such as the scale with which these estimates can be acquired and the fine grain size of the estimates (especially at very low values), there are costs. First, there is contradictory evidence about whether GPT-2’s probabilities can qualitatively predict the presence or absence of N400 effects (Michaelov et al., 2021; Lindborg and Rabovsky, 2021). But perhaps more importantly, there is an a priori reason to prefer cloze probabilities over language model probabilities. First, cloze probabilities are themselves psycholinguistic measures that inform researchers what information about the sentence comprehenders have extracted, prioritized and utilized to compose meaning. While accessing GPT-2’s probabilities is “free,” it is not a substitute for humans’ subjective probabilities about upcoming words, which are guided by the human language processing algorithm, which psycholinguists are actually interested in (Smith and Levy, 2011; Staub et al., 2015).
CHAPTER 7

CONCLUSION

This dissertation has explored lexical predictability in sentence processing by zooming in on thematic role-reversal sentences. The aims of the dissertation were to investigate why it is that language users predict verbs that assign thematic roles anomalously, resulting in violations of world knowledge, and how we may reconcile the differing conclusions about the extent to which this occurs derived from different measures. The proposed answers to these questions are summarized in turn below.

7.1 Mechanisms Underlying Lexical Preactivations

As the experiments in this dissertation, and many others, have demonstrated, sentence contexts can preactivate lexical items. Some of these preactivated words are plausible sentence continuations but some are ultimately incompatible with how the world usually works. One of the goals of this dissertation is to broadly characterize the set of mechanisms that preactivate both appropriate and anomalous words.

Thematic role reversals are a fitting test case because predicting a role-anomalous verb implicates there being some “failure” on the part of the predictor to use all of the available sentence and world-knowledge information to constrain predictions. Moreover, cloze responses are apt measures of these predictions because they most clearly align with the presumed goal of the predictor: guessing upcoming words.

I have argued that verb preactivations are the result of distinct mechanisms. First, in
chapter 3, by comparing Experiments 2 and 3, I demonstrated that a small amount of lexical preactivation is attributable to simple Bag of Words associations. Then, by comparing Experiments 2 and 4, I argued that more preactivation flows from a mechanism I am calling a Bag of Unassigned Arguments mechanism where arguments that have not yet been associated with a predicate are prioritized to spread activation to verbs. Both of these mechanisms were argued for to explain the presence of anomalous predictions; however, a parsimonious account of preactivation holds that they are also partially responsible for the preactivation of appropriate predictions in the Canonical contexts.

The remaining preactivations are attributable to predictions derived from sentence compositional meanings. In the framework laid out in chapter 1, the comprehender composes meanings together to arrive at predictions about likely events and situations being described. Aspects of these events may be lexicalized which leads to the preactivation of situentially relevant words.

In chapter 1 I sketched out a model inspired by the lemma model (Levelt et al., 1999). With this model I was able to more explicitly characterize how lexical preactivation can occur as a result of concept-to-concept spreading activation and from inferred events derived from sentence compositional meanings. The Bag of Unassigned Arguments can also be couched in this model. At the point in time that the comprehender has encountered arguments that are not yet assigned to a predicate (they are in need of a verb), their incrementally built parse of the incomplete sentence can be used to guide expectations about the event that will ultimately be described. That is, if the comprehender has encountered both ‘lifeguard’ and ‘child,’ and both are in need of predication, the comprehender may begin to conjure up ideas for what event(s) these may ultimately be participants in. In terms of the model presented in Figure 2, information about the sentence fragment is being fed up from the parse to the non-linguistic event concept. Importantly, the thematic roles that will ultimately be assigned to these preverbal arguments are temporarily unknown and there may be multiple parses entertained for
the sentence fragment at this time. For each of these parses, there are many candidate events that could be the intended message conveyed by the sentence, and so the comprehender can entertain many different (and sometimes competing) events. With these events in mind, the system works as it does for the lexical predictions derived from sentence compositional meanings. These events contain concepts that have not yet been uttered; these concepts are preactivated in parallel. Among these preactivated concepts is the non-linguistic concept for the action itself which can be lexicalized as a verb. Finally the preactivation of these concepts preactivates the corresponding linguistic representations further down the hierarchy.

At present it looks as though cloze responses are guided primarily by sentence compositional meanings which generate plausible continuations. However, roughly 5% of cloze responses may be attributable to simple Bag of Words style associative priming and roughly 10% may be attributable to a slightly more sophisticated Bag of Unassigned Arguments mechanism.

