A COMPUTATIONAL MODEL OF COGNITIVE CONSTRAINTS IN SYNTACTIC LOCALITY

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Marisa Ferrara Boston

January 2012

A COMPUTATIONAL MODEL OF COGNITIVE CONSTRAINTS IN SYNTACTIC LOCALITY

Marisa Ferrara Boston, Ph.D.

Cornell University 2012

This dissertation is broadly concerned with the question: how do human cognitive limitations influence difficult sentences? The focus is a class of grammatical restrictions, locality constraints. The majority of relations between words are local; the relations between question words and their governors are not. Locality constraints restrict the formation of these non-local dependencies. Though necessary, the origin, operation, and scope of locality constraints is a controversial topic in the literature.

The dissertation describes the implementation of a computational model that clarifies these issues. The model tests, against behavioral data, a series of cognitive constraints argued to account for locality. The result is an explanatory model predictive of a variety of cross-linguistic locality data. The model distinguishes those cognitive limitations that affect locality processing, and addresses the competence-performance debate by determining how and when cognitive constraints explain human behavior. The results provide insight into the nature of locality constraints, and promote language models sensitive to human cognitive limitations.

BIOGRAPHICAL SKETCH

Marisa Ferrara Boston (née Marisa Angela Ferrara) was born January 26, 1981 in Painesville, Ohio. She received a Bachelor of Arts degree in Individualized Studies from Eastern Michigan University in 2001, and a Master of Arts degree in English Linguistics from Eastern Michigan University in 2005. She also received a Certificate in Language Technology from Eastern Michigan University in 2005. She was awarded an IGERT Cognitive Science fellowship from Michigan State University from 2005 to 2008. She finished her graduate studies at Cornell University, where she has served as a Research Assistant in the Department of Linguistics.

To my father, Salvatore Ferrara, who taught me to work hard,

my husband, Jason Boston, who taught me to never give up,

and my son, Felix Boston, who taught me to enjoy the journey.

ACKNOWLEDGEMENTS

This dissertation is meaningful because of those who have shared in the journey, and I would like to acknowledge the many who have helped me along the way. First on the list is my guide, John Hale. The foundations of this dissertation, and of much of my approach to science, stem from John and the incredible amount of enthusiasm, curiosity, and knowledge he brings to his work. John pushed me to embrace the subjects I found most challenging, and designed creative coursework and projects that taught me to approach linguistics from a broader perspective. I am grateful that I gave that young upstart a chance all those years ago, and that he gave me one as well.

My approach to research and linguistics has also been shaped by my collaborations with Shravan Vasishth. By some strange bit of luck Shravan was placed on my path, and he helped to broaden my perspective even further. Shravan's particular talent is his patience in advising students in those aspects of research so difficult for students to learn. Whether it is writing papers, developing a research plan, or being more academic, Shravan is ready with some book or motivating advice. I also admire his strict code of honor requiring an intimate relationship with the tiniest of details in research, a code I strive to continue.

Because the dissertation is a personal journey, it can feel at times that you haven't seen another person for miles. That is why I particularly cherish my encounters with people who have taken the time to understand this work and offer comments, suggestions, and technical advice. Mats Rooth and Claire Cardie round out my dissertation committee, and were unfailingly knowledgeable, insightful and helpful throughout the various stages of this work.

Rick Lewis and Sam Epstein ensured that my interpersonal skills didn't whither away once I moved away from Cornell. They invited me take part in their Rational Behavior and Minimalist Inquiry class at the University of Michigan. The class allowed me to take a step back from the implementation and think deeply on competence and performance issues in linguistics. What began as a way to keep myself current ended up having a profound effect on the design of this work, and I am most grateful for their support and intellectual stimulation.

It turns out that although the dissertation is a personal journey, many of the topics themselves are well-worn paths. I owe much to the experience of those who came before me. This work has benefited greatly from the suggestions of Colin Phillips, Philip Hofmeister, and Masaya Yoshida, researchers who worked with the competence and performance issues of syntactic locality much longer than I have. Joakim Nivre was my gentle dependency parsing mentor before I even knew what dependency parsing was, and his ideas have helped me to fall in love with the approach. Finally, Marco Kuhlmann so kindly and patiently helped me navigate the steep learning curve towards mathematical linguistics, leading me to areas of research I never knew I had an interest in.

Although this journey ends at Cornell, it began at Michigan State. The IGERT community at MSU introduced me to cognitive science, and it was this introduction that motivated me to stay on the path towards this degree. I am indebted to the MSU professors on my original dissertation committee, Joyce Chai, Yen-Hwei Lin, Marcin Morzycki, and Alan Munn, for their early encouragement.

I was first introduced to the field of linguistics at Eastern Michigan University, where I was a graduate assistant at the LINGUIST List. My managers at LINGUIST List, Helen Aristar-Dry and Tony Aristar, persuaded me to continue on towards a PhD despite my reluctance. They gave me that proverbial push from the nest, and I thank them for it. Sometimes.

If I had to find a true beginning to this journey, it would date back to when I was an undergraduate student at Eastern Michigan University. There I was, trying to somehow combine a French major with an interest in aviation, when my cultural diversity professor, Liza Cerroni-Long, took pity on me. She saw in me a kindred spirit, and our annual dinners since have provided me with a refreshing reminder of what I am striving for. You have always been what I like best about this profession, Liza.

And what is any journey without your fellow pilgrims? I must acknowledge my academic soul sisters, three girls who helped me through three different phases of this process. Effi Georgala was the fashionable Euro girl at Cornell who I thought for sure would hate me when

I arrived. Instead, she loved my shoes. And I loved hers. Our daily (yes, daily) emails over the past two years kept us honest and on-task as we both worked on our dissertations. So much of this journey has been shared with Effi, and although I perhaps should have packed tennis shoes, I know that there is at least one girl who approves of my stylish kicks.

Secondly, there is Sabrina Gerth from the University of Potsdam. Conference buddy and all-around amiga, Sabrina and I shared conference adventures and heart-to-hearts that would probably beat out the best that any high school sleepover could offer. To this day, there is no one I would rather share lunch with on a train to Bielefeld.

Finally, Tanya Sydorenko and I shared daily white-knuckled and mostly snowy commutes across the state of Michigan for two full years. Along the way, she helped me through those difficult early days of graduate school when the destination seemed impossibly far away. You were right, Tanya: we did eventually make it. And I even went the speed limit.

I probably would not have made it to this destination without the daily laughs from great lab mates: Zhong Chen, Kyle Grove, and Jiwon Yun; or the once-a-year laughs from the best of conference buddies: Umesh Patil, Titus von der Malsburg, Paul Engelhardt, Matt Husband, and Nikole Patson-Huffman; or the IGERT brothers who helped me through my first Java project: Matt Gerber and Zahar Prasov (I told you guys I wouldn't forget); or those real-world friends who now have an unhealthy amount of computational linguistics knowledge for the average citizen: Sarah Ziegler, Susan Smith, and Kim Dunlap; or that one sister who learned to predict my dissertation progress by the outfit I was wearing: Diana Ferrara; or that one flower I stopped to pick along the way: Felix Boston; or the support of my parents, siblings, grandparents, in-laws, aunts, uncles, cousins, and various other family who accepted me for who I am before, during, and even after this.

Most importantly of all, I am incredibly fortunate to be able to look back on this journey and realize I did not lose sight of what is most important to me. Jason Boston, you have kept me grounded throughout the worst, you have celebrated with me throughout the best, and you held my hand no matter how far away from each other we had to be. I may have written this,

but you made it something that I can be proud of.

Thank you all.

TABLE OF CONTENTS

Biographical Sketch . . . iii
Dedication . . . iv
Acknowledgements . . . v
Table of Contents . . . ix
List of Tables . . . xii
List of Figures . . . xiii

1 Introduction 1
1.1 Introduction . . . 1
1.2 Overview of the problem . . . 2
1.3 Significance . . . 4
1.4 Chapter overview . . . 8

2 Syntactic Locality 11
2.1 Introduction . . . 11
2.2 Dependency grammar . . . 11
2.3 Syntactic locality phenomena . . . 14
2.3.1 Islands . . . 15
2.3.2 Superiority violations . . . 20
2.4 Gradience in locality . . . 22
2.5 Explanations from linguistic theory . . . 24
2.5.1 A syntactic explanation for gradience . . . 24
2.5.2 A semantic explanation for gradience . . . 25
2.6 Limitations of linguisic explanations . . . 27
2.7 Conclusion . . . 28

3 Cognitive Constraints 29
3.1 Introduction . . . 29
3.2 Working memory in sentence processing . . . 30
3.3 Activation and Interference . . . 31
3.4 Implemented theories . . . 32
3.4.1 Activation theories . . . 33
3.4.2 Interference theories . . . 34
3.4.3 Combination theories . . . 37
3.5 Conclusion . . . 40

4 A computational model 41
4.1 Introduction . . . 41
4.2 Why a dependency parser? . . . 43
4.2.1 Why dependency grammar? . . . 43
4.2.2 Why the Nivre transition system? . . . 45
4.2.3 Why a non-projective model? . . . 46
4.3 The Nivre non-projective transition system as a cognitive model . . . 49
4.4 The oracle . . . 53

4.4.1 Treebanks . . . 53
4.4.2 ① and ②: Treebank transformations . . . 55
4.4.3 ③ and ④: From treebanks to state-action banks . . . 56
4.4.4 ⑤ and ⑥: From state-action banks to feature banks . . . 56
4.4.5 ⑦, ⑧, and ⑨: From feature banks to probabilistic features . . . 58
4.5 Cognitive theories as probabilistic features . . . 61
4.5.1 Activation . . . 63
4.5.2 Interference . . . 65
4.5.3 Composite . . . 67
4.6 Complexity metrics . . . 68
4.6.1 Surprisal . . . 68
4.6.2 Retrieval . . . 72
4.7 Dependency length and difficulty . . . 78
4.8 Accuracy . . . 79
4.9 Conclusion . . . 81

5 Experiments **82**
5.1 Introduction . . . 82
5.2 Measures . . . 84
5.2.1 Syntactic judgments . . . 84
5.2.2 Acceptability and Magnitude Estimation . . . 85
5.2.3 Reading Time and Residual Reading Time (RRT) . . . 85
5.3 CNP . . . 86
5.3.1 Classic . . . 86
5.3.2 Gradience . . . 87
5.3.3 Challenges . . . 89
5.4 WHI . . . 93
5.4.1 Classic . . . 93
5.4.2 Gradience . . . 94
5.4.3 Challenges . . . 96
5.5 SUV . . . 97
5.5.1 Classic . . . 97
5.5.2 Gradience . . . 98
5.5.3 Challenges . . . 100
5.6 Experiments that were not modeled . . . 103
5.7 Conclusion . . . 105

6 A methodology for cognitive modeling **106**
6.1 Introduction . . . 106
6.2 Encoding practice . . . 109
6.2.1 Assigning dependency analyses to experimental sentences . . . 110
6.2.2 Determining the crucial arcs . . . 112
6.3 Evaluation process . . . 114
6.3.1 ① Transform into dependency analyses . . . 114
6.3.2 ② Run through the parser . . . 116

6.3.3 ③ Obtain difficulty measures for each cognitive factor . . . 116
6.3.4 ④ Transform into percentage difficulty by condition . . . 117
6.3.5 ⑤ Compare to human difficulty . . . 118
6.4 Conclusion . . . 121

7 Results **122**
7.1 Introduction . . . 122
7.2 Results by study . . . 122
7.2.1 Introduction . . . 122
7.2.2 CNP . . . 123
7.2.3 WHI . . . 133
7.2.4 SUV . . . 140
7.3 Results by phenomenon . . . 149
7.3.1 CNP . . . 150
7.3.2 WHI . . . 152
7.3.3 SUV . . . 155
7.4 Discussion . . . 158
7.4.1 Comparison to other work . . . 158
7.4.2 CNPs: Argument from a null result . . . 161
7.4.3 WHIs: A reductionst phenomenon . . . 163
7.4.4 SUVs: Insight from a computational model . . . 163
7.5 Conclusion . . . 165

8 Conclusions **166**
8.1 Introduction . . . 166
8.2 Summary of findings . . . 166
8.3 Contributions . . . 167
8.4 Future work . . . 171
8.5 Conclusion . . . 173

A List of Abbreviations **174**

References **176**

LIST OF TABLES

4.1 Transitions for the Nivre non-projective transition system. 50
4.2 These treebanks inform parser probabilities. 53
4.3 Fragment of a Brown sentence in CoNLL format. 54
4.4 Treebank part-of-speech transformations. 57
4.5 Feature specification. :: indicates concatenation. 63
4.6 How time is determined in the parser. 75

7.1 CNP Summary Table. 151
7.2 WHI Summary Table. 153
7.3 SUV Summary Table. 156

LIST OF FIGURES

2.1 A DG example. . . . 12
2.2 A discontinuous sentence in Czech. . . . 12
2.3 A grammatical non-local wh-word. . . . 14
2.4 An ungrammatical non-local wh-word. . . . 14
2.5 Question formation in the generative framework. . . . 16
2.6 An example of a CNP violation. . . . 17
2.7 An ECP violation. . . . 19
2.8 An example of a WHI violation. . . . 20
2.9 An example of a SUV violation. . . . 21
2.10 Experimental evidence of SUV gradience. . . . 23
2.11 Experimental evidence of CNP gradience. . . . 23
2.12 Experimental evidence of WHI gradience. . . . 24

3.1 The integration site for long-distance dependencies. . . . 30

4.1 The parts of a cognitive model. . . . 42
4.2 A non-projective analysis of an English sentence. . . . 47
4.3 A non-projective experimental sentence. . . . 48
4.4 An example parse of a CNP experimental sentence. . . . 52
4.5 Parsing non-projective dependencies. . . . 60
4.6 How likely is an attachment between *who* and *captured*? . . . 64
4.7 Surprisal calculation. . . . 70
4.8 Retrieval calculation. . . . 77

5.1 An organizational chart of syntactic locality experiments. . . . 83
5.2 CNP classic example 1: English (measure: RRT). . . . 87
5.3 CNP gradience example 1: English (measure: RRT). . . . 88
5.4 CNP gradience example 2: English (measure: acceptability). . . . 89
5.5 CNP challenge example 1: English (measure: acceptability). . . . 90
5.6 CNP challenge example 2: English (measure: acceptability). . . . 91
5.7 CNP challenge example 3: German (measure: acceptability). . . . 92
5.8 CNP challenge example 4: Swedish (measure: acceptability). . . . 93
5.9 WHI classic example 1: English (measure: RT). . . . 94
5.10 WHI gradience example 1: English (measure: RT). . . . 94
5.11 WHI gradience example 2: English (measure: acceptability). . . . 95
5.12 WHI gradience example 3: German (measure: acceptability). . . . 95
5.13 WHI challenge example 1: English (measure: acceptability). . . . 96
5.14 WHI challenge example 2: Swedish (measure: syntax). . . . 97
5.15 SUV classic example 1: English (measure: syntax). . . . 98
5.16 SUV gradience example 1: English (measure: RRT). . . . 98
5.17 SUV gradience example 2: English (measure: acceptability). . . . 99
5.18 SUV gradience example 3: German (measure: acceptability). . . . 100
5.19 SUV challenge example 1: German (measure: syntax). . . . 101
5.20 SUV challenge example 2: German (measure: acceptability). . . . 102

5.21 SUV challenge example 3: Russian (measure: acceptability). 103

6.1 A sample graph comparing cognitive predictions to human difficulty. 106
6.2 The evaluation process. 108
6.3 A WHI gradience example. Measure: acceptability. 111
6.4 The evaluation process explained. 115
6.5 A WHI gradience example. Measure: acceptability. 118
6.6 The resulting graph. 120

7.1 CNP classic example 1: English (measure: RRT). 124
7.2 Full CNP classic results. 124
7.3 CNP gradience example 1: English (measure: RRT). 126
7.4 Full CNP gradience results. 127
7.5 CNP gradience example 2: English (measure: acceptability). 128
7.6 CNP challenge example 1: English (measure: acceptability). 129
7.7 Full CNP challenge results. 130
7.8 CNP challenge example 2: English (measure: acceptability). 131
7.9 CNP challenge example 3: German (measure: acceptability). 131
7.10 CNP challenge example 4: Swedish (measure: acceptability). 132
7.11 WHI classic example 1: English (measure: RT). 133
7.12 Full WHI classic results. 134
7.13 WHI gradience example 1: English (measure: RT). 135
7.14 Full WHI gradience results. 136
7.15 WHI gradience example 2: English (measure: acceptability). 137
7.16 WHI gradience example 3: German (measure: acceptability). 137
7.17 WHI challenge example 1: English (measure: acceptability). 138
7.18 Full WHI challenge results. 138
7.19 WHI challenge example 2: Swedish (measure: syntax). 139
7.20 SUV classic example 1: English (measure: syntax). 140
7.21 Full SUV classic results. 141
7.22 SUV gradience example 1: English (measure: RRT). 142
7.23 SUV gradience example 2: English (measure: acceptability). 143
7.24 Full SUV gradience results. 144
7.25 SUV gradience example 3: German (measure: acceptability). 145
7.26 SUV challenge example 1: German (measure: syntax). 146
7.27 Full SUV challenge results. 147
7.28 SUV challenge example 2: German (measure: acceptability). 148
7.29 SUV challenge example 3: Russian (measure: acceptability). 149
7.30 Dependency length is greater in CNP-violating contexts. 150
7.31 The dependency structure is different for WHI-violating Swedish sentences. . . 154
7.32 An updated organizational chart of syntactic locality experiments. 157

CHAPTER 1

INTRODUCTION

1.1 Introduction

Is syntactic locality a processing phenomenon? This topic has been central to the competence and performance debate in linguistics for half a century, leading to advances in both grammatical and processing theories. Yet it has also resulted in sometimes contentious disagreements among linguists and psycholinguists regarding the origin and nature of the phenomenon. These disagreements underscore the increasing importance of understanding how competence and performance interact in language comprehension.

This dissertation employs a human sentence processing model armed with cognitive constraints to predict syntactic locality difficulty. The cognitive constraints test **reductionist**, or performance-based, accounts of syntactic locality. The broad-coverage nature of the model allows for an examination of multiple syntactic locality phenomena, represented by experimental data across a variety of languages, in a single architectural model. The aim is to determine whether locality is a processing phenomenon, a competence phenomenon, or both.