In chapter 2, I demonstrated that simply displacing a verb prediction site from its associated syntactic object does not significantly diminish target verb preactivation. While this runs counter some findings in the role reversal literature, I argued that this is in line with other findings that have been leveraged to claim that predictions are maintained through time.

Taken together, these findings suggest that the Bag of Words and Bag of Unassigned Arguments lead to relatively long lasting preactivation of lexical units. Indeed this is in line with other findings suggesting that predictions are maintained, even after the context that preactivated them has concluded in some cases.

In chapter 4, I also demonstrated that in role-reversal contexts that have been shown to elicit the more typical N400 pattern, role-anomalous verbs are nevertheless still preactivated (above a baseline); they have simply been surpassed by another verb making them relatively less active. Thus, the findings of Ehrenhofer et al. (2019) need not be
viewed as exceptional. Rather, these sentences suggest that the strength of the associations between nouns and verbs in role reversal sentence contexts matters a great deal. Strong individual associations between a verb's syntactic subject and the verb leads to anomalous verbs losing out to more active alternatives, but the preactivation of the previously preactivated, and anomalous, verb does not necessarily dissipate.

In sum, it appears as though verbs are preactivated using a range of mechanisms with varying degrees of sophistication and that these predictions are maintained through time with little degradation.

7.2 Aligning Predictability Effects Across Methodologies

In Experiment 6 in chapter 5, it was found that reading times closely mirror the ERP findings in the role reversal literature. Role-anomalous verbs are initially read as quickly in Reversed conditions as when they are appropriate, in the corresponding Canonical conditions. The anomaly is not immediately detected; reading times in later measures and on spillover regions show an effect, while early measures such as the first fixation duration show no effect.

If one compares these findings with the cloze data, the magnitude of the preactivation of role-anomalous verbs appears to differ dramatically. While in the cloze experiments, role-anomalous verbs were preactivated, they were markedly less preactivated than when they were appropriate (in the Canonical conditions). In Experiment 6, they were equivalent. I argued that is due to the fact that reading times and ERPs are more prone to reflect predictions about words that the comprehender thinks may surface sooner or later in the context. Words that are implausible in a given position, but still possibly related to the event being described, may nevertheless be quite preactivated in anticipation of their eventual surfacing. Meanwhile cloze responses are more constrained to be fitting in the position that is left blank. I argued that either a "cloze filter" mechanism or a kind of appropriateness activation boost leads to the decreased cloze
responses in the Reversed conditions.

Thus the discrepancies between cloze, reading times and the N400 can be explained by task demand differences. Comprehenders in non-cloze tasks do not have the same demands to linearize their predicted lexical items. After all, predictions derived from event expectations are unordered because events themselves are unordered. There is nothing about the agent of an action or the patient of an action or the action itself that inherently suggests any kind of ordering.

This too can be implemented in the model I sketched out in chapter 1. The mechanisms outlined in the immediately preceding section (Bag of Words, Bag of Unassigned Arguments, etc.) are mechanisms for preactivating concepts and words. Since these mechanisms work by simple association or via the event concept space, these mechanisms preactivate concepts in an unordered fashion. To reiterate, events are themselves unordered things; there is nothing about the action, the actor, or anything else involved that suggests any kind of inherent ordering. Language however necessarily orders bits of information. So the preactivations derived from these mechanisms are, at first, sooner-or-later predictions about upcoming material.

For the cloze task, and sentence continuation tasks more generally, where the participant must consider the location of the blank in the sentence, a “cloze filter” or “appropriateness boost” comes online. In terms of the lemma-inspired model in chapter 1, one way this can be cashed out is in terms of connections from the “parse” box to the “lemma” box. Essentially, once position information must be considered (as in production), the parse of the sentence fragment be used to guide which preactivated lemmas are fitting for the current position. Again it is not clear whether these connections are inhibitory or facilitatory in nature.

Importantly, one does not need to allude to stages of linguistic information coming online to account for the difference between more online and offline measures; it need not be the case that argument-role mappings are not yet online in the participant’s mind
by the time the N400 is elicited but are by the time they produce a cloze response. Chow et al. (2016a) proposed that in online experiments (such as ERP experiments) the N400 is simply elicited before argument-role mappings have come online. And cloze responses, being a more offline measure, reflect predictions after this information has come online. They also use the findings from Momma (2016) and Chow et al. (2018) to argue that with more time, anomalous verb predictions decrease, though again see Experiment 1 in chapter 2 for reason to doubt that this occurs across the board. Rather than this staged model of verb prediction, I am proposing that differing task demands can give rise to this N400 (and now eye tracking) versus cloze discrepancy. Because the tasks ask for different levels of engagement with the linearization of the stimulus, participants are simply engaging with their predictions in different ways, with cloze requiring the consideration of the positional fitness of words in a way that other tasks don't. Future research can and should further test these ideas.

7.3 Limitations, Future Directions and Closing Remarks

An obvious limitation of the work in this dissertation is it focused on a very specific kind of relationship to investigate lexical preactivation: thematic relationships between a verb and its arguments. To more generally investigate the role of various mechanisms on lexical preactivation, other kinds of relations would need to be utilized.

A connection that has not been heavily drawn upon, but is potentially quite fruitful, is work with negation and prediction, where a similar cloze versus reading times and ERPs pattern has been observed (Fischler et al., 1983; Nieuwland and Kuperberg, 2008; Mayer et al., 2019). With negation, other elements of the sentence can be probed: nouns, adjectives, etc. in addition to verbs. Since negated sentences are almost identical to their affirmative counterparts, they are useful for testing the relative weighting of simple Bag of Words associations and sentence compositional meanings.

Consider (54). It has been found that the continuation ‘bird’ leads to a reduced N400
compared to a plausible and truthful continuation (e.g. ‘tree’), and that ‘bird’ is read as if it is predicted. On the face of it this would appear to be a case where comprehenders are not effectively using the sentence compositional meaning to constraint their predictions. First, it has been shown that given a richer context that constrains towards opposite adjectives, as in (55), the more typical pattern arises (Nieuwland and Kuperberg, 2008).

54. A robin is not a . . .

55. With the proper equipment, scuba diving is usually (not) considered to be very . . .

However, let’s return to the idea of sooner or later predictions versus next word predictions. In the context of (54), ‘bird’ is not a bad expectation for a word that might surface at some point, it is only an implausible continuation at that specific point. Consider if the sentence continued ‘blue bird’ or ‘large bird.’ Both of these would be true, plausible sentences and both of them have the word ‘bird’ in them. Thus, if the comprehender is preactivating words that may appear in the context at some point, ‘bird’ truly is a fine word to expect. It is only when they must consider the position, as they must do in the cloze task, when ‘bird’ becomes an anomalous response.

Work on negation, quantification (Nieuwland et al., 2010) and discourse connectives (Xiang and Kuperberg, 2015) offer a similar window into the mechanisms that lead to lexical preactivation in sentence contexts. Since the set of words that make up the contexts can be held constant (as is done for role reversals), the relative impact of the Bag of Words mechanism and predictions derived from sentence compositional meanings can be investigated further.

Ultimately, the work in this dissertation may be used to better understand how words become preactivated throughout the course of sentence and discourse comprehension. This is theoretically interesting in and of itself, but it also has practical applications such as improving language models to better match human performance.
I hope the ideas in this dissertation spark brilliant minds to pursue this line of research further and I invite collaborators to work with me to further our understanding of the amazing and mysterious linguistic machinery we humans are endowed with.
Appendices
APPENDIX A

EXPERIMENTAL STIMULI

All experimental materials are provided in a repository at:

https://osf.io/f8hkt/?view_only=3319c399cbb642eeb66d2b875c5b3e

- Experiment 1 Sentences: experiment1_stims.csv
- Experiment 1 Targets: experiment1_targets.csv
- Experiment 2 Sentences: experiment2_stims.csv
- Experiment 2 Targets: experiment2_targets.csv
- Experiment 3 Sentences: experiment3_stims.csv
- Experiment 3 Targets: experiment3_targets.csv
- Experiment 4 Sentences: experiment4_stims.csv
- Experiment 4 Targets: experiment4_targets.csv
- Experiment 5 Sentences: experiment5_stims.csv
• Experiment 6 Sentences: experiment6_stims.csv
APPENDIX B

SUPPLEMENTAL EYE TRACKING ANALYSES FOR EXPERIMENT 6

B.1 Analyses of the Pre-Target Region

\[ \text{Skip} \sim \text{Argumenthood} \times \text{Order} + (1 + \text{Argumenthood} \mid \text{Participant}) + (1 + \text{Order} \mid \text{Item}) \]