The results demonstrate that a model that uses solely cognitive constraints can model weak islands and superiority violations. More specifically, weak islands are a result of multiple working memory factors, including activation and interference. Superiority violations, on the other hand, are a result of interference only. The cognitive constraints are not able to model strong islands, indicating this particular locality violation is a result of either a grammatical constraint or even an untested cognitive factor.

The remaining sections of this chapter provide more details that motivate this dissertation topic and its results. Section 1.2 provides further background into syntactic locality and the

problem this dissertation addresses. Section 1.3 discusses the significance of the results to the syntactic locality debate, as well as to other areas such as computational linguistics and psycholinguistics. Section 1.4 provides an overview of each of the chapters in this dissertation.

1.2 Overview of the problem

This work develops a computational model of a pervasive aspect of language comprehension, syntactic locality. The majority of the relations between words are local; however, the relation between question words and their governors are not. (1) demonstrates one such sentence, where *what* is distant from its governor, *read*.

(1) **What** do you think that Diego read?

Not all non-local dependencies between words are possible in natural language. Many are constrained. For example, sentences like (2) are considered difficult by English speakers.

(2) ***What** do you wonder *whether* Diego read?

Here, there is a constraint on the long-distance dependency because the wh-word has a governor that is within a clause headed by another wh-word, *whether*. Historically, these **locality** constraints have been considered a central component of language competence (Ross, 1967; Chomsky, 1973; Rizzi, 1990; Cinque, 1990).

This work explores whether syntactic locality constraints may be better explained by a language-independent factor, memory. A parser implements reductionist theories (Clifton & Frazier, 1989; Clifton, Fanselow, & Frazier, 2006; Arnon, Snider, Hofmeister, Jaeger, & Sag, To Appear; Hofmeister, 2007; Hofmeister, Jaeger, Sag, Arnon, & Snider, 2007; Hofmeister &

Sag, 2010), and acts as an interpretation of what is happening in the brains of listeners as they hear locality-violating sentences such as (2). It is a working model of human sentence processing that is consistent with a variety of memory and syntactic locality theories. Rather than use grammatical constraints, the parser uses language-independent cognitive constraints derived from general cognitive research (Anderson, 1976, 2005; Wanner & Maratsos, 1978; Gibson, 2000; Lewis & Vasishth, 2005).

The parser's difficulty is calculated as it interprets English, German, Swedish, and Russian violations, including strong islands (represented by complex-NP island violations, **CNPs**), weak islands (represented by wh-island violations, **WHIs**), and non-island locality (represented by superiority violations, **SUVs**). The difficulty predicted by each of the cognitive constraints is compared to human difficulty. If the parser finds the same patterns of difficulty as the human without requiring explicit grammatical constraints, the model supports reductionist claims. If the parser does not find the same patterns, it indicates either a problem with the reductionist claims or that the phenomenon requires grammatical constraints.

The results demonstrate that the cognitive theories model weak and non-island locality without requiring specific grammatical constraints. Specifically, weak islands are best-modeled by reductionist theories that incorporate both activation-based and interference-based difficulty. SUVs, on the other hand, are a result of solely interference-based difficulty. The cognitive constraints do not, however, model the strong island data. As these results are supported by a working model, a rethinking of syntactic locality, and the competence-performance divide, is needed.

1.3 Significance

These results are directly applicable to work in linguistics, psycholinguistics, and computational linguistics. In linguistics, the results help illuminate the nature of syntactic locality, and highlight which phenomena are best explained by a grammatical approach. The goal of syntax is to encode a speaker's knowledge of language, or competence, in the simplest possible terms (Chomsky, 1965). Reductionist approaches take this aim to heart, demonstrating that some phenomena are not caused by a speaker's knowledge, but rather how that knowledge is put to use. By providing explanations that center on general cognitive constraints, reductionist accounts minimize the phenomena competence is accountable for, thereby simplifying the grammar. Yet these approaches are often met by criticism because they do not ground difficulty to explicit cognitive mechanisms, or they use questionable experimental methods that can not distinguish competence from performance factors (Phillips, In Press).

This work seeks to arbitrate this debate by providing explicit definitions of cognitive theories on a working computational model. Further, the work considers a variety of results that argue both for and against reductionist approaches. Testing explicit cognitive theories against a variety of cross-linguistic experimental data offers a broad examination of reductionist explanations for syntactic locality. This not only addresses the concerns of proponents of competence approaches, but it also provides an implemented test of the reductionist accounts.

The results indicate that those in favor of competence accounts of locality should focus their attention more on the nature of strong islands than on weak islands and superiority violations. Mathematically, a grammatical bound of movement is advantageous: unconstrained movement would create a grammar that over-generates and allows for constructions that are outside the realm of natural language (Stabler, 1997; Gärtner & Michaelis, 2007). Yet this constraint should be centered on phenomena like strong islands rather than weak islands or SUVs. The grammar can be simplified by relegating the latter to performance, as they are

modeled by general cognitive factors and also follow from working memory claims. Those in favor of a full competence account of all syntactic locality phenomena should focus on grammaticized constraints of specific working memory factors like activation and interference. A comparison of claims like Relativized Minimality (Rizzi, 1990; Cinque, 1990) to activation and interference may also lead to fruitful competence-based research.

On the other hand, psycholinguists and linguists in favor of performance-based accounts of syntactic locality should focus on the positive result for weak islands and superiority violations. These two phenomena not only behave as the cognitive theories predict, but they also demonstrate how phenomena long considered distinct in linguistic theory can also be distinguished by cognitive accounts.

Although reductionist arguments for strong islands are not supported in this work, this could be a result of the specific cognitive theories considered here. The retrieval-based theories work well for weak islands and superiority violations; this is not surprising because long-distance retrievals and similarity-based interference are both available for these phenomena. With strong islands, interference is not a factor, particularly for those cases considered in this work. Further, experimenal evidence indicates that a retrieval may not even be occurring (Stowe, 1986). In this case, a better cognitive model of strong islands would implicate storage difficulty rather than retrieval difficulty.

Determining which cognitive factors contribute to processing difficulty can often be as contentious in psycholinguistics as the competence-performance debate is. This research provides a methodology for testing the specific constraints on a broad-coverage parser. For example, it can be difficult to understand the differences between the Dependency Locality Theory (**DLT**) (Gibson, 2000) and retrieval (Lewis & Vasishth, 2005), two psycholinguistic theories sensitive to both activation and interference difficulty. Theoretically, the two theories take into account different factors to predict difficulty, yet they can often make the same predictions. A broad-coverage computational model can quickly compute predictions for these and other

theories, helping to define the similarities and differences, and the strengths and weaknesses, of each of these theories against real experimental data.

This research is also significant to psycholinguistics because it helps to strengthen the level of explicitness in performance-based accounts. Experimental results may implicate a quantity like decay or interference in human sentence processing difficulty for some phenomenon. But what is decay in the human parser? Is a simple metric like distance adequate to model the data? Or do the results require a more sophisticated model that operates on general cognitive predictions from connectionist calculations? Similarly, what is the nature of interference? Do specific types of words and phrases always create difficulty, or is difficulty the result of too many similar structures in memory? The methodology reported here provides a framework that can answer these questions for specific experimental results. This provides a level of explicitness required to address concerns against reductionist accounts of linguistic phenomena.

Finally, because this research is computational in nature, it takes advantage of many advances in natual language processing (**NLP**) and computational psycholinguistics. However, it is hoped that it offers something in return. First of all, cognitive constraints for syntactic locality may provide an important tool for parsing unbounded dependencies. Unbounded dependencies have recently become an important topic in statistical NLP because they are a challenge for parsing. As in language, the majority of the relationships in treebanks are local. This creates a skewed data set, which can be problematic for accuracy since short, local dependencies are given the same rating as long, non-local dependencies (Nivre, Rimell, McDonald, & Gómez-Rodríguez, 2010).

However, when only the long dependencies are considered, parser accuracy is dismally low on many of the state-of-the-art parsers, including the transition system considered in this work (Nivre et al., 2010). Humans do not have difficulty with these long-distance dependencies, particularly the grammatical examples that are provided in treebanks. Perhaps what is required is more human-like probabilistic features, which take into account the level of work-

ing memory difficulty a human is predicted to have with each unbounded dependency. The cognitive constraints considered here operate remarkably well on the unbounded dependencies for the syntactic locality data. This would further support previous approaches that have found human-like memory helpful in broad-coverage parsing (Schuler, AbdelRahmen, Miller, & Schwartz, 2010).

One important challenge facing computational psycholinguistics is a standardization of training, testing, and development sets for broad-coverage models. Broad-coverage models are preferable because they are full-scale working models, not toy grammars which only work for a few phenomena or experimental sets (Crocker, 2005). The same model can be used to test a variety of experimental data, and can also be challenged to account for a variety of phenomena. Standard data sets would not only help verify accuracy in human sentence processing models, but also bring computational psycholinguistics in line with computational linguistic practices (Keller, 2010).

Unfortunately, these data sets are not available at the time of this study. Further, it's not clear when they will be available. This methodology reflects an attempt to keep the broad-coverage model accountable even without the standardized data sets. The data tested consists of a variety of experimental results. In includes data that both supports and refutes the claims it is testing, and it incorporates a variety of theories that make specific and often contrasting predictions. Overfitting is not possible on this data set, particularly given the way that accuracy is measured. This methodology may be useful for other human sentence processing models that lack standardized data sets.

1.4 Chapter overview

Chapter 2: Syntactic locality

This chapter provides an overview of syntactic locality, and its central role in the competence and performance debate. It introduces Dependency Grammar (DG), and then discusses standard locality constraints for strong islands, weak islands, and superiority violations. Subsequent sections detail syntactic and semantic approaches to explaining data that challenges these constraints, and introduce arguments against a competence approach.

Chapter 3: Cognitive constraints

The second chapter discusses the major claims made by reductionists in support of a performance-based approach to syntactic locality. It provides background on the role of working memory in processing difficulty, and then introduces the specific cognitive constraints implicated in reductionist accounts. This chapter lays the foundation for the activation, interference, and combination constraints that are encoded in the computational model.

Chapter 4: A computational model

This chapter provides the implementation details of the human sentence processing model used in this work, a statistical dependency parser. It motivates the use of the Nivre non-projective transition system (2009) as a model. Although the parser is central to the work, there are a variety of peripheral implementation details crucial to this research. This chapter also details these, including training on treebank data, encoding cognitive theories probabilistically, and encoding complexity metrics that measure parser difficulty.

Chapter 5: Experiments

This work takes into account experimental data that either support or challenge reductionist approaches. This chapter discusses the categorization these experiments, organized by the specific phenomena considered in this work. For each phenomenon (strong islands, weak islands, and superiority violations), a series of experimental data was collected: classic experimental data that demonstrate the typical judgments of difficulty, gradience data that support reductionist accounts, and challenging data that is problematic for reductionist accounts. These experiments sample human difficulty across four languages (English, German, Swedish, and Russian), and also incorporate a variety of measures, including reading times and acceptability judgments.

Chapter 6: A methodology for cognitive modeling

This research is novel because it tests many factors at once. It uses a broad-coverage, cross-linguistic dependency parser, it tests against a variety of experimental data types, and it considers many different cognitive theories. The broad range of data, methodologies, and testable theories requires an explicit methodology that addresses the central question: which reductionist theories account for which syntactic locality phenomena? This chapter outlines the methodology, discussing how experimental data is prepared, how the parser is run, and how parser difficulty is compared to human difficulty.

Chapter 7: Results

The results section details how each of the cognitive constraints fare against each of the experimental results. These results are aggregated by phenomenon to offer a higher-level discussion on which cognitive theories best explain strong islands, weak islands, and superiority

violations. The results indicate that strong islands can not be modeled by the cognitive theories, but weak islands and superiority violations can. The broad implications of these results are then discussed.

Chapter 8: Conclusion

The final chapter of this dissertation provides a summary of the findings, as well as a comparison of these findings to other research. I then discuss how these findings contribute to the syntactic locality and competence-performance debate, and directions for future research.

CHAPTER 2

SYNTACTIC LOCALITY

2.1 Introduction

This chapter describes syntactic locality as a linguistic phenomenon, detailing its origins, how it has challenged prevailing theories, and how linguistic theories have expanded to encompass locality constraints.

Before discussing locality itself, this chapter provides a brief background on Dependency Grammar (DG) (Tesnière, 1959; Hays, 1964), the grammar used throughout this research. Section 2.3 discusses classic examples of the syntactic locality phenomena considered here: Complex Noun Phrase Islands (CNPs), Wh-Islands (WHIs), and Superiority Violations (SUVs). Section 2.4 then discusses exceptions to the constraints, and necessary accommodations within the theories. Section 2.6 outlines several limitations of grammatical theories that have led to various reductionist arguments provided in the next chapter.

2.2 Dependency grammar

DG analyzes sentence on the basis of word-to-word dependencies. Figure 2.1 depicts a dependency analysis of a simple English sentence. In this sentence, there is a **dependency** between *sailed* and *Diego*, with *sailed* being the **head** of *Diego* and *Diego* being the **dependent** of *sailed*. All the words in the sentence have a head except for the main verb *sailed*; this word is called the **root** of the sentence, and has special status. Dependencies generally follow standard linguistic patterns, with subjects and objects depending on the main verb, and constituents depending on phrasal heads.

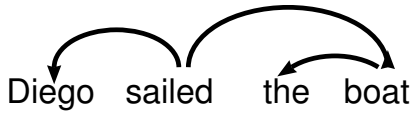


Figure 2.1: A DG example.

One way in which DG differs from more popular formalisms is that it does not have phrasal nodes; in fact, DG ultimately places little importance on constituency, and instead focuses on the obvious connections between words. The main advantage of this system is that it makes language description easier, particularly for non-configurational languages. The example in Figure 2.2 is from Czech, a language that has freer word order than English. The sentence translates to "A strong individual will obviously withstand a high risk better than a weak individual", but notice the discontinuity among dependencies. This discontinuity is easily handled by DG since constituents can appear anywhere in the sentence. The only complication is crossing dependencies, which become a problem for computational approaches as is described in Chapter 4.2.3.

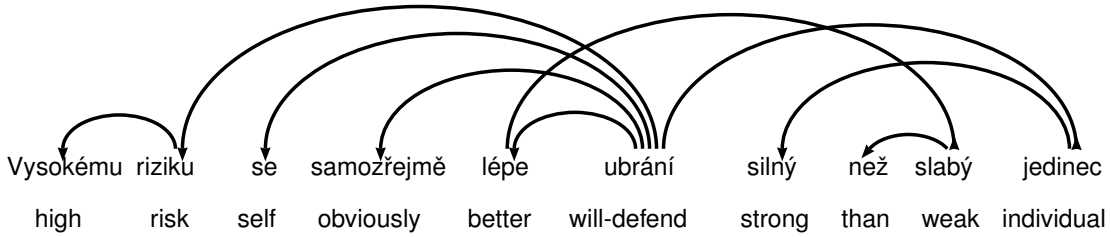


Figure 2.2: Czech, a non-configurational language, has sentences with discontinuous constituents. These sentences are difficult for phrase structure systems.

Though simple, DG is one of the oldest formalisms: evidence suggests that Pāṇini used dependencies to describe Sanskrit in the 4th century B.C. Its modern linguistic roots date back to the late 1950s, when Tesnière described sentences as sequences of "word-to-word connections" (1959). This simple analysis of syntactic structure was also found to be advantageous for computational applications, and was the syntactic component in early machine translation (Hays, 1964).

Mainstream linguistics came closest to embracing DG in the 1970s when lexicalized formalisms became popular. In X-bar theory (Chomsky, 1970), phrasal nodes are created from heads, which are often existing words within a sentence. Although X-bar theory made use of phrasal nodes and covert information, such as traces, it also shared with DG a focus on relationships between overt words.

In the 1980s, DG was recognized as an excellent tool for language description, particularly for non-configurational languages. Mel'čuk (1988) explored DG as a grammar for Czech, a language that poses many challenges for configurational grammars, as Figure 2.2 demonstrates. Through both Mel'čuk's work and the Word Grammar approach of Hudson (n.d.), DG was expanded to a full-scale descriptive and explanatory language theory.

Currently, DG is undergoing a resurgence in statistical NLP. This simple formalism can be easily inferred from treebanks, and has led to state-of-the art results[1]. Objectors argue that the formalism is too simple to describe the full complexity of language. In fact, despite efforts by Hudson and Mel'čuk, DG can not easily handle constituency, and it may be mathematically impossible for it to handle crucial notions in language such as c-command[2].

Although the grammar is too simple to accurately describe natural language, this could in fact be a feature. Dependencies are an incontrovertible aspect of language structure: all grammars must formalize the notion of dependency, and it provides a common thread across formalisms. This is particularly beneficial for a human sentence processing model, as the simplest possible grammar can be tested. Secondly, as was discussed above, DG will continue to be an important descriptive tool for non-configurational languages, as more restrictive formalisms are still not able to accurately handle many structures. Finally, DG has had and continues to provide important contributions to both computational and mathematical linguistics, as is further discussed in Chapter 4.2.

[1] See Chapter 4.2.1 for details.
[2] See Chapter 7.4.2 for details.

2.3 Syntactic locality phenomena

The majority of the dependencies between words in a sentence are local, but some are more distant. One common long-distance dependency occurs in questions, as in Figure 2.3. Here, the question word is at the front of the sentence, even though it should appear at the end of the sentence as the object of the verb *sailed*.

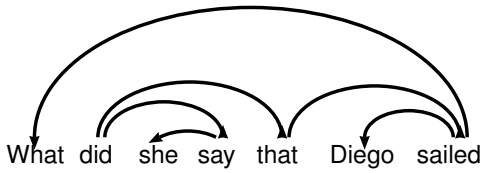


Figure 2.3: In English questions, the question word appears at the beginning of the sentence.

Locality conditions provide guidelines, or constraints, on long-distance dependencies. Figure 2.4 shows an ungrammatical sentence that includes an illegal long-distance dependency.

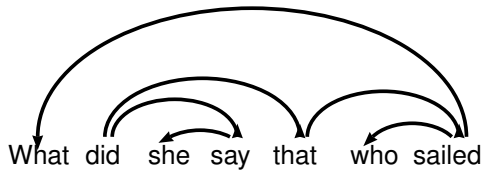


Figure 2.4: If another wh-word intervenes, the sentence becomes difficult.

The intervening wh-phrase, in this case *who*, prevents the long-distance dependency from being acceptable. Characterizing why this is the case is a central topic of Ross's seminal dissertation (1967), and has since become a central topic in theoretical linguistics and syntax. This dissertation considers three phenomena that violate locality conditions, leading to sentence difficulty. Two are examples of islands, and are described further in Section 2.3.1. The other phenomenon details SUVs, and is discussed further in Section 2.3.2.

2.3.1 Islands

Although explanations for syntactic locality exist in a variety of formalisms like Generalized Phrase-Structure Grammar (Fodor, 1992), Tree-Adjoining Grammar (Frank, n.d.), and Optimality Theory (Legendre, Wilson, Smolensky, Homer, & Raymond, 2006), this section focuses on research from the transformational grammar tradition. In this tradition, questions and other long-distance dependencies are conceptualized as movement. The idea is that the wh-word originates in its correct place thematically, usually near its head word. Then, because of discourse or other factors, the word is extracted from its base position and moves to a higher position in the sentence. A typical account of the sentences above in this framework is as in Figure 2.5. As can be seen here, *what* moves from a position in the complement of the VP to the beginning of the sentence, and leaves behind a trace, t_i[3].

The term **island** falls out naturally from this movement conception of language: islands are barriers to movement, and do not allow wh-words to pass through them (Ross, 1967). This work considers two types of islands: strong and weak. The main difference between the two is what can pass through: weak islands allow a prepositional phrase to extract, but strong islands do not (Cinque, 1990). Another distinction, more relevant given the experimental literature considered here, is based on what kinds of wh-phrases can extract. Weak islands allow some arguments to extract, but not adjuncts. Strong islands allow neither (Szabolcsi & den Dikken, 2002). (3) through (6) show examples from Szabolcsi and den Dikken (2002) that demonstrate this distinction. (3) and (4) are weak islands, where the island headed by *whether* is denoted within the brackets. Although extracting the adjunct *why* leads to difficulty in (4), extracting an argument like *when* in (3) does not. In the strong island examples in (5) and (6), on the other hand, both are difficult.

(3) ?When did John ask [whether to fire him]?

[3]Because the focus here is on aspects of syntactic locality pertinent to DG analyses, details like the Subject-Aux inversion are left out. See Adger (2003, p.341) for more details.

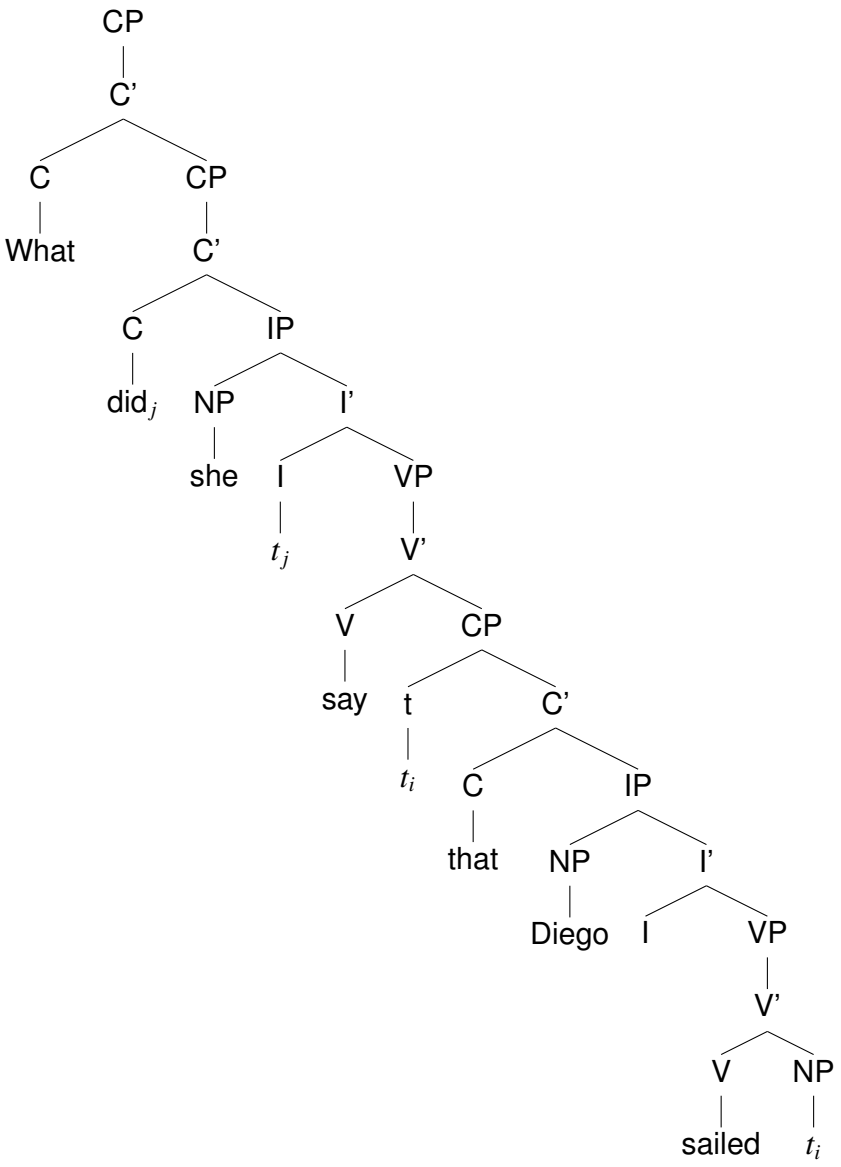


Figure 2.5: Question formation in the generative framework.

(4) *Why did John ask [whether to fire him]?

(5) *When did John bring [the girl who asked]?

(6) *Why did John bring [the girl who asked]?

The strong and weak island distinction does not reflect a difference in processing difficulty: strong islands are not considered harder than weak islands. Rather, it reflects what kinds of

phrases are allowed to extract.

There is often controversy regarding what is an argument and what is an adjunct, exactly how strong and weak islands differ, and what kinds of phrases can extract in different languages. Experimental evidence supports this controversy (Kluender, 1992; Keller, 1996; Yoshida, 2006). Even though the distinction between strong and weak islands is not always clear, many agree that it does exist (Szabolcsi & den Dikken, 2002); further, psycholinguistic work and even this computational study supports a distinction between the two on the basis of the competence and performance divide (Kluender, 1992, 1998). The following two subsections describe the particular strong and weak islands considered, CNPs and WHIs.

Strong islands

The strong island considered in this work is the Complex Noun Phrase Island, which doesn't allow extractions from within a definite noun phrase with a relative clause. A classic example is provided in Figure 2.6 (Hofmeister & Sag, 2010). Here, the dependency between *who* and *captured*, depicted with a dashed blue line, is illegal. This is because *who* is within a complex noun phrase with a relative clause, the phrase *the report that we had captured...*.

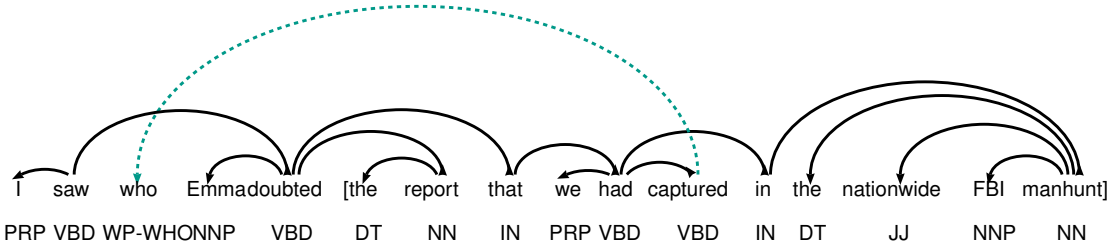


Figure 2.6: An example of a CNP violation.

There are other types of strong islands, including complex noun phrases with complement clauses, subject islands, and adjunct islands. The reader is referred to Szabolcsi and den Dikken (2002) for more information on these islands. This work focuses on this strong island

because the majority of the experimental evidence, particularly for English, includes CNPs with relative clauses.

Various locality hypotheses seek to explain why extraction from CNP islands is not acceptable. The first, from Ross (1967), is provided in (7):

(7) **Complex NP Constraint:** No element contained in a sentence dominated by a noun phrase with a lexical head noun may be moved out of that noun phrase by transformation (Ross, 1967, p.127).

Later, as linguists sought to create omnibus constraints that handled more cross-linguistic data, CNP islands were explained by subjacency (Chomsky, 1973, 1977). Subjacency argues that moved elements can not cross more than one bounding node. Bounding nodes for English are the determiner phrase (DP), or noun phrase, and S, the sentential phrase (currently IP).

The idea of bounding nodes is central to the current explanation for CNPs, the **ECP** (Empty Category Principle) (Aoun, Hornstein, & Sportiche, 1982). The ECP requires traces, or gaps in psycholinguistic terminology, be properly governed by their antecedents, or fillers. Proper government in this case is antecedent-government. Antecedent-government requires two things: that the filler c-command its gap (Reinhart, 1976), and that there be at most one barrier on the path between the filler and its gap. Figure 2.7 shows a fragment from one of the CNP island violations in an experimental study by Hofmeister and Sag (2010). The full island is *I saw who Emma doubted reports that we had captured...*. Given this analysis, we can see that *who* does c-command its trace, because the trace is a descendent of *who*'s sister node, IP. However, there are two barriers, or IPs, on the path between *who* and its trace t_i. Therefore, the sentence is in violation of the ECP.

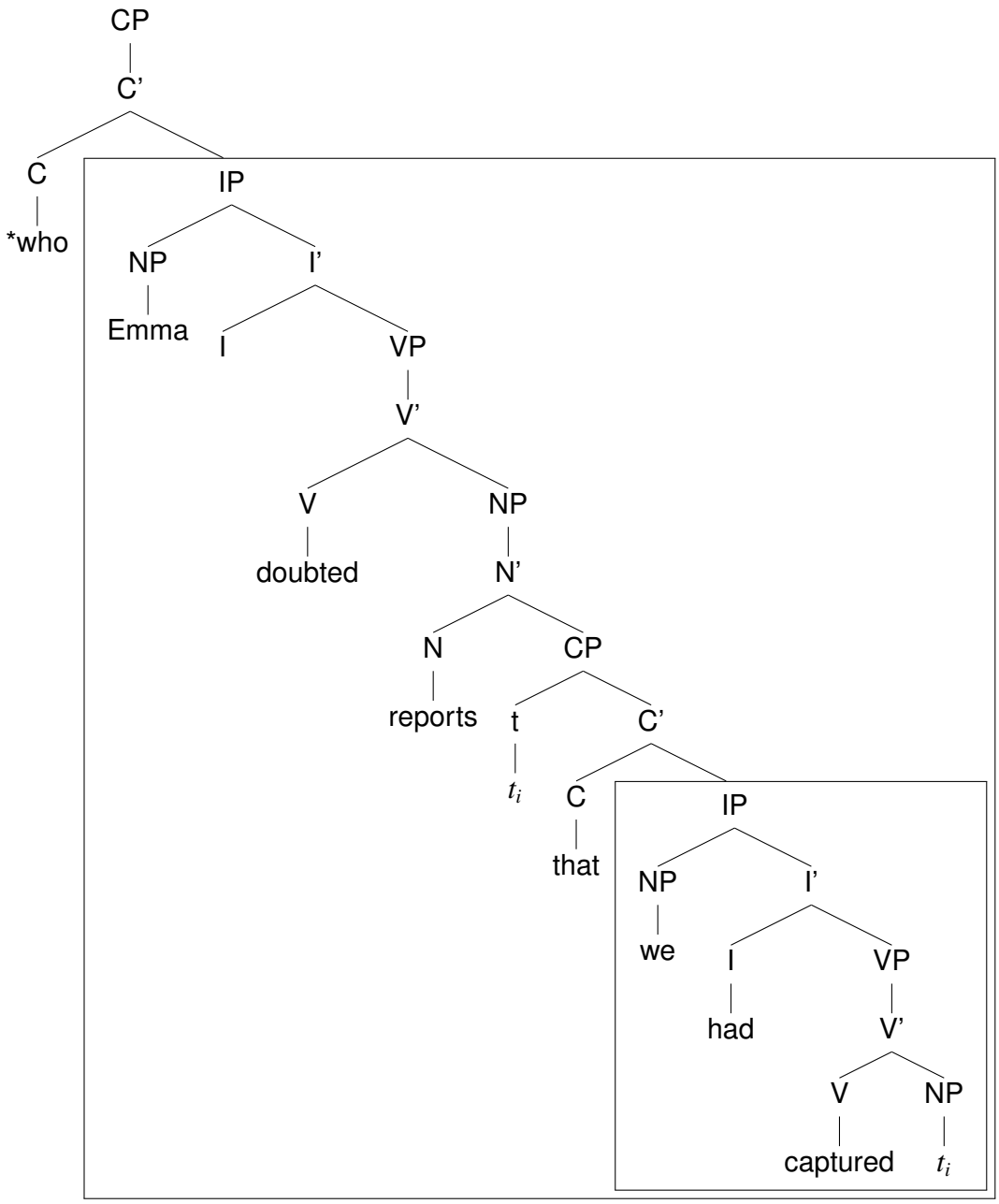


Figure 2.7: There are too many barriers between *who* and its trace, in violation of the ECP.

Weak islands

The standard weak island is the Wh-Island, where phrases headed by wh-words are islands to movement. Figure 2.8 provides an example from experimental data (Hofmeister & Sag, 2010). In this example, the wh-word *who* is moved past the wh-island phrase that begins with

whether.

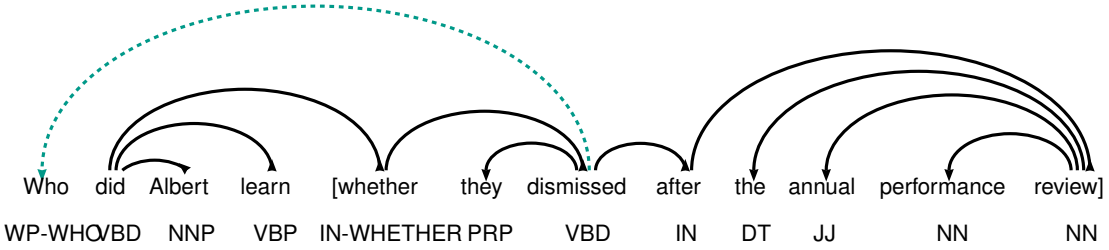


Figure 2.8: An example of a WHI violation.

The first definition of WHIs, from Ross (1967), was relatively broad, as (8) demonstrates:

(8) **Wh-Island Constraint:** Wh-words can not be moved across across other wh-words (Ross, 1967)

This definition is too broad because it encompasses two separate locality phenomena, WHIs and SUVs. Although Subjacency and the ECP restrict movement across the wh-phrase boundary, they are too constricting since weak islands allow some words to "escape". Therefore, theories like Relativized Minimality (Rizzi, 1990; Cinque, 1990) are considered to be more applicable to WHIs. These theories are discussed further in the next section on gradience.

2.3.2 Superiority violations

Unlike islands, SUVs do not require discussion of bounding nodes. Rather, the focus is on structural roles: a wh-word can not be moved across a "structurally higher" wh-word (Chomsky, 1973). Figure 2.9 shows an example. Here the object *what* is moved across the subject *who*.

The definition from Ross in (8) above accurately captures the intuition behind SUVs. This Superiority constraint is controversial because of the many well-known exceptions (Karttunen,

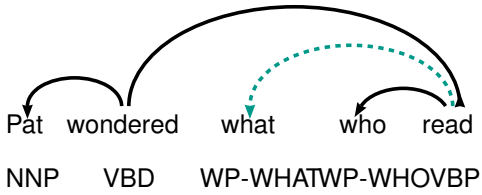


Figure 2.9: An example of a SUV violation.

1977). However, SUVs are considered part of this study because their difficulty in English is that SUVs are difficult in English is well-documented in the psycholinguistic literature.

2.4 Gradience in locality

One of the biggest sources of gradience in all three locality phenomena comes from the filler-type. Example (9) shows an SUV that is difficult because the wh-word *when* crosses *who*[4]. As noted by Karttunen (1977), replacing *who* with a *which-N* construction as in (10) leads to a more acceptable sentence.

(9) *When$_i$ did who eat the sandwich t_i?

(10) Which sandwich will who eat?

Pesetsky (2000) notes that this occurs whether the *which-N* is the extracted element or the intervenor, as demonstrated in (11).

(11) What will which contestant eat?

In both (10) and (11), the superiority constraint is violated, yet speakers report the sentences are more acceptable than the SUV in (9). These intuitions are supported by experimental evidence. Arnon and her colleagues (To Appear) examine sentences like those in Figure 2.10, confirming Karttunen's and Pesetsky's intuitions. Here, the which-N cases are more acceptable, given the experimental measure of Residual Reading Times (RRT), than the bare examples[5].

SUVs are even evident in naturally-occurring text. Examples (12) and (13) were discovered in Internet articles by Sag, Hofmeister, Arnon, Snider, and Jaeger (2008).

[4]It should be noted that in the generative tradition *who* also moves. However, because this movement does not affect the surface ordering, it is left out of these examples.

[5]Further details on this experiment are available in Chapter 5.5.2.

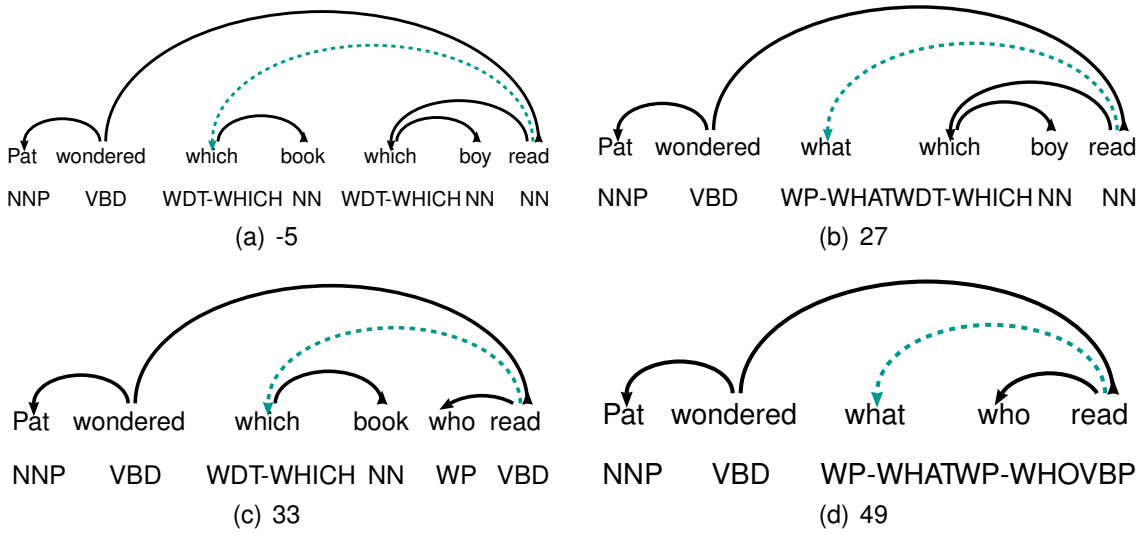


Figure 2.10: Experimental evidence of SUV gradience from RRT measurements.