<table>
<thead>
<tr>
<th>Fixed Effect</th>
<th>Estimate (ms)</th>
<th>z-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-1.454</td>
<td>-7.429</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Argumenthood</td>
<td>-3.151</td>
<td>-11.934</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Canonical vs. Reversed</td>
<td>0.003</td>
<td>0.026</td>
<td>0.980</td>
</tr>
<tr>
<td>Canonical vs. Substitution</td>
<td>0.001</td>
<td>0.004</td>
<td>0.997</td>
</tr>
<tr>
<td>Canonical vs. Reversed : Argumenthood</td>
<td>0.093</td>
<td>0.407</td>
<td>0.684</td>
</tr>
<tr>
<td>Canonical vs. Substitution : Argumenthood</td>
<td>-0.069</td>
<td>-0.301</td>
<td>0.764</td>
</tr>
</tbody>
</table>

Table 22: Summary of the statistical analysis of the probability of skipping the pre-target region for Experiment 6. Bolded rows indicate significant effects (at \( \alpha = 0.0125 \)). The model specification is provided above the table.

\[ \text{First Fixation Duration} \sim \text{Argumenthood} \times \text{Order} + (1 \mid \text{Participant}) + (1 \mid \text{Item}) \]

<table>
<thead>
<tr>
<th>Fixed Effect</th>
<th>Estimate (ms)</th>
<th>t-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>247.217</td>
<td>45.147</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Argumenthood</td>
<td>18.265</td>
<td>2.766</td>
<td>0.006</td>
</tr>
<tr>
<td>Canonical vs. Reversed</td>
<td>1.662</td>
<td>0.358</td>
<td>0.720</td>
</tr>
<tr>
<td>Canonical vs. Substitution</td>
<td>-3.881</td>
<td>-0.831</td>
<td>0.406</td>
</tr>
<tr>
<td>Canonical vs. Reversed : Argumenthood</td>
<td>-13.879</td>
<td>-1.495</td>
<td>0.135</td>
</tr>
<tr>
<td>Canonical vs. Substitution : Argumenthood</td>
<td>-4.498</td>
<td>-0.482</td>
<td>0.630</td>
</tr>
</tbody>
</table>

Table 23: Summary of the statistical analysis of the first fixation duration of the pre-target region for Experiment 6. Bolded rows indicate significant effects (at \( \alpha = 0.0125 \)). The model specification is provided above the table.
### First Pass Time ~ Argumenthood * Order + (1 | Participant) + (1 | Item)

<table>
<thead>
<tr>
<th>Fixed Effect</th>
<th>Estimate (ms)</th>
<th>t-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>296.005</td>
<td>31.200</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Argumenthood</td>
<td>109.310</td>
<td>10.268</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Canonical vs. Reversed</td>
<td>−6.342</td>
<td>−0.848</td>
<td>0.397</td>
</tr>
<tr>
<td>Canonical vs. Substitution</td>
<td>−14.055</td>
<td>−1.865</td>
<td>0.0623</td>
</tr>
<tr>
<td>Canonical vs. Reversed : Argumenthood</td>
<td>−34.011</td>
<td>−2.272</td>
<td>0.023</td>
</tr>
<tr>
<td>Canonical vs. Substitution : Argumenthood</td>
<td>−19.205</td>
<td>−1.275</td>
<td>0.202</td>
</tr>
</tbody>
</table>

Table 24: Summary of the statistical analysis of the first pass time of the pre-target region for Experiment 6. Bolded rows indicate significant effects (at $\alpha = 0.0125$). The model specification is provided above the table.

### Go Past Time ~ Argumenthood * Order + (1 + Argumenthood | Participant) + (1 | Item)

<table>
<thead>
<tr>
<th>Fixed Effect</th>
<th>Estimate (ms)</th>
<th>t-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>411.59</td>
<td>20.497</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Argumenthood</td>
<td>141.75</td>
<td>4.334</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Canonical vs. Reversed</td>
<td>24.27</td>
<td>1.157</td>
<td>0.247</td>
</tr>
<tr>
<td>Canonical vs. Substitution</td>
<td>−18.73</td>
<td>−0.888</td>
<td>0.375</td>
</tr>
<tr>
<td>Canonical vs. Reversed : Argumenthood</td>
<td>−31.04</td>
<td>−0.740</td>
<td>0.459</td>
</tr>
<tr>
<td>Canonical vs. Substitution : Argumenthood</td>
<td>−5.72</td>
<td>−0.136</td>
<td>0.892</td>
</tr>
</tbody>
</table>