(12) What did who know and when did they know it?

(http://www.antigonishreview.com/bi-113/113curb.html)

(13) What did who say and who did the asserting?

(http://www.thenation.com/doc/20030512/cockburn)

CNPs and WHIs also exhibit this kind of gradience. Figure 2.11 shows that *which convict* has faster residual reading times (and is therefore more acceptable) than the bare extraction, *who*, for the CNP. Figure 2.12 shows that the same pattern holds for WHIs with reading times. The reader is referred to Chapter 5.1 for details on these experiments from Hofmeister and Sag (2010).

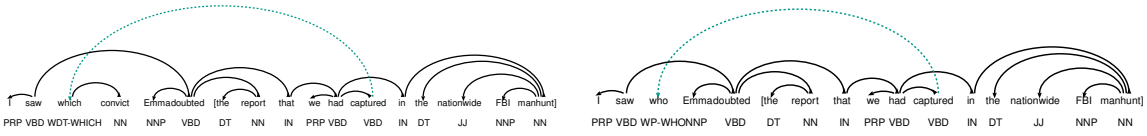


Figure 2.11: Experimental evidence of CNP gradience from RRT measurements.

(a) 243.74 (b) 255.96

Figure 2.12: Experimental evidence of WHI gradience from RT measurements.

2.5 Explanations from linguistic theory

The examples from the previous section demonstrate that the locality constraints can be violated. An adequate explanations has to take this into account. Much of the linguistic literature that accounts for examples of this kind of filler gradience focuses on WHIs. This section details this research.

2.5.1 A syntactic explanation for gradience

One hypothesis that addresses WHI violations is at the interface between syntax and semantics/pragmatics. The previous section provided examples demonstrating the acceptability of extracting *which-N* constructions from wh-islands. Pesetsky's D-linking (or *Discourse-Linking*) hypothesis (1987) argues that this is because *which-N* phrases delimit a set of possible entities, making them easier to extract than other wh-words. For example, the sentence in (10), repeated in (14), is acceptable for native English speakers even though *which sandwich* crosses the wh-island *who*.

(14) Which sandwich$_i$ will who eat t_i?

According to Pesetsky, this is because the content of the extracted element is a delimited set: it is a set of sandwiches. This is easier than an unrestricted set, such as the set referred to by *what* in (15).

(15) *What$_i$ will who eat t_i?

Here, *what* can refer to sandwiches or cookies or pizza or any kind of food. According to Pesetsky, *which-N* phrases are easier because they are linked to a limited set by the discourse.

Pesetsksy's D-linking hypothesis is extended by Rizzi (1990) and by Cinque (1990) to account for a variety of linguistic dependencies, including island violations. Their hypotheses use D-linking to differentiate between *referential* and *non-referential* determiner phrases (DPs) (Cinque, 1990). *Which-N* constructions are referential, or refer to "specific members of a preestablished set" (Cinque, 1990, p.8). Cinque and Rizzi both argue that referential DPs can be extracted from WHIs, and other structures, because their traces do not require a local co-index. In DG terms, words that refer to members of a preestablished set can be governed by words outside their immediate constituent. Non-referential DPs, on the other hand, must be governed by words that are within the constituent.

Taking the ungrammatical example in (15) above, Rizzi and Cinque would argue that the island will be violated because *what*'s trace is not governed locally within the phrase *who eat*. This results in ungrammaticality because *what* does not refer to specific members within a preestablished set. *Which sandwich*, on the other hand, does not need to be locally governed because it refers to the members of a preestablished set, {sandwiches}.

2.5.2 A semantic explanation for gradience

The referentiality argument promoted by Rizzi and Cinque provides syntactic motivation for gradience in wh-islands. However, the account prompts the next logical question: why are referential DPs special? This section describes approaches that provide semantic explanations.

de Swaart (1992) and Kiss (1993) claim that WHIs exist because they ensure wh-phrases

can receive the wide scope interpretation. The difference between wide and narrow scope is exemplified in (16). The sentence is ambiguous between two readings, one in which *everyone* takes wide scope over *which sandwich*, and the other where it does not. The wide scope reading gives rise to an interpretation where a respective sandwich is eaten by each individual. In the narrow scope reading, *everyone* does not take scope over *which sandwich*, leading to a reading where a single type of sandwich was eaten by each individual.

(16) Which sandwich$_i$ did everyone eat t_i?

a. Wide Scope: Diego ate ham and cheese, Phoebe ate tuna, Jason ate a club.

b. Narrow Scope: Ham and cheese.

According to Kiss and de Swaart, WHI violations occur because the wh-word can not receive the wide scope interpretation. Acceptable violations occur when words that do not negate the scope-taking, so-called "harmless intervenors" (Szabolcsi & Zwarts, 1993), are available. These harmless intervenors, usually verbs, allow for more acceptable violations.

Szabolcsi and Zwarts use this analysis to explain the variation in extractability in wh-words, which provides a basis for referential DP arguments. They posit a hierarchy of wh-words, provided in (17), based on how easily they extract from wh-islands (Szabolcsi & Zwarts, 1993, p.249). *Which-N* is easier to extract than *who*, which is in turn easier to extract than the rest of the hierarchy.

(17) which person(s) > who >what> who/what the hell > how, why

Szabolcsi and Zwarts (1993) explain this hierarchy in terms of set range: words like *which-N* can extract easily because they denote a set of unordered individuals. *How* and *why* generally do not denote individuals. They also often refer to ordered sets. For example, the phrase *how much money* from the ungrammatical sentence in (18) below refers to the ordered set of money.

(18) *How much money was Diego wondering whether his roommate would gamble?

This results in a new characterization of WHI gradience based on the set a wh-word refers to: rather than "specific members of a preestablished set" (Cinque, 1990), the set is further constrained to contain unordered individuals. This does not yet explain why unordered sets of individuals are easier to extract. Szabolcsi and Zwarts (1993) provide a somewhat cognitive explanation: they claim a wh-word that denotes a set of unordered individuals makes for faster lookup of that set. For example, the term *which sandwich* refers to an unordered set; using the wh-word, even when it is ungrammatical, helps the listener figure out the unordered set of sandwiches being asked about. This is not useful for other wh-words because they denote more ordered sets. In this case, extraction is not worth the violation.

There are other semantic accounts of island violations, most notably those of Kroch (1989) and Comorovski (1989) based on the plausibility of the presuppositions of sentences containing violations. However, these approaches have the same limitation: they can only explain gradience on the basis of unspecific claims of cognitive effects. The next subsection discusses this issue further.

2.6 Limitations of linguisic explanations

The previous subsection demonstrated a few linguistic accounts of WHI violation acceptability. However, they could only explain the variation in terms of the acceptability of *which-N* constructions; none could describe the naturally-occurring evidence reported in (12) and (13). Although this limitation is a problem, proponents of reductionist explanations of locality refer to a more striking limitation: competence can not motivate why gradience should exist at all. Hofmeister (2007), following Chung (1994), explains the issue as follows:

> ...underlying all these observations is the question of why the content (or amount of content) in the extracted element should alter acceptability, especially given that this contrast appears most strikingly in the context of supposedly universal syntactic constraints. (Hofmeister, 2007, p.2)

The limitation of the competence perspective is that it does not easily handle gradience. Addendums and restrictions to universal claims can lead researchers to claim the original constraints are “ad hoc and without independent motivation” (Hofmeister et al., 2007, p.187).

This criticism poses a problem for any competence-based explanation of gradience in syntactic locality. The grammatical constraints can explain why island violations are ungrammatical, but they can not easily explain exceptions to these constraints. Linguistic theories spanning syntax, semantics, and pragmatics can account for the gradience, but they can not explain why the gradience exists in the first place. If the gradience, and syntactic locality itself, can be explained by general cognitive constraints, the grammar can be simplified. This position is put to the test by the cognitive model considered here.

2.7 Conclusion

This chapter provides an overview of syntactic locality from a linguistic perspective. The focus is theories that provide a deeper understanding of locality principles, and can help challenge reductionist assumptions. However, as the last section highlights, a reductionist analysis could simplify the grammar and provide a natural explanation for the gradience in violations that is often found. The next chapter discusses how.

CHAPTER 3

COGNITIVE CONSTRAINTS

3.1 Introduction

The last chapter highlights several characteristics that promote syntactic locality's status as an important topic in linguistics. Some, like gradience, make islands and SUVs particularly suitable for a cognitive approach. This would not only offer a language-independent explanation, but also simplify the grammar. This **reductionist** approach to syntactic locality "reduces" the role of the grammar and uses processing factors to account for the difficulty (Fodor, 1978; Deane, 1991; Pritchett, 1992; Kluender, 1992; Kluender & Kutas, 1993; Kluender, 1998; Hawkins, 1999; Arnon et al., To Appear; Hofmeister, 2007; Hofmeister et al., 2007; Hofmeister & Sag, 2010). Yet, it is equally possible that this gradience in acceptability judgments comes from grammatical illusions (Phillips, In Press), and many experiments do not address this concern.

A computational model has the potential to address concerns against reductionism by explicitly defining cognitive explanations and testing them against the experimental data on syntactic locality. In the process, the computational model can either bolster or detract from reductionist accounts by addressing two important questions: does gradience fall out naturally from independently-motivated cognitive principles? And if so, do these cognitive principles accurately model human difficulty?

This chapter discusses the cognitive principles that support reductionist arguments for syntactic locality. Section 3.2 and Section 3.3 discuss the role of working memory, and more specifically activation and interference, in processing difficulty. Section 3.4 discusses specific activation, interference, and combination theories that have the potential to model syntactic locality.

3.2 Working memory in sentence processing

Working memory maintains the words, structures, and other information necessary to build an analysis as a sentence is parsed. But, it has a well-known limited capacity (Miller, 1956) which can lead to processing difficulty (Yngve, 1960; Lewis, 1996).

Psycholinguists usually separate working memory constraints into two categories: storage costs and processing or integration costs. Storage costs calculate the cost of maintaining a word, or in many cases an incomplete dependency, in memory (Kimball, 1973; Hakuta, 1981; Gibson, 1991; Lewis, 1993; Stabler, 1994; Gibson, 1998, 2000; Schuler et al., 2010). Processing and integration costs, on the other hand, are concerned with the cost of integrating a word into an analysis (Pickering, Barton, & Shillcock, 1994; Gibson, 1998, 2000; Lewis, 1999; Lewis & Vasishth, 2005).

The majority of the theories this work considers focus on integration costs because syntactic locality difficulty arises at the head of the wh-word (Stowe, 1986; Arnon et al., To Appear; Hofmeister & Sag, 2010). Experiments that use word-based measures, such as reading times, measure difficulty at the embedded verb, which would be *fire* in Figure 3.1. This is the integration site: where the long-distance dependency between the wh-word and its head is created. It is here that one would expect to see difficulty caused by memory decay and interference from other words, the two quantities most implicated in reductionist accounts of syntactic locality. The next section discusses each in more detail.

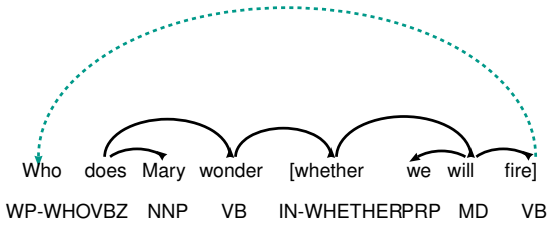


Figure 3.1: The integration site for the long-distance dependency is at *fired*.

3.3 Activation and Interference

The cost of integrating an item into an analysis can be high for various reasons, but the most interesting for syntactic locality has to do with inaccessibility[1]. While parsing a non-local dependency, for example the sentence in Figure 3.1, retrieving the wh-word and integrating it with the verb is difficult. There are two potential causes: the difficulty is because the word hasn't been accessed for a long time, or it is because there are other words in the sentence that interfere with it.

The first problem is one of activation (Brown, 1958), where an item decays in memory the longer it remains unused. Using the example sentence in Figure 3.1, the wh-word hasn't been accessed in memory for a long time. If it were the head of some other word in memory, it may have been reactivated and easier to access. But as it stands, the word is not needed until its head is parsed, which is seven words later in the sentence.

The WHI example demonstrates another possible cause of the difficulty, interference with other material (Lewis, 1996, 1999; Caplan & Waters, 1999; Gordon, Hendrick, & Johnson, 2001; Gordon, Hendrick, & Levine, 2002). The word *who* in Figure 3.1 has several features, one of which likely encodes the fact that it is a question word. If there are other question words in the sentence, as there are here, they would all share this feature, and would all likely be activated should a word be a suitable head. This leads to interference because in an associative memory, where words are accessed based on their features, it is difficult to access the correct wh-word (McElree, Foraker, & Dyer, 2003).

While activation and interference are relatively uncontroversial costs for working memory, it is not always possible to implicate one over the other (Anderson, 2002). A variety of theories posit how activation and interference affect sentence processing. The next section surveys the

[1] Another integration cost may come from expectation: the parser expects a certain word or type of word, but instead receives another. Chapter 4.6.1 discusses how the use of surprisal (Hale, 2001) as a complexity metric can help model this particular cost

theories that can explain syntactic locality difficulty.

3.4 Implemented theories

Reductionist accounts argue that what appear to be grammatical locality violations are in fact sentences with severe processing difficulty. Gradience in locality difficulty then naturally falls out from the cognitive approach: it is the result of alleviating difficulty. (19) lists a series of constraints that have the potential to explain locality processing difficulty and gradience, along with references citing their use in sentence processing. Many have been argued to affect locality constraints in particular (Hofmeister & Sag, 2010).

(19) **Distance:** Wanner and Maratsos (1978); Joshi (1990); Rambow and Joshi (1994); Hawkins (1990).

Activation: Deane (1991); Just and Carpenter (1992); Vosse and Kempen (2000); Just and Varma (2002); Lewis and Vasishth (2005).

Filler Load: McElree et al. (2003); Lewis and Vasishth (2005); Lewis, Vasishth, and Van Dyke (2006); Hofmeister (2007).

Intervenor Load: Gibson (1991); McElree et al. (2003); Hofmeister (2007).

Interference: Lewis (1996, 1999); Vosse and Kempen (2000); Gordon et al. (2001, 2002); Gordon, Hendrick, and Johnson (2004); Gordon, Hendrick, Johnson, and Lee (2006); Van Dyke and Lewis (2003); Warren and Gibson (2002); Lewis and Vasishth (2005); Lewis et al. (2006); Van Dyke and McElree (2006).

DLT Gibson (1998, 2000); Alexopoulou and Keller (2007); Demberg and Keller (2008); Demberg-Winterfors (2010).

Retrieval Lewis and Vasishth (2005).

There are a few notable exceptions to this list. For example, Bourdages (1992) argues that wrap-up effects are a key part of the difficulty in locality processing. Similarly, the difficulty in identifying the head of a wh-word has long been considered a main cause of unacceptability (Frazier, 1979; Fodor, 1979; Stowe, 1986; Pickering et al., 1994; Phillips, 2006). Each of these constraints was tested in initial versions of this research and did not perform well against the experimental data. This is likely due to a problem with the translation to probabilistic features,

and the architectural constraints posed by the parser. These theories are also incorporated into some of the higher level theories considered here, such as the DLT and retrieval. For this reason, they are left out of the discussion. The following subsections discusses those theories that are considered.

3.4.1 Activation theories

Distance

Distance as measured by string position, or the number of words between a governor and a dependent, has been argued to affect processing difficulty (Wanner & Maratsos, 1978; Joshi, 1990; Rambow & Joshi, 1994; Hawkins, 1990; Gibson, Pearlmutter, Canseco-Gonzalez, & Hickok, 1996; Gibson, 1998; Gibson & Thomas, 1999; Gibson, 2000). The hypothesis is that the further away a word is from its dependent, the more difficult it will be to create the dependency. Experimental results demonstrate that this has an effect on processing difficulty (Gibson, 1998; Gibson & Pearlmutter, 1998; Hawkins, 1999; Gibson, 2000). However, it can not explain all data. For example, many questions have long dependencies, but they are easier to process than some shorter sentences. For this reason, a more sophisticated version of activation is considered as well, decay.

Decay

Activation decay is sensitive to the decay that items held in memory are subject to, particularly if they have not been accessed for a long time. Previous research has considered activation of words and structures in memory, though most of this research focuses on connectionist and neural network systems (Elman, 1990, 1991; Stevenson, 1994; Tabor, Juliano, & Tanenhaus,

1997; Christiansen & Chater, 1999). The implementation of decay that this work considers is based on principles from the general cognitive framework **ACT-R** (Adaptive Control of Thought-Rational) (Anderson, 1976; Anderson & Lebiere, 1998; Anderson, 2005), following work by Lewis and Vasishth (2005). Here, an item in memory becomes less active (i.e., decays) as time passes in a sentence. If the word is used again, for example if it is required to build some sub-structure, it can be reactivated and is therefore easier to retrieve later in the sentence. The precise equations that calculate a word's decay, as well as how they are translated into a retrieval time, are discussed further in Section 3.4.3 and Chapter 4.6.2.

3.4.2 Interference theories

Filler load

Linguists have long argued that certain types of wh-words are more acceptable for extraction than others (see Section 2.5 for details). One possible psycholinguistic explanation for this variation is termed *filler load*, where filler refers to the extracted element. Rather than basing the variation on referentiality or semantic properties, the psycholinguistic arguments are based on how easy it is to retrieve a filler from memory.

Experimental evidence for this variation is provided by Hofmeister (2007), repeated below in (20). The WHI violation in (20a) is the most difficult sentence, and the non-island in (20c) is the easiest. The *which-N* condition in (20b) falls in between.

(20) a. **Who** did Albert learn *whether* they dismissed after the annual performance review?

b. **Which employee** did Albert learn *whether* they dismissed after the annual performance review?

c. **Who** did Albert learn *that* they dismissed after the annual performance review?

Hofmeister explains this variation in terms of *filler load*: more informative fillers, like (20b), decrease processing difficulty because they have more features in memory. For example, **which employee** is more informative than **who**; it therefore has more features in memory than **who**. When the processor is trying to retrieve the wh-word from memory, it is easier to remember a word with more features than one with less (McElree et al., 2003; Lewis & Vasishth, 2005; Lewis et al., 2006). Therefore, the type of filler has an effect on how difficult it is to process the wh-word.

Intervenor load

One of the most prominent theories in psycholinguistics, the Dependency Locality Theory (DLT) (Gibson, 1998, 2000), bases memory difficulty on the number of unresolved discourse referents. Gibson uses this measure because psycholinguistic results suggest that the types of words intervening between a dependency affect its processing. In other words, a parsed word that requires a lot of memory may contribute to processing difficulty as syntactic relations are built, even if these relations do not involve the word. This provides an argument for **intervenor load** being a source of processing difficulty.

The intervenor load can be determined in various ways. Following Gundel, Hedberg, and Zacharski (1993), Gibson (1998) argues for an **Accessibility Hierarchy** of discourse referents. There are a variety of ways to refer to people in a discourse. They can be referred to as indexical pronouns (*you*), short names (*Luke*), full noun phrases (*the doctor*), or referent pronouns (*they*). Each of these is demonstrated in the sentences in (21) from Gibson's experimental conditions. In the conditions, the discourse referents intervene in the dependency between *student* and *copied*.

(21) a. The student who the professor who **I** collaborated with had advised copied the article.

b. The student who the professor who **Jen** collaborated with had advised copied the article.

c. The student who the professor who **the scientist** collaborated with had advised copied the article.

d. The student who the professor who **they** collaborated with had advised copied the article.

Using acceptability judgments, Gibson found that the indexical pronouns, (21a), are "significantly easier to process than any of the other three conditions" (Gibson, 1998, p.18).

Although this experiment does not involve locality data, Hofmeister (2007) argues that intervenor load affects locality processing. This cognitive constraint would explain gradience because intervenors that are more accessible, in terms of Gibson's hierarchy, are easier to process. Because they're easier to process, they interfere less with the dependency between a wh-word and its governor, leading to more acceptable judgments.

Interference effects

It can be difficult to differentiate the memory-based constraints posited in the literature. DLT includes a measure of intervenor difficulty which is different from intervenor load; both have something to do with *interfering* with dependencies, but do not specify how. These constraints are in fact related, but they are sensitive to different aspects of memory. The DLT is a calculation, or metric, of intervenor load that relates to memory difficulty. Intervenor load specifies that some words are just more difficult as intervenors, and is sensitive to these words. And interference effects arise when words, regardless of whether they are discourse referents, cause difficulty because they share the same features as other words (McElree et al., 2003; Lewis & Vasishth, 2005; Lewis et al., 2006).

In the case of syntactic locality, a word that shares the same syntactic, lexical, or semantic features as the wh-word would add processing difficulty during parsing. This has been argued to be a possible cognitive factor in syntactic locality difficulty (Hofmeister et al., 2007). This hypothesis has not been explicitly tested against locality data, although there is ample evidence of interference effects in other domains of processing (Gordon et al., 2001, 2002, 2004, 2006; Van Dyke & Lewis, 2003; Warren & Gibson, 2002; Lewis et al., 2006; Van Dyke & McElree, 2006).

In this work, interference is encoded as both a soft constraint in the parser, as the other cognitive constraints presented in this section, and as part of a complexity metric, retrieval. For both, interference is calculated as similarity-based interference (**SBI**), a quantity based on the diminished activation a word has if there are other similar words in the sentence (Anderson, 1976; Lewis, 1999; Anderson, 2005; Lewis & Vasishth, 2005). This hypothesis will likely be most important for sentences that have multiple wh-words, like WHIs and SUVs, and less important for those that do not.

3.4.3 Combination theories

DLT

The DLT (Gibson, 1998, 2000) links a preference for local attachments (i.e., activation) and integration costs (i.e., interference) into a prolific complexity metric for sentence processing. The hypothesis is given in (22).

(22) "The structural integration complexity depends on the distance or *locality*[2] between the

[2]Gibson employs the term *locality* to mean linear distance. This is different from the original notion of syntactic locality that is used as a cover-term for such things as island phenomena in linguistics. The linguistic version refers to *structural* distance. Though roughly related, the two terms do not refer to the same thing. In this dissertation, the term *locality* refers to the syntactic definition of locality.

two elements being integrated" (Gibson, 2000, p.102).

Although the DLT refers to the distance between words, this distance is not measured by string position. Rather, it is measured in terms of the number of discourse referents introduced between a dependent and its governor. For example, in the sentences provided in (23) through (25), *the nurse* and *the administrator* are new discourse referents. They therefore increase the DLT integration cost of attaching *nurse* as a dependent of a later verb. This DLT cost at each word is depicted below the sentences, and is supported by experimental evidence (Grodner & Gibson, 2005).

(23) The nurse **supervised** the administrator while...
0 1 **1** 0 1 1

(24) The nurse from the clinic **supervised** the administrator while...
0 1 0 0 1 **2** 0 1 1

(25) The nurse who was from the clinic **supervised** the administrator while...
0 1 0 1 0 0 1 **3** 0 1 1

Note that the cost of new discourse referents can be thought of as a storage cost: the DLT measure goes up once the discourse referent is introduced, and is unloaded after it has found its head. However, this work only considers difficulty at the head of the verb, which translates to retrieval difficulty. Although the DLT combines the two, it is encoded as a retrieval metric.

Retrieval

The most complete model of integration effects in language is provided by the retrieval theory (Lewis & Vasishth, 2005). Retrieval is sensitive to both activation and interference, but unlike many sentence processing models, it is based on general cognitive principles. Lewis and Vasishth (2005) build upon the ACT-R framework, a hybrid model featuring both a symbolic

and subsymbolic (or connectionist) component. At the lower level, a series of mathematical equations calculate the difficulty imposed by activation and interference during processing. At the higher level, a production system determines what has to be retrieved from memory, and integrates new words with the structure being built. ACT-R is a full cognitive framework; it has been applied to complex human behavior such as problem solving, learning, individual differences, and attention.

Lewis and Vasishth (2005) apply this cognitive framework to language, specifically sentence processing. They build a production system for sentences, and use the same subsymbolic equations for activation and interference used in other ACT-R models to determine the amount of time it takes to retrieve items from memory. The system includes predictions for how long it takes to do other language-related tasks (for example, decide to retrieve an item from memory, and build an analysis), and the end result is a prediction in milliseconds for how long it takes to read a sentence. Their model offers a concise memory-related complexity metric that can be used for comparison to human data, and results indicate it is a successful model of human difficulty (Lewis & Vasishth, 2005; Lewis et al., 2006; Vasishth, Brüssow, Lewis, & Drenhaus, 2008; Boston, Hale, Vasishth, & Kliegl, 2011).

Because of its combination of activation and interference quantities, retrieval has been argued to be an accurate model of syntactic locality by other researchers (Hofmeister & Sag, 2010). This work is the first test of this prediction. Chapter 4 provides details on how the theory is implemented in this parser, both as a probabilistic feature and as a complexity metric akin to the original Lewis and Vasishth model.

3.5 Conclusion

This chapter highlights working memory constraints that can explain syntactic locality from a cognitive perspective. Many of these constraints have a long tradition in psycholinguistics, and are at the heart of many theories. This research encodes the sentence processing theories that have the best potential to explain syntactic locality, and then tests them against the psycholinguistic results themselves. The following chapters describe how.

CHAPTER 4

A COMPUTATIONAL MODEL

4.1 Introduction

This chapter discusses the implementation of a non-projective dependency parser as a cognitive model of syntactic locality. Although the parser is the central component of the cognitive model, the implementation includes many tools and processes, as Figure 4.1 demonstrates. These processes address peripheral considerations for determining how well cognitive constraints explain syntactic locality. For example, the model has to incorporate probabilities from treebanks in four different languages (steps (1) through (4), discussed in Section 4.4). It also has to convert cognitive theories into these probabilistic features to inform parser decision (steps (5) through (9), discussed in Section 4.5), and derive parser difficulty measures over these features (steps (10) and (11), discussed in Section 4.6). Section 4.7 discusses a tendency for longer dependencies to be more difficult, and how this can be over-ridden in the parser. A cognitive model must have a method for determining parser accuracy, which is detailed in Section 4.8. Before discussion of each of these crucial implementation features begins, though, Section 4.2 details why this specific dependency parser was chosen as the cognitive model.

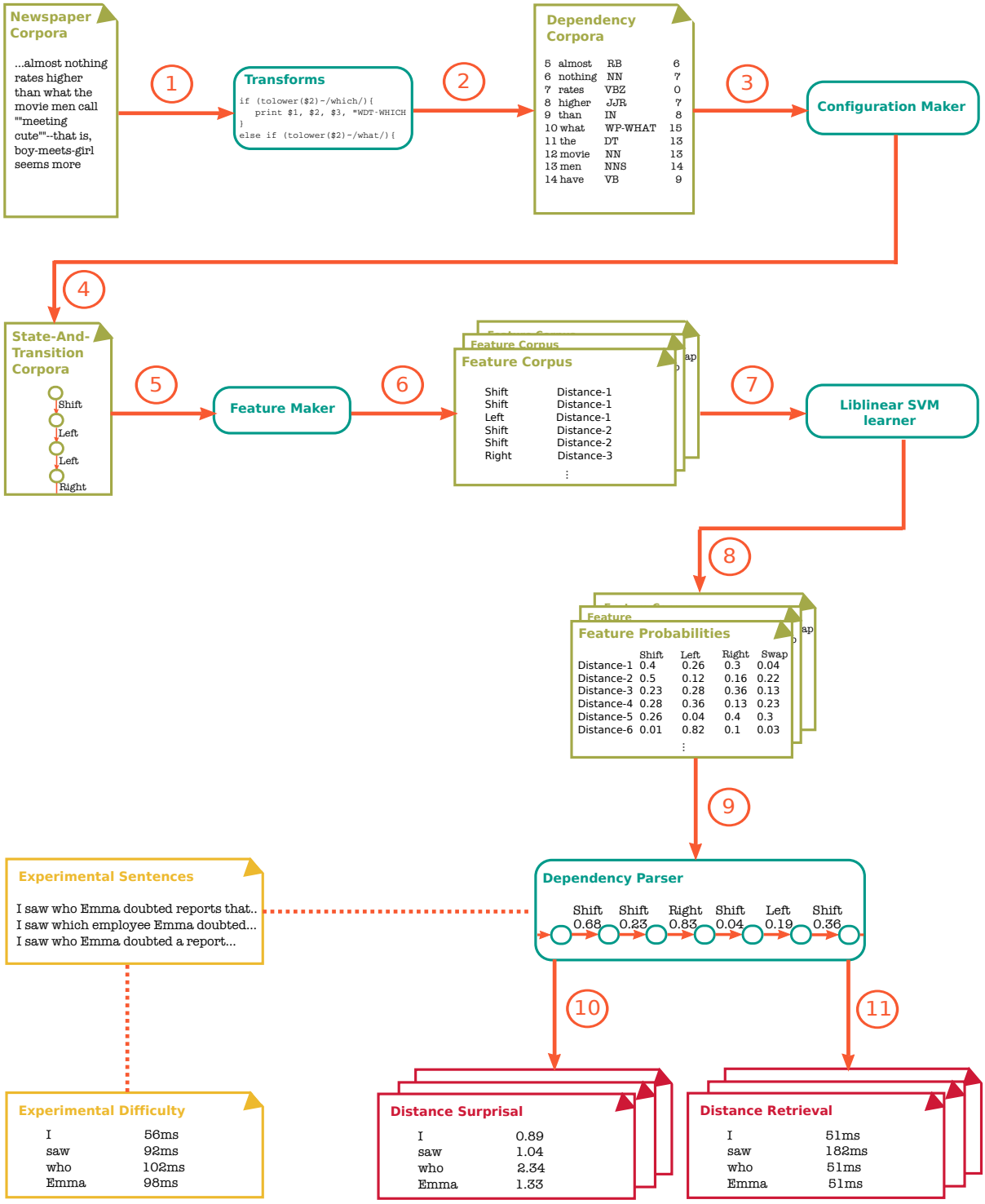


Figure 4.1: The parts of a cognitive model.

4.2 Why a dependency parser?

There are several language models that could have been used for this research, many already implemented. Instead, this work implements a non-projective, transition-based parser operating on dependency grammar. The following subsections outline the suitability of this type of parser to address these research questions.

4.2.1 Why dependency grammar?

Most grammar-based human sentence processing models use context-free grammars (**CFG**) or grammars that follow from that tradition because they are well-supported in computational linguistics (Roark, 2001; Hale, 2001; Levy, 2008)[1]. A variety of algorithms provide good performance, and the context-free grammar provides adequate coverage for many psycholinguistic phenomena. Human sentence processing models that depart from this tradition gravitate towards more restrictive mildly context-sensitive (**MCSG**) formalisms that are in keeping with human grammar abilities (Hale, 2003; Demberg-Winterfors, 2010; Grove, 2011).

This parser instead uses dependency grammar, a formalism that is too simple to adequately model the mildly context-sensitive nature of human grammar. Because dependency grammar relies solely on word-to-word dependencies, it does not model the hierarchy and constituency available in acceptable sentences. Chapter 2 provides more details on the limitations of dependency grammar; this section details several advantages it provides for cognitive modeling.

The first advantage is that dependency grammar avoids linguistic controversies. Although MCSGs are mathematically equivalent, they are descriptively different in terms of how they

[1] There is also a class of human sentence-processing models that do not use grammars at all, such as those based on neural networks (Elman, 1990, 1991; Tabor et al., 1997; Tabor & Tanenhaus, 1999; Christiansen & Chater, 1999). This discussion focuses on symbolic models.

analyze linguistic sentences. But, the notion of dependency is inherent within CFG and all mildly context-sensitive formalisms, a fact formalized by the dependency-generative nature of grammars (Kuhlmann, 2007). Dependency grammar therefore takes into account an aspect of grammatical analysis that everyone agrees is a part of grammar; it avoids the more controversial question of what else is required.

Dependency grammar provides another advantage for cognitive models: simplicity of design. One can readily see what each parser state holds within the Nivre transition system, and can analyze the respective contributions of particular analyses without the complications that a more complex grammar, and hence more complex parser, would have. This level of explicitness in the model allows for an easy examination of the contributions of the different cognitive hypotheses in explaining syntactic locality.

Finally, previous work with this dependency parser indicates that a parser based on dependency grammar is adequate in modeling psycholinguistic difficulty (Boston, Hale, Patil, Kliegl, & Vasishth, 2008; Boston et al., 2011). Boston et al., 2011 demonstrate that surprisal and retrieval values from this parsing model predict fixation durations in a German eye-movement corpus, indicating that the parser models sentence processing difficulty. This research also demonstrates that dependency grammar can model aspects of difficulty that a CFG parser can not: Boston et al., 2008 compare surprisal values from this parsing model to surprisal values from a **PCFG** (probabilistic CFG) parser (Levy, 2008) on the same German corpus. Results demonstrate that the dependency model accounts for different types of difficulty than the PCFG model. Finally, Buch-Kromann, 2006 discusses a psycholinguistically-motivated version of dependency grammar, which may be turned into a promising model.

There are a variety of practical reasons for choosing a parser based on dependency grammar, ranging from simplicity to state-of-the-art performance in standard NLP tasks. Previous research that demonstrates the adequacy of dependency grammar in modeling human sentence processing, as well as its ability to model differently from CFG parsers, helped to solidify

the decision to continue development with dependency grammar for this research.

4.2.2 Why the Nivre transition system?

Although dependency parsing has a long tradition in computational linguistics (Hays, 1964; Milward, 1994; Eisner, 1996b), interest within the NLP community grew once parsers using this simple grammar delivered state-of-the-art parsing accuracy (Yamada & Matsumoto, 2003). This resurgence gave rise to a variety of methods for dependency parsing. Some, like the graph-based parsing techniques of the MST parser (McDonald, Pereira, Ribarov, & Hajiĉ, 2005) and the k-best Maximum Spanning Tree parser (Hall, 2007; Hall, Havelka, & Smith, 2007) can not be used as psycholinguistic models because they are not incremental: they build dependencies by taking into account the entire sentence.

That leaves grammar-based and transition-based dependency parsers. Grammar-based parsers work on an explicitly-defined grammar, and operate similarly to standard CFG parsers by using the same algorithms, such as the Cocke-Kasami-Younger (CKY) and Earley parsing algorithms (Earley, 1970; Eisner, 1996b, 2000). They include the pioneering work of Hays, 1964 as well as over a decade of work by Eisner and his research group(Eisner, 1996b, 1996a, 2000; Eisner & Smith, 2005, 2010). A grammar-based dependency parser could have worked well as a cognitive model, particularly because of its similarity to the CFG parsers often used as human sentence processing models.

Instead, this work develops a transition-based parser that follows from the language modeling traditions of the 1970s and 1980s (Kaplan, 1972; Kimball, 1973; Marcus, 1980; Berwick & Weinberg, 1984). There have been a variety of transition-based dependency parsers developed: Yamada and Matsumoto (2003) provide an arc-standard algorithm that provided the state-of-the-art parsing accuracy results for English, and Attardi (2006) created an incremental,

non-projective parser from this model. Covington (2001) highlights a variety of transition-based algorithms for dependency parsing as well.

Although many of these grammar-based and transition-based dependency parsers can be adequate models of human sentence processing, this work uses the transition system defined by Nivre (2004, 2006, 2008, 2009). The original decision to work with the Nivre transition system as a parsing model was based on its simplicity and its incrementality: Nivre's interest in human sentence processing contributes to the parser's adaptability to modeling psycholinguistic data. Further, previous research demonstrates its potential for modeling difficult and psycholinguistically-interesting sentences (Boston et al., 2008, 2011).