Table 25: Summary of the statistical analysis of the go past time of the pre-target region for Experiment 6. Bolded rows indicate significant effects (at $\alpha = 0.0125$). The model specification is provided above the table.
### B.2 Analyses of the Post-Target (Spillover) Region

Skips \( \sim \) Argumenthood * Order + (1 + Argumenthood | Participant) + (1 + Argumenthood | Item)

<table>
<thead>
<tr>
<th>Fixed Effect</th>
<th>Estimate (ms)</th>
<th>z-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-3.524</td>
<td>-13.921</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Argumenthood</td>
<td>-0.243</td>
<td>-0.553</td>
<td>0.580</td>
</tr>
<tr>
<td>Canonical vs. Reversed</td>
<td>-0.257</td>
<td>-1.515</td>
<td>0.130</td>
</tr>
<tr>
<td>Canonical vs. Substitution</td>
<td>-0.266</td>
<td>-1.562</td>
<td>0.118</td>
</tr>
<tr>
<td>Canonical vs. Reversed : Argumenthood</td>
<td>0.599</td>
<td>1.757</td>
<td>0.079</td>
</tr>
<tr>
<td>Canonical vs. Substitution : Argumenthood</td>
<td>-0.106</td>
<td>-0.310</td>
<td>0.756</td>
</tr>
</tbody>
</table>

Table 26: Summary of the statistical analysis of the probability of skipping the spillover region for Experiment 6. Bolded rows indicate significant effects (at \( \alpha = 0.0125 \)). The model specification is provided above the table.

First Fixation Duration \( \sim \) Argumenthood * Order + (1 + Argumenthood | Participant) + (1 + Argumenthood | Item)

<table>
<thead>
<tr>
<th>Fixed Effect</th>
<th>Estimate (ms)</th>
<th>t-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>236.977</td>
<td>54.473</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Argumenthood</td>
<td>7.832</td>
<td>1.416</td>
<td>0.159</td>
</tr>
<tr>
<td>Canonical vs. Reversed</td>
<td>0.173</td>
<td>0.053</td>
<td>0.958</td>
</tr>
<tr>
<td>Canonical vs. Substitution</td>
<td>2.768</td>
<td>0.848</td>
<td>0.396</td>
</tr>
<tr>
<td>Canonical vs. Reversed : Argumenthood</td>
<td>-11.640</td>
<td>-1.783</td>
<td>0.075</td>
</tr>
<tr>
<td>Canonical vs. Substitution : Argumenthood</td>
<td>-4.076</td>
<td>-0.624</td>
<td>0.533</td>
</tr>
</tbody>
</table>

Table 27: Summary of the statistical analysis of the first fixation duration of the spillover region for Experiment 6. Bolded rows indicate significant effects (at \( \alpha = 0.0125 \)). The model specification is provided above the table.
First Pass Time $\sim$ Argumenthood * Order + (1 + Order | Participant) + (1 + Argumenthood | Item)

<table>
<thead>
<tr>
<th>Fixed Effect</th>
<th>Estimate (ms)</th>
<th>t-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>358.807</td>
<td>30.619</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Argumenthood</td>
<td>−37.004</td>
<td>−1.964</td>
<td>0.053</td>
</tr>
<tr>
<td>Canonical vs. Reversed</td>
<td>14.151</td>
<td>1.705</td>
<td>0.092</td>
</tr>
<tr>
<td>Canonical vs. Substitution</td>
<td>19.909</td>
<td>2.476</td>
<td>0.015</td>
</tr>
<tr>
<td>Canonical vs. Reversed : Argumenthood</td>
<td>−5.157</td>
<td>−0.363</td>
<td>0.717</td>
</tr>
<tr>
<td>Canonical vs. Substitution : Argumenthood</td>
<td>5.604</td>
<td>0.394</td>
<td>0.694</td>
</tr>
</tbody>
</table>

Table 28: Summary of the statistical analysis of the first pass time of the spillover region for Experiment 6. Bolded rows indicate significant effects (at $\alpha = 0.0125$). The model specification is provided above the table.