Nivre and his research group have implemented this transition system as the MaltParser, which provides state-of-the-art results in dependency parsing shared tasks (Nivre et al., 2007). One possible research direction would have been to use the MaltParser for modeling syntactic locality because of its high level of accuracy. But, this implementation was created and extended for several reasons. First, this model is demonstrably useful in psycholinguistic modeling despite its lower accuracy. Further, the MaltParser is tuned for accuracy in parsing standard corpora. Although it can be used incrementally for human sentence processing modeling, it is unclear how many of these tweaks make assumptions that go against human sentence processing theory. Finally, it would be difficult if not impossible to get word-by-word complexity metrics like surprisal and retrieval scores from the MaltParser because of the way it has been implemented.

4.2.3 Why a non-projective model?

One advantage of using DG for linguistic analysis is that it allows for *non-projective* structures, or sentences with crossing dependencies. Non-projective structures are often difficult

for other formalisms, but DG allows word-to-word dependencies anywhere in the sentence. Figure 4.2 shows an example of an English sentence that has these crossing dependencies (Nivre, 2008). The head *hearing* is discontinuous from its dependent *on the issue*. The word *scheduled* intervenes, even though *scheduled* is not itself a dependent of *hearing*. Similarly, the dependency between *scheduled* and *today* is discontinous because it is interrupted by the phrase *on the issue*. These discontinuous dependencies give rise to what looks like crossing dependencies in the arcs, which can be very difficult to analyze with rewriting rules. This is despite the fact that most speakers have no difficulty understanding this sentence.

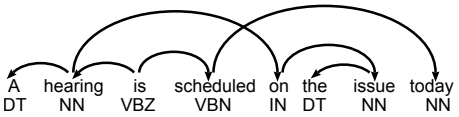


Figure 4.2: A non-projective analysis of an English sentence.

Non-projective sentences have a greater potential for ambiguity than projective sentences because of this discontinuity (Satta, 1992): if all the dependents of a head were required to be adjacent to the head or its dependents (i.e., continuous), the number of potential heads for a dependent is much smaller. If, on the other hand, dependents need not be adjacent to their heads, the number of potential heads for a word increases, as does the search space. Projective dependency parsing can be done in linear time, or contingent directly on the number of words that are in the sentence. Non-projective dependency parsing, on the other hand, can in the worst case take twice as long as the number of words in the sentence (Covington, 2001). This does not bode well for a cognitive model, as we know that humans are able to understand language in linear, not exponential time.

The original Nivre transition system is limited in that it can only parse projective structures (Nivre, 2004). This allows it to parse sentences in linear time, as a human would, but it also means that the parser can not model sentences like the one in Figure 4.2. This is a problem for the research planned here because some experimental sentences are non-projective, as demonstrated by CNP example from the Hofmeister and Sag (2010) experiment in Fig-

ure 4.3. The dependency between *captured* and *who* crosses the dependency between *saw* and *doubted*.

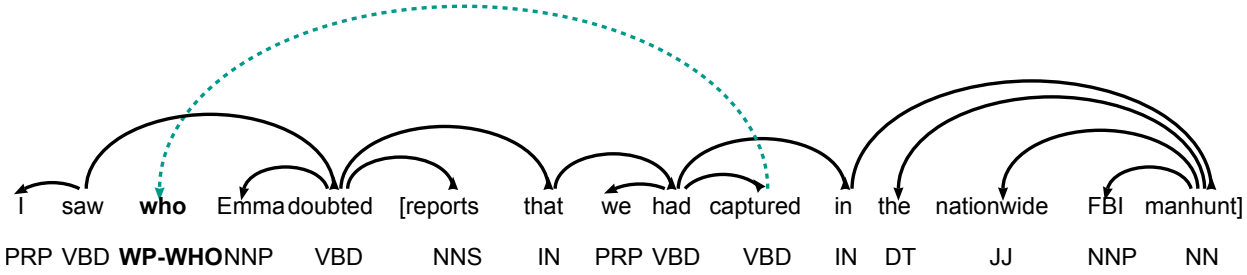


Figure 4.3: A non-projective experimental sentence.

It is therefore necessary to build a parser that can handle non-projective structures, even if the parser will not be able to parse in linear time. For the Nivre transition system, there are two possibilities: pseudo-projective parsing (Nivre & Nilsson, 2005) combines a dependency-changing "lift" operation (Kahane, Nasr, & Rambow, 1998) with post-processing to turn non-projective structures into projective structures and then back again. The idea behind pseudo-projective parsing is that the arc that causes the non-projective analysis (i.e., the crossing dependency) is lifted along its ancestors until it is projective. In other words, the child node becomes a dependent of the grandparent, the great-grandparent, etc., successively until it is part of a continuous dependency.

Although this change creates a projective structure from the non-projective sentence, the linguistic analysis has changed. In Figure 4.3 *who* would successively become the dependent of *had, that* and *doubted* before it would be projective. This is a problem for cognitive theories like retrieval, which depend on the difficulty in building analyses based on how far apart two words are. Although a post-process turns the projective sentence into the original non-projective sentence, at parse-time retrieval for *who* would be relatively simple because *doubted* is the next word.

Pseudo-projective parsing is therefore not the best way to model non-projective experimental sentences cognitively. That leaves Nivre's non-projective transition system, which will

be described in the next section.

4.3 The Nivre non-projective transition system as a cognitive model

Nivre (2004, 2006) defines a state transition system for incremental dependency parsing, which he later extends for non-projective parsing Nivre (2008, 2009). This cognitive model is based on the non-projective system.

Nivre defines each parser state as consisting of a tuple, in (26).

(26) state = (Σ, B, A)

Σ consists of a stack of already-parsed words that still require heads or dependents. B is a buffer of upcoming words, and A holds the dependency analysis information. The transition system handles non-projective analyses by allowing already-parsed words to be pushed back onto the buffer B so that the sentence can be reordered and attachments can be made locally. One problem with this system, from a cognitive model perspective, is that buffered, already-parsed and unparsed words are held in one data structure, B, which could lead to problems for incrementality. For example, the parser should be blind to any words that haven't been parsed yet, but already-parsed words should be available. A data structure is therefore added to Nivre's tuple, T, as in (27). Following Nivre (2004), T holds all words that have not yet been parsed. B only holds words that have been taken off the stack for reordering.

(27) state = (Σ, B, T, A)

To transition from one state to the next, Nivre defines four actions: `Left-Arc`, `Right-Arc`,

Shift, and Swap. Aside from the distinction between $\texttt{Shift}_\beta$ and $\texttt{Shift}_\tau$, the definitions in Table 4.1 are directly from Nivre (2009, p.353). Following his conventions, Σ's top appears on the right and B and T's tops appear on the left. Dependencies can be formed only between the

Transition			**Definition**	**Condition**
Left-Arc	$([\sigma\|i, j], B, T, A)$	$\Rightarrow$	$([\sigma\|j], B, T, A \cup \{(j, i)\})$	$i \neq 0$
Right-Arc	$([\sigma\|i, j], B, T, A)$	$\Rightarrow$	$([\sigma\|i], B, T, A \cup \{(i, j)\})$	
$\texttt{Shift}_\beta$	$(\sigma, [i\|\beta], T, A)$	$\Rightarrow$	$([\sigma\|i], \beta, T, A)$	$\beta \neq \emptyset$
$\texttt{Shift}_\tau$	$(\sigma, B, [i\|\tau], A)$	$\Rightarrow$	$([\sigma\|i], B, \tau, A)$	$\tau \neq \emptyset$
Swap	$([\sigma\|i, j], \beta, T, A)$	$\Rightarrow$	$([\sigma\|j], [i\|\beta], T, A)$	$0 < i < j$

Table 4.1: Transitions for the Nivre non-projective transition system, where Σ is the stack memory, B is the buffer for swapped elements, T is the input list, and A holds the dependency analysis so far.

two top elements in the stack, σ_1 and σ_2. For Left-Arc transitions, σ_1, or j, becomes the head of σ_2, i, and i is removed from the stack. For Right-Arc transitions, the opposite occurs: σ_1, or j, becomes the dependent of σ_2, i, and j is removed from the stack. In the implementation, Shift is one action: until β is empty, Shift pops elements off of β and pushes them onto σ. Once β is empty, Shift pops elements off of τ. In this way, the extra data structure does not change the way the parser itself works.

Finally, the Swap action is what gives this parser its ability to handle non-projective structures. Essentially, Swap reorders elements so that two discontinuous elements can be side-by-side in the stack as σ_1 and σ_2. Swap pops σ_2, in this case i, off of σ and pushes it onto β. This allows a new word, originally σ_3, to be available for dependencies with σ_1. Additionally, this reorders the sentence: if σ_2 is pushed back onto the stack, the two words are inverted. Figure 4.4 demonstrates how these parser actions work on an experimental sentence.

The diagram shows the sequence of parser actions required to parse the CNP example sentence in Figure 4.3. The diagram zooms in on the states required to parse the word *captured*. The left portion of the diagram shows the σ, τ, and β data structures, and the right portion shows the dependency analysis in A. Words already deleted from the stack are marked

in gray; these are words that require no further parsing. The current top stack elements are marked in bold. A `Shift` action brings *captured* onto the stack σ. Although there is a dependency between *had* and *captured*, this dependency can not be created yet because *captured* has not yet found all of its dependents. In particular, it needs to attach the extracted wh-word *who* from the beginning of the sentence. In order to get to *who*, which is further down on the stack, a sequence of `Swap` transitions need to take place.

Captured is first swapped with *had*, then *that*, and then *doubted*. Notice that with each swap, the swapped word (marked in green) is pushed onto β until *captured* and *who* are σ_1 and σ_2 respectively. At this point, a `Left-Arc` transition occurs, connecting *who* to *captured*.

Although the sentence can remain in this order in an NLP model, a cognitive model requires as much of the original input as possible. Therefore, once the non-projective dependency has been made, a series of `Shift` and `Swap` transitions reorder the sentence to its original order. This is done after every non-projective dependency; although this increases the parse time for the sentence, it ensures that measures like distance and activation remain as they do for humans.

This parser keeps with the Nivre transition system as much as possible, which has a worst-case time complexity of $O(n^2)$ for his transition system. Although the additional re-ordering after each non-projective sentence adds to this time complexity, the additional time is linear: rather than an additional `Shift` for every `Swap`, there are an additional three `Shifts` and one `Swap`.

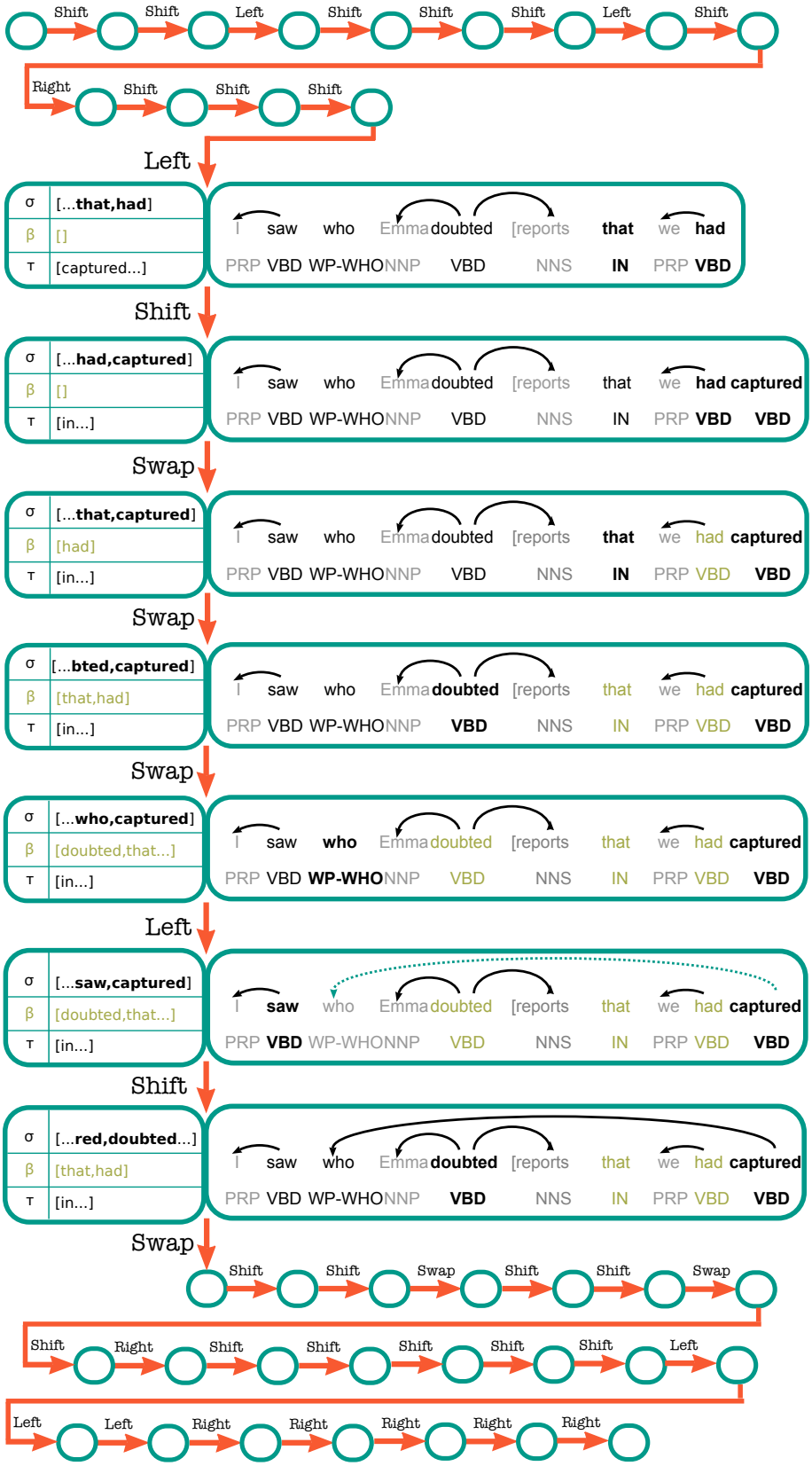


Figure 4.4: An example parse of a CNP experimental sentence.

4.4 The oracle

But how does the parser know which actions to take to create the correct dependency analysis? Nivre (2009) calls this function the **oracle**: given a parser state, it provides a transition to a subsequent parser state. Developing the oracle forms a large part of the cognitive model, as can be seen in the parsing diagram provided at the beginning of the chapter, and repeated in Figure 4.5. This section explains steps (1) through (9) in the diagram. Section 4.4.1 and Section 4.4.2 describes the treebanks used to define a probabilistic grammar, and transformations to the treebank that are required to provide sensitivity to syntactic locality phenomena. Section 4.4.3 describes how these treebanks are then converted into state-action banks, providing a sequence of parser states and actions for each sentence in the treebanks. Section 4.4.4 and Section 4.4.5 then discuss how the state-action banks are converted into probabilistic features that are used in the parser.

4.4.1 Treebanks

The syntactic locality experimental data is in four languages: English, German, Swedish, and Russian. The model therefore requires dependency treebanks in each of these languages to inform parser decisions. These treebanks are listed in Table 4.2.

Name	Language	Sentences	Text type	Format	Projectivity
Brown	English	19,395	Balanced	CFG	Projective
Negra	German	20,602	Newspaper	CFG	Non-projective
Talbanken05	Swedish	8,834	Newspaper	Dependency	Non-projective
SynTagRus	Russian	32,950	Balanced	Dependency	Non-projective

Table 4.2: These treebanks inform parser probabilities.

In two of the languages, Swedish and Russian, dependency treebanks already exist. For Swedish, a portion of the Swedish Talbanken 2005 (Nivre, Nilsson, & Hall, 2006) which had

been prepared for a CoNLL (Computational Natural Language Learning Conference) shared task is used. This data was already in the useful CoNLL format, which is used as a base format for all dependency treebanks for the parser. Table 4.3 shows an example of this format. The crucial information for this task is the word, its position, its part-of-speech (**POS**), and the position of its head. The Russian SynTagRus corpus (Nivre, Boguslavsky, & Iomdin, 2008) is also a dependency corpus, and it had been formatted for a CoNLL shared-task as well.

ID	Form	Lemma	Coarse POS	POS	Features	Head	Dependency Relation	Projective Head	Label
1	In	_	IN	_	_	7	ADV	_	_
2	American	_	JJ	_	_	3	NMOD	_	_
3	romance	_	NN	_	_	1	PMOD	_	_
4	,	_	,	_	_	7	P	_	_
5	almost	_	RB	_	_	6	NMOD	_	_
6	nothing	_	NN	_	_	7	SBJ	_	_
7	rates	_	VBZ	_	_	0	ROOT	_	_
8	higher	_	JJR	_	_	7	OPRD	_	_

Table 4.3: Fragment of a Brown sentence in CoNLL format.

The Brown (Francis & Kucera, 1979) and Negra treebanks (Brants et al., 2004) inform English and German transitions. Both are large-scale treebanks often used in broad-coverage parsing. Whereas the Brown corpus is balanced, containing sentences from a variety of media, including newspapers, novels, and magazines on a variety of topics, the Negra treebank contains only newspaper text. It was chosen because it is the most freely-available large-scale German treebank.

Both the Brown and Negra treebanks are converted from CFG format to DG format with headfinder tools. Headfinder tools provide a list of ”main children” for each node in a CFG tree. The lists are ordered so that the main child will be the first in the list to be found within that node. (28) shows an example from Dubey’s headfinder tool for Negra (2004). The head of an S node (and hence the root of a sentence) will be the first verb found, with finite verbs (VVFIN) preferred. If no verbs are available before a verb phrase is encountered (VP), then the head will be the head of that verb phrase, and so on.

(28) “S”→[“VVFIN”; “VMFIN”; “VAFIN”; “VVIMP”; “VAIMP”; “VMPP”; “VVPP”;“VP”; “CVP”; “S”; “CS”; “VVINF”; “VAIMP”; “NP”; “PP”];

This work uses Dubey’s headfinder rules in a conversion tool that creates dependencies for Negra. For Brown, the freely-available pennconverter (Johansson & Nugues, 2007) is used.

4.4.2 ① and ②: Treebank transformations

There are two main differences between the experimental data and the treebanks. First, the experimental data is non-projective, whereas some of the treebanks, like Brown, are projective. Secondly, the experimental data includes island violations, whereas the treebanks themselves have very few island violations[2]. The Swedish Talbanken includes eight WHI violations, which is the most of any treebank. Although there are more examples of CNP violations across languages (Swedish has 30, Russian has 8, English has 3, and German has 1), they are too infrequent to effect probabilities. Some amount of the probability mass, though, must be allotted to island-violating and non-projective contexts.

Therefore, the data is transformed to create island violations in the treebanks, shown in (1) in the diagram. The first step is to find examples of relative clauses for CNPs and complement clauses headed by *whether* for WHIs in the treebanks. The second step is to search for a preceding wh-word in the sentence and change that wh-word’s dependency to the embedded verb within the relative/complement clause. This not only creates an island-violating context, but it also often creates non-projective sentences, necessary for Brown. These new sentences provide probabilities that allow the parser to consider island violations.

Many of the experiments for syntactic locality differentiate the specificity of the wh-word.

[2]SUVs are difficult to find because the parser is unlexicalized and unlabeled: for the parser, any two wh-word dependents of a later verb forms an SUV.

For example, in English, *what* and *which* share a POS tag, WDT. But, experimental evidence suggests that *what* and *which-noun* have different acceptability in an SUV context (Arnon et al., To Appear; Fedorenko & Gibson, Submitted). To differentiate the words in the POS-based parser, the POS tags for all wh-words were changed in each language, as Table 4.4 shows.

The island-violation and POS transformations provide the necessary data to make the parser sensitive to both islands and wh-word specificity. The result of these two transformations on the four raw treebanks are the base-level dependency treebanks in ②.

4.4.3 ③ and ④: From treebanks to state-action banks

Unlike grammar-based parsers, transition-based parsers do not base probabilities on the grammar itself. Rather, they base probabilities on state-action pairs. Therefore, each of the treebanks are converted to sequences of state and action pairs for each sentence. ConllToConfigMaker implements this converter, which takes in treebanks in CoNLL format and provides treebanks in Nivre configuration (or state) format. For each sentence, the sequence of parser states necessary to build the correct dependency analysis is listed, as well as the correct actions to take to get to the next parser state. This treebank of state-action pairs is then used to train probabilistic features, discussed in the next section.

4.4.4 ⑤ and ⑥: From state-action banks to feature banks

The state-and-action banks contain all the information in each parser state: all the words in the stack, all the words in the buffer, all the dependency analyses created. One problem is that to generalize from the treebank data to the experimental data at test time, this is too much information; it is difficult to find within the treebank data the exact parser state encountered

Original POS	New POS	Affected words
	Brown	
WP	WP-WHO	who, whom
WDT	WDT-WHICH	which
WDT	WDT-WHAT	what
WP	WP-WHAT	what
WRB	WRB-WHEN	when
WRB	WRB-WHERE	where
WRB	WRB-WHY	why
WRB	WRB-HOW	how
IN	IN-WHETHER	whether
	Negra	
PWS	PWS-WER	wer, wen (who)
PWAT	PWAT-WELCHE	welche (which)
PWS	PWS-WAS	was (what)
PWAV	PWAV-WANN	wann (when)
PWAV	PWAV-WO	wo (where)
PWAV	PWAV-WARUM	warum (why)
PWAT	PWAT-WIEVIEL	wieviel, wievielen (how)
PWAV	PWAV-WIE	wie (how)
PWAV	PWAV-WIEVIEL	wieviel (how)
KOUS	KOUS-OB	ob (whether)
	Swedish	
PO	PO-VEM	vem, vems (who)
PO	PO-VILKEN	vilken, vilket, vilka (which)
PO	PO-VAD	vad (what)
AB	AB-NAR	när (when)
AB	AB-DAR	där (where)
PO	PO-VAR	var (where)
AB	AB-VARFOR	varför (why)
AB	AB-HUR	hur (how)
UK	UK-OM	om (whether)
	Russian	
S	S-WHO	кто, кого, кому, кем, ком (who)
A	A-WHICH	какой, какая, какое (which)
S	S-WHAT	что, чего, чему, чем, чём (what)
CONJ	CONJ-WHEN	когда (when)
ADV	ADV-WHERE	где, куда, откуда (where)
ADV	ADV-WHY	почему (why)
ADV	ADV-HOW	как (how)
PART	PART-WHETHER	ли (whether)

Table 4.4: Question word parts-of-speech (POS) were changed to make the parser sensitive to gradience in wh-words.

during parsing. This is particularly true considering difficult sentences like those that demonstrate syntactic locality violations. If the exact parser state is not found, how can the parser decide what action is best based on its treebank experience?

The answer is in using only specific **features** of the parser state to inform parser decisions. Each feature bases parser decisions on a select amount of information from the parser state. One simple probabilistic feature is Distance: it considers how far apart two words are in a sentence, and bases decisions on that information. For the Nivre parser, the two words are σ_1 and σ_2. They can be arbitrarily far apart in the sentence; after all, the `Right-Arc`, `Left-Arc`, and `Swap` transitions manipulate σ so that the two words at the top of the stack need not be neighbors within the sentence. At runtime, the parser can query the Distance feature to see which parser action is best to take when σ_1 and σ_2 are one word apart, or two words apart, and so on.

This information must be translated from the state-action bank format into a series of feature banks to be usable at runtime. The feature bank for the Distance feature is created by printing out, for each parser state, the transition that was taken and the distance between σ_1 and σ_2. This is done for each of the probabilistic features considered here, further detailed in Section 4.5. The next section discusses how these feature banks are turned into probabilities that inform parser decisions.

4.4.5 ⑦, ⑧, and ⑨: From feature banks to probabilistic features

The feature banks consist of transitions and feature instances, as shown in the Feature Corpus in the diagram in Figure 4.5. But, the parser requires probabilities for taking each action shown in the Feature Probabilities. Machine learning provides a method for learning the patterns in the feature banks, which result in weights for each transition and feature instance.

The machine learning implementation used here is liblinear 1.5 (Lin, Weng, & Keerthi, 2008), a tool that uses support-vector machines (SVMs). Unlike other SVM implementations, this tool provides fast learning over the large amounts of treebank data required for this research. And, because it is based on SVMs, it provides high accuracy for transition-based dependency parsers, and the Nivre system in particular (Yamada & Matsumoto, 2003; Attardi, 2006; Nivre, 2009). Liblinear's output is a list of feature instances, along with weights for each parser action. These weights are then normalized: each weight is divided by the sum of the four action weights, resulting in probabilities for each parser action[3].

The feature probabilities following (8) in the diagram show the predictions for DISTANCE. When σ_1 and σ_2 are one word apart, or neighbors, all four actions are roughly equally probable. But when they are six words apart, the most probable action is `Shift`. DISTANCE is a simple feature that has the advantage of being generalizable: the parser will likely not encounter a state in the experimental data that has not been encountered in the treebank data. One downfall of a generalizable feature is that it is less informative: the probabilities are likely to be even or not very helpful in many cases, and accuracy is usually low. The next section discusses how the cognitive theories are implemented as informative probabilistic features for the syntactic locality data.

[3] I normalize weights at run-time within the parser, but for ease of description the normalized weights are shown in the Feature Probabilities list that is a result of (8).

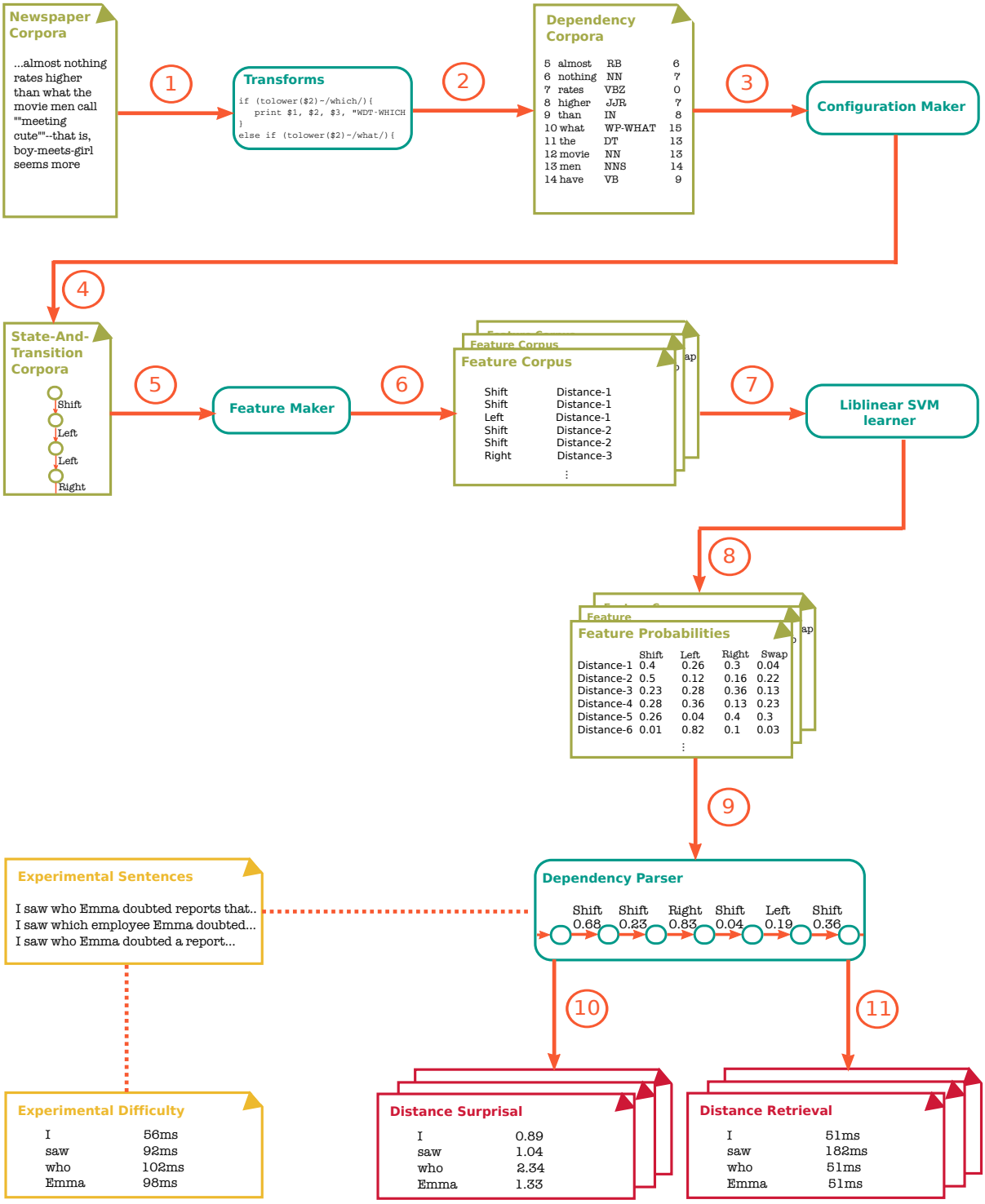


Figure 4.5: Parsing non-projective dependencies.

4.5 Cognitive theories as probabilistic features

There are a variety of ways that cognitive theories can be implemented in a human sentence processsing model. First, they could simply be implemented as metrics in their own right. For example, retrieval (Lewis & Vasishth, 2005) is a metric that provides a measure of difficulty in terms of milliseconds. It can be directly compared to data. The DLT (Gibson, 2000) can also be directly compared to data since it provides a numerical reading of difficulty in terms of open discourse referents. In fact, one could convert each of the memory theories considerd in Chapter 3 directly into metrics to compare to data.

One of the reasons the theories are not encoded as metrics in the parser is because it would make comparison across metrics difficult. Because the metrics are different, it's not clear that a direct comparison would be informative. A direct comparison would also lack the benefit of the broad-coverage, working cognitive model. Researchers have argued for some time that human sentence processing models should move away from small models that work on relatively few sentences to broad-coverage models (Crocker & Brants, 2000; Crocker, 2005; Keller, 2010). Encoding the cognitive theories as metrics would not ensure the benefit of this approach.

The main reason the cognitive theories are encoded as features rather than metrics is because features can help *inform* parser decisions, rather than simply be a result of them. Probabilistic features based on cognitive theories, particularly those focused on memory limitations in the human processor, provide a more adequate human sentence processing model not only for syntactic locality, but for all phenomena. The dependency parser used in previous work had an unlimited memory: words from arbitrarily far back in the sentence could be accessed, no matter how much time (or parse steps) had passed. An accurate cognitive model requires memory limitations; these can be achieved by changing the architecture, but they can also be achieved by using probabilistic features. Although the metrics may provide a direct

implementation of the cognitive theories, implementing them as probabilistic features provides a model of memory limitations informed by human-based theories.

The theories could have also been encoded as hard constraints in the parser. For example, one could construct a hard DLT constraint that prohibits attachments once the DLT value is higher than 4. But, this seems to go against the variable nature of cognitive constraints: violations of cognitive constraints depend on how much memory is available. If the memory burden on the rest of the sentence, for example from lexical information, is relatively low, then perhaps attachments can be made at a higher DLT. Further, it's not exactly clear where to draw the line between acceptable and unacceptable values for any of these metrics. This would require extensive experimental and statistical analyses that, though interesting, fall beyond the scope of this work.

The cognitive theories are therefore encoded as soft constraints, which in this case are probabilistic features. Soft constraints have several advantages over hard constraints: they are naturally violable and cumulative, which could help in addressing gradience patterns in the syntactic locality data. They are sensitive to probabilities derived from corpus data, and can therefore take into account cross-linguistic variation. And they can be directly compared: not only can the cognitive theories be rated based on their accuracy in determining dependency analyses for syntactic sentences, but surprisal (Hale, 2001) can be used to determine how much difficulty the cognitive theories allot to different sentences.

It is this last factor that makes implementing cognitive theories as probabilistic features most appealing for better understanding syntactic locality: this directly compares the cognitive theories on the exact same broad-coverage parsing mechanism. The only difference is what aspects of memory the parser is sensitive to when deciding which action to choose. It could be sensitive to activation-based information, such as how far apart two words are (DISTANCE), or interference-based information, such as whether any similar words occur between the two words considered for attachment (INTERFERERS). It could even take into account both factors and

Feature	Feature Type	Includes
DISTANCE	String Position	$\sigma_1 - \sigma_2$
ACTIVATION	Value	baselineActivation(σ_2)
FILLER	POS	$\sigma_{wh-word}$
INTERVENORS	POS	nominalIntervenors($\sigma_1 ... \sigma_2$)
INTERFERERS	POS	interferers(σ_2)
SBI	Value	SBI(σ_2)
DLT	Count	nominalIntervenors($\sigma_1 ... \sigma_2$)
RETRIEVAL	Time (ms.)	retrievalTime(σ_2)

Table 4.5: Feature specification. :: indicates concatenation.

decide which parser action is best by how long it would take to retrieve the word from memory (PROBRETRIEVAL).

The rest of this section details how to define probabilistic features based on each of the cognitive hypotheses considered in Chapter 3. The feature definitions are provided in Table 4.5. Section 4.5.1 describes the two activation-based features, DISTANCE and DECAY. Section 4.5.2 discusses the interference-based features, FILLER, INTERVENORS, INTERFERERS, and SBI. Section 4.5.3 discusses the two theories that take into account both quantities, DLT and PROBRETRIEVAL. More details on the cognitive theories that inform these features, as well as why they were selected, is available in Chapter 3.

4.5.1 Activation

Distance

The DISTANCE feature captures the intuition that the distance between two words can affect the processing difficulty in integrating them. The parser is made sensitive to distance by basing the likelihood of parser actions on how far apart two words are, as in (29).

(29) DISTANCE: String position of $|\sigma_1 - \sigma_1|$

The distance between *who* and *captured* for the CNP sentence in Figure 4.6 is 7. The parser would determine whether to create this attachment by considering the probability of attaching words that are 7 words apart in the training data. If the training data has more short-distance dependencies, it may advise the parser not to make this attachment.

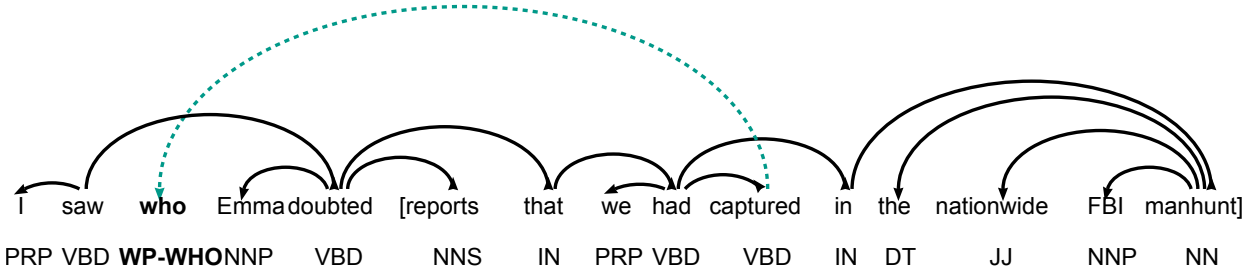


Figure 4.6: How likely is an attachment between *who* and *captured*?

Decay

The DECAY feature is one of the addends in the Lewis and Vasishth (2005) retrieval equation. It represents the baseline activation for a word that is to be retrieved, and is meant to capture the memory decay of words that have already been heard in the sentence. Words that are further back, and which have not been re-activated with recent attachments, have lower baseline activations; within the retrieval equation, this causes an increase in retrieval time since it takes longer to retrieve a word that has been subject to memory decay. Details on the baseline activation and retrieval equations, as well as how they were translated into the parser, are available in Section 4.6.2. For the purposes here, the probabilistic feature for activation consists simply of the quantity returned by the baseline activation equation for retrieval, as in (30). Baseline activation quantities were rounded to the hundredths position. The parser calculates the DECAY of *who* for the CNP example above as -3.37.

(30) DECAY: Value of baselineActivation(σ_2)

4.5.2 Interference

Filler

The FILLER feature is sensitive to the POS of the first-encountered wh-word that is still unresolved, defined as in (31). It is considered an interference feature because unresolved fillers are a memory burden that can interfere with dependencies between unrelated words (Charles Clifton & Frazier, 1986). For all parser actions that occur between when *who* is first encountered and when it is finally attached to *captured*, the value of FILLER is WP-WHO.

(31) FILLER: POS of $\sigma_{wh-word}$

Intervenors

The INTERVENORS feature is sensitve to the POS of any intervenors, or discourse referents, that occur between two words, as in (32). It can be considered a more-informative version of the DLT because it considers the type, rather than just the number, of intervening discourse referents. Even though it is considered a more informative version of DLT, it is classified as an interference feature because it doesn't explicitly take into account distance or activation: it simply lists the words that interfere with processing, according to the DLT. For the CNP example, the value of the feature is PRP|NNS|NNP, for each of the nominal intervenors that have occurred between the two words, beginning with the most recent, *we*.