Go Past Time $\sim$ Argumenthood * Order + (1 + Argumenthood | Participant) + (1 + Argumenthood | Item)

<table>
<thead>
<tr>
<th>Fixed Effect</th>
<th>Estimate (ms)</th>
<th>t-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>564.34</td>
<td>16.813</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Argumenthood</td>
<td>−46.90</td>
<td>−0.796</td>
<td>0.427</td>
</tr>
<tr>
<td>Canonical vs. Reversed</td>
<td>103.14</td>
<td>3.313</td>
<td>0.001</td>
</tr>
<tr>
<td>Canonical vs. Substitution</td>
<td>91.85</td>
<td>2.949</td>
<td>0.003</td>
</tr>
<tr>
<td>Canonical vs. Reversed : Argumenthood</td>
<td>−80.65</td>
<td>−1.295</td>
<td>0.195</td>
</tr>
<tr>
<td>Canonical vs. Substitution : Argumenthood</td>
<td>−87.07</td>
<td>−1.398</td>
<td>0.162</td>
</tr>
</tbody>
</table>

Table 29: Summary of the statistical analysis of the go past time of the spillover region for Experiment 6. Bolded rows indicate significant effects (at $\alpha = 0.0125$). The model specification is provided above the table.

### B.3 Analysis of the Target Word with Response Type as a Predictor

These models only included data from the Reversed conditions
### Table 30: Summary of the Statistical Analysis of the Probability of Skipping the Target Word in the Reversed Condition by Response Type for Experiment 6. Bolded rows indicate significant effects (at $\alpha = 0.0125$). The model specification is provided above the table.

<table>
<thead>
<tr>
<th>Fixed Effect</th>
<th>Estimate (logits)</th>
<th>z-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-1.981</td>
<td>-13.342</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Argumenthood</td>
<td>0.155</td>
<td>0.693</td>
<td>0.489</td>
</tr>
<tr>
<td>Response</td>
<td>0.118</td>
<td>0.671</td>
<td>0.502</td>
</tr>
<tr>
<td>Argumenthood : Response</td>
<td>-0.475</td>
<td>-1.386</td>
<td>0.166</td>
</tr>
</tbody>
</table>

### Table 31: Summary of the Statistical Analysis of the First Fixation Duration of the Target Word in the Reversed Condition by Response Type for Experiment 6. Bolded rows indicate significant effects (at $\alpha = 0.0125$). The model specification is provided above the table.

<table>
<thead>
<tr>
<th>Fixed Effect</th>
<th>Estimate (ms)</th>
<th>t-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>238.390</td>
<td>56.300</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Argumenthood</td>
<td>-12.915</td>
<td>-2.770</td>
<td>0.006</td>
</tr>
<tr>
<td>Response</td>
<td>4.640</td>
<td>0.766</td>
<td>0.447</td>
</tr>
<tr>
<td>Argumenthood : Response</td>
<td>10.469</td>
<td>1.105</td>
<td>0.269</td>
</tr>
</tbody>
</table>

### Table 32: Summary of the Statistical Analysis of the First Pass Time of the Target Word in the Reversed Condition by Response Type for Experiment 6. Bolded rows indicate significant effects (at $\alpha = 0.0125$). The model specification is provided above the table.

<table>
<thead>
<tr>
<th>Fixed Effect</th>
<th>Estimate (ms)</th>
<th>t-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>270.605</td>
<td>40.736</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Argumenthood</td>
<td>-30.164</td>
<td>-3.881</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Response</td>
<td>8.510</td>
<td>1.110</td>
<td>0.267</td>
</tr>
<tr>
<td>Argumenthood : Response</td>
<td>5.082</td>
<td>0.350</td>
<td>0.726</td>
</tr>
</tbody>
</table>
Go Past ~ Argumenthood * Response + (1 | Participant) + (1 | Item)

<table>
<thead>
<tr>
<th>Fixed Effect</th>
<th>Estimate (ms)</th>
<th>t-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>456.79</td>
<td>16.919</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Argumenthood</td>
<td>39.08</td>
<td>1.086</td>
<td>0.278</td>
</tr>
<tr>
<td>Response</td>
<td>20.52</td>
<td>0.541</td>
<td>0.589</td>
</tr>
<tr>
<td>Argumenthood : Response</td>
<td>−17.22</td>
<td>−0.235</td>
<td>0.814</td>
</tr>
</tbody>
</table>

Table 33: Summary of the Statistical Analysis of the Go Past Time of the Target Word in the Reversed Condition by Response Type for Experiment 6. Bolded rows indicate significant effects (at $\alpha = 0.0125$). The model specification is provided above the table.

Arehalli, S. and Linzen, T. (2020). Neural language models capture some, but not all, agreement attraction effects.


164


R Core Team (2021). R: A language and environment for statistical computing.