(32) INTERVENORS: POS of intervenors($\sigma_1 ... \sigma_2$)

Interferers

Like, INTERVENORS, INTERFERERS takes into account the POS of certain words that occur between σ_1 and σ_2. Rather than considering nominal intervenors, this feature focuses on the words that SBI (Lewis & Vasishth, 2005) considers to interfere with the retrieval of σ_2. As with DECAY, the full specification of SBI is included in Section 4.6.2. But for the interesting cases in syntactic locality, SBI will consider any wh-word from the list in Table 4.4 as interfering with the extracted wh-word being retrieved for syntactic locality. For the CNP example, though, there is no wh-word between *who* and *captured*. In the case where these two words are at the top of the stack, INTERFERERS would be NULL.

(33) INTERFERERS: POS of interferers(σ_2)

SBI

The SBI feature is the numerical quantity that the Lewis and Vasishth (2005) retrieval equation assigns for SBI. It takes into account how many words interfere with the retrieved word and returns a value. The SBI value operates such that retrieved words that have more interference from other words in the sentence have less activation and therefore take more time to retrieve. The reader is referred to Section 4.6.2 for more details on how SBI is implemented in the parser. The feature is defined as in (34), and the values are rounded to the hundredths decimal point. Its value for the CNP example is 1.50.

(34) SBI: Value of SBI(σ_2)

4.5.3 Composite

DLT

The DLT feature translates the intuitions from the DLT theory into a probabilistic feature. The DLT captures the intuition that longer dependencies are problematic in sentence processing. However, long dependencies are not all equally difficult: it depends on what kinds of words intervene. For the DLT, open discourse referents, like nominals, increase the difficulty of long-distance dependencies because they need to be held in memory until their head verb is encountered. (35) shows the feature definition: it returns a count of all nominal intervenors between σ_1 and σ_2. For the CNP example, this would be 3.

(35) DLT: Count of nominalIntervenors(σ_1...σ_2)

ProbRetrieval

The most complicated feature is derived from the most complex and encompassing of the memory theories considered. ProbRetrieval (probabilistic retrieval) bases sentence processing difficulty on how difficult a word is to retrieve from memory. Although all of the memory theories considered aim to quantify this idea, retrieval as defined by Lewis and Vasishth (2005) arguably encompasses them all. It bases retrieval difficulty on two explicit quantities, activation and interference, which are calculated from principles of an independent general cognitive framework. Full details on the retrieval theory and its calculation in the parser are available in Chapter 3.4.3 and Section 4.6.2. To turn the theory into a probabilistic feature, the retrieval time of σ_2 is estimated and rounded to the nearest whole number. For the CNP example, retrieval time for *who* is estimated as 91ms, which is the value of the feature.

(36) ProbRetrieval: Value of retrieval(σ_2)

Each of these feature definitions are used in the FeatureMaker tool diagrammed in steps (7) and (8) and described in Section 4.4.4. Once these feature treebanks are submitted to the SVM learner, the output is a set of probabilities for each of the parser actions, based on the feature's instance value.

4.6 Complexity metrics

The point of the working cognitive model is that it should be able to predict when humans will find certain sentences difficult. But, in order to do this, there must be a way to measure parser difficulty. There are two complexity metrics considered in this cognitive model to predict human difficulty. The first, surprisal (Hale, 2001), considers changes in the probability space as a sentence is parsed. The advantage of this metric is that it can directly compare the difficulty that probabilistic features are sensitive to, and determine which cognitive theories perform best as probabilistic features. The second, retrieval (Lewis & Vasishth, 2005), is the direct encoding of the retrieval complexity metric on this parser. Retrieval is the most explicit of the cognitive theories considered, and encoding it as its own metric allows for a direct measure of the reductionist hypothesis. This section discusses how both are implemented, beginning with surprisal in Section 4.6.1.

4.6.1 Surprisal

Surprisal offers a complexity metric for broad-coverage human sentence processing models. It not only eliminates the need for hand-coding small models, but it also incorporates frequency and grammatical information that is known to affect sentence processing. It is able to do this by being sensitive to the probabilistic space that the parser explores as it is parsing a sentence– that is the probabilistic space explored by any probabilistic model, operating on any grammar,

informed by any parsing strategy. Previous work in this paradigm demonstrates surprisal's adaptability by introducing a definition of surprisal appropriate for transition systems rather than grammars (Boston et al., 2011). This section discusses this implementation.

Surprisal is based on how the **prefix-probability** changes from one word to the next during a parse. A word's prefix-probability is calculated as the sum of all transitional probabilities t considered during a parse, as in Equation 4.1.

$$\text{prefix-probability}(w, G) \quad = \sum_{t \in NIVRE(\mathcal{D}(G, wv))} \text{Prob}(t) \tag{4.1}$$

This work focuses on a serial version of the parser; therefore, the prefix-probability considers only one t. This transitional probability t must be in the set of allowable Nivre derivations between grammars and strings, $NIVRE(\mathcal{D})$. The derivations must also be complete: for any prefix w, there must be a suffix v that would result in a grammatical sentence of the grammar G. The transitional probability t is calculated by multiplying all the transitional probabilities, or action probabilities, that have led to the word being parsed. This is illustrated by the diagram in Figure 4.7.

This diagram shows a parse of the running example sentence of a CNP violation. The snippet shows the transitions from when the relative clause is introduced by *that* until the embedded verb (and head of the extracted wh-word, *who*) is parsed. The y axis shows probabilities and the x axis the words in the sentence. The blue boxes are states, and listed within the state are the transitions that led to the state as well as the transitional probability to that point. The probabilities are actual probabilities from the Distance feature discussed in Section 4.5.1. As can be seen in the diagram, each state's prefix-probability is calculated by multiplying the transition probability, such as 0.314 for `Shifting` *we*, by the previous word's prefix-probability (2.965e-06 for *that*). Because this is a serial parser, only one prefix-probability is considered for the surprisal calculation, and this is the prefix-probability before the next word is shifted onto the stack. Word boundaries are delineated by the gray vertical lines between words.

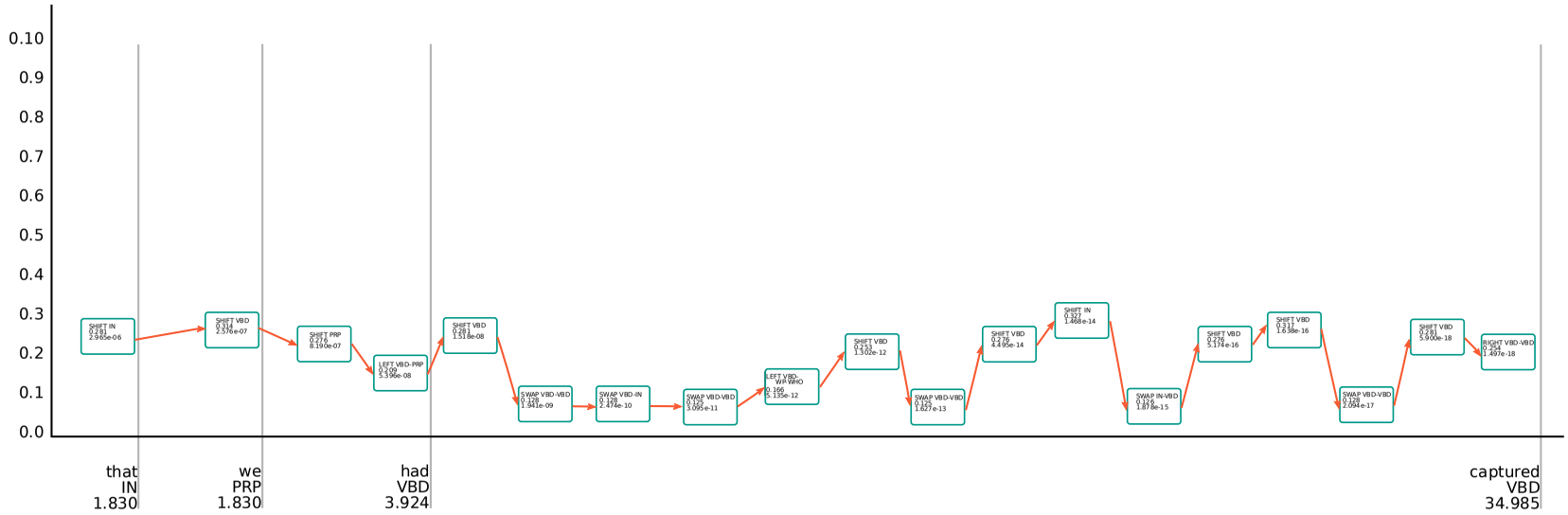


Figure 4.7: Surprisal calculation.

Surprisal is then calculated as the change in prefix-probability between words, formalized as in Equation 4.2. It is the negative log of the current word's prefix-probability divided by the previous word's prefix-probability; in other words, the change in the probability space once the current word has been parsed. This equation captures the intuition that transitions that lead to "surprising" words or analyses have lower probabilities; these lower probabilities indicate an increase in surprisal, or difficulty.

$$\text{surprisal}(n) = -\log_2 \left(\frac{\text{prefix-probability}(w_n, G)}{\text{prefix-probability}(w_{n-1}, G)} \right) \tag{4.2}$$

In the diagram in Figure 4.7, surprisals at *that*, *we*, and *had* are relatively low. In fact, *that* and *had* appear to have the same surprisal, although that is simply an effect of the rounding done for the figure; the transition probability does not affect the prefix-probability very much, indicating that encountering a personal pronoun (PRP) after a complementizer (IN) is not surprising for the Distance feature. On the other hand, surprisal is higher once the verb *captured* is encountered. This is a result of two things: first, it is because the process of attaching a non-projective dependency from *captured* to *who* requires many states and transitions. For each transition, the prefix-probability is reduced by multiplying probabilities, leading to a lower final prefix-probability for *captured* (1.497e-18) than for *had* (5.396e-08). Further, many of the probabilities for these transitions are low: the `Swap` transitions in particular are not predicted by the Distance feature.

The surprisal value for *captured* is higher than for other words, indicating that creating the CNP island-violating dependency is difficult. This is a good thing for the cognitive model. But, there is a potential problem. Non-projective sentences are more difficult than projective sentences because of the large number of actions required to make them. But, non-projective sentences are not always more difficult than projective sentences (Levy, Fedorenko, Breen, & Gibson, Submitted), and this parsing model would not be able to correctly model this result. On

the other hand, because wrap-up effects generally lead to higher processing difficulty (Just & Carpenter, 1980; Rayner, Kambe, & Duffy, 2000), this is considered a feature rather than a bug of the system. Further, all of the non-projective sentences considered here are in strong island and highly unacceptable contexts for human speakers, so the parser should report difficulty when modeling them.

Despite this potential problem, there is one important benefit of using surprisal for this study. It allows a standard of comparison across probabilistic features and hence cognitive theories. Consider a situation where the parser is attempting to decide whether to attach σ_1 and σ_2. The words happen to be two words apart: there is one intervening word. Because the words are only two words apart, the activation features like Distance and Decay would consider a `Left-Arc` and `Right-Arc` transition likely: these features are highly local and attachment probabilities decrease with length. But what if the word that intervenes interferes with σ_2? For activation features, this would make no difference. If the words are only two words apart, σ_2's activation is high, attachment probabilities are high, and surprisal will be low.

Interference features, on the other hand, would be sensitive to this intervenor's interference, and will have lower attachment probabilities and higher surprisal when compared with a case where the intervening word does not interfere. If humans also have more difficulty with the interfering case, then one could say that the better model is the interference model rather than the activation model. This allows the model to address the question of which cognitive theory best explains syntactic locality: activation, interference, or both.

4.6.2 Retrieval

Retrieval, as developed by Lewis and Vasishth (2005), bases sentence comprehension difficulty on the need to retrieve structures from working memory. Experimental evidence suggests

that there is a cost associated with this retrieval process (Caplan & Waters, 1999; Gibson, 1998; Gordon et al., 2004; Just & Carpenter, 1992; Just & Varma, 2002; Lewis, 1999; Warren & Gibson, 2002; Grodner & Gibson, 2005), and retrieval explicitly links this difficulty to working memory limitations using the general cognitive framework Adaptive Control of Thought-Rational (ACT-R) (Anderson & Lebiere, 1998; Anderson, 2005). The central insight of this approach is that strains on working memory translate to sentence processing difficulty.

The parsing system implemented in Lewis and Vasishth (2005) uses condition-action pairs informed by a phrase structure grammar to drive parsing. The system also uses a series of memory buffers to represent long-term and short-term storage of elements. The architecture makes use of parallel, associative retrieval (McElree et al., 2003), activation fluctuations of elements already in memory, and similarity-based retrieval interference. Together, they determine the amount of time it takes to process each word, and these time predictions provide a metric of sentence processing difficulty.

Abstracting from the implementation-particular aspects of the Lewis and Vasishth (2005) formulation, the constraint on working memory that most directly translates to sentence processing difficulty is the amount of time it takes to retrieve a word (Vasishth, Brüssow, Lewis, & Drenhaus, 2008). This is accomplished in the dependency parser by translating the ACT-R formulation of retrieval to the Nivre transition system: a retrieval occurs whenever the parser draws a dependency arc. The retrieval time of the word that is to be attached (or attached to) is based on that word's *activation*, which is calculated as in Equation 4.3 (Lewis & Vasishth, 2005).

$$\mathrm{A}_i = B_i + \sum_j W_j S_{ji} \tag{4.3}$$

Activation is based on two separate quantities. One is the word's baseline activation B_i, which calculates activation decay due solely to the passage of time, as in Equation 4.4. In

particular, baseline activation for word$_i$ is the summation over the time since the j^{th} retrieval of word$_i$, t_j. The parameter d is set to 0.5, as in most ACT-R models (Lewis & Vasishth, 2005).

$$B_i = \ln\left(\sum_{j=1}^{n} t_j^{-d}\right) \tag{4.4}$$

The second variable that is used in determining a word's activation is the amount of SBI that occurs with other words that have been parsed, given by the second addend in Equation 4.3. SBI is estimated by the weighted strengths of association between the word to be retrieved and other words already parsed, depicted in Equation 4.5.

$$S_{ji} = S - \ln(fan_j) \tag{4.5}$$

In Equation 4.5, word j is a word similar to word i. In this formulation, similarity is determined by part-of-speech classes, with, for example, nouns being able to interfere with other nouns, but not with verbs. In terms of syntactic locality, the only interference that is of interest is between wh-words. Therefore, all wh-word POS listed in Table 4.4 are considered "similar". If word j is similar, the amount it interferes with word i is determined by fan_j, or the number of words already associated with j. S equals the maximum associative strength of 1.5 (Lewis & Vasishth, 2005).

This interference variable has a weight, W_j, associated with the number of elements in the goal chunk. It is formalized as G/j by Lewis and Vasishth (2005), where G is a constant set to 1.0 and j is the number of cues in the goal chunk. Because the parser is unlexicalized, there will only be one cue for each retrieved item, the POS. Therefore, W_j is simplified to equal 1.0 for all retrievals.

The addition of a word's baseline activation plus any SBI provides the word's activation, A_i. This activation is used to calculate the sentence processing metric used in this study, the time to retrieve the word. The equation is given in 4.6, where F is estimated at 0.14, following the

Lewis and Vasishth (2005) implementation, and *e* is Euler's constant.

$$\mathsf{T}_i = Fe^{-A_i} \tag{4.6}$$

The ACT-R formulation of retrieval also incorporates fixed "production" rules of 50 ms. for taking actions, where productions directly translate to actions programmed in the ACT-R planner. In this formulation, productions are registered for taking a `Shift`, `Left-Arc`, and `Right-Arc` action, but not for a `Swap` action because this does not explicitly change the dependency analysis. Productions are also registered for deciding to retrieve items, as in the ACT-R formulation. The time it takes to integrate a word into the analysis is therefore determined on the basis of both productions and retrievals, as specified in Table 4.6. As in the ACT-R formulation of retrieval, the amount of time it takes to read a word is set to 1ms.

`Left-Arc`	50ms. + 50ms. + Retrieval Time.
`Right-Arc`	50ms. + 50ms. + Retrieval Time.
`Shift`	50ms.
`Swap`	0ms.

Table 4.6: How time is determined in the parser.

Figure 4.8 diagrams how the retrieval calculation works on the running CNP example. Notice that the state and transitions are exactly the same as the surprisal calculation in Figure 4.7. The main difference is that now the y axis is the amount of time in milliseconds, and the x axis shows the words and their parse times based on retrieval. It would seem that surprisal and retrieval make complementary predictions about sentence processing difficulty, but as can be seen by comparing the two calculation diagrams, this is not exactly correct in this parsing model. If it were so, the diagrams should appear to be the inverses of each other, with actions that require high parse times for retrieval being low-probability for surprisal. Although there are a few places where this almost appears to be so, it's not the case; this confirms the intuitions of Boston et al. (2011) that there is some overlap in the difficulty attributions of the two quantities. This is also likely because the surprisal calculations are based on DISTANCE, a probabilistic fea-

ture that, though often used for standard dependency parsing (Eisner & Smith, 2005, 2010), is also the basis for a strong cognitive theory.

It should be noted that aside from the translation from ACT-R productions and retrievals to the Nivre dependency system, the calculation of retrievals follows directly from the ACT-R principles in the Lewis and Vasishth (2005) implementation. Also, all numerical constants in Equations 4.3-4.6 are kept at the default values in ACT-R. These equations are constant not only for the retrieval metric, but also for the DECAY, SBI, and PROBRETRIEVAL probabilistic features discussed in Section 4.5.

Finally, consider the differences in the two implementations of retrieval used in this work. The discussion in this section has centered on the retrieval metric, which most directly encodes the retrieval discussed in Lewis and Vasishth (2005). It provides a parsing time estimate for each word in a sentence based on the memory burden required to build analyses. But, retrieval is also a probabilistic feature, PROBRETRIEVAL. The probabilistic feature also seeks to encode the intuition behind retrieval, but less directly. In this case, the amount of retrieval time affects the likelihood of attaching words in a sentence: in a way, it drives parsing rather than being the outcome of a parse. Although the two quantities should behave similarly, it is likely they will make conflicting predictions in some cases. This is because the probabilistic PROBRETRIEVAL is sensitive to the treebank information. For example, there may be an attachment that the retrieval theory would consider particularly difficult, such as with a word that is far away. But, probabilistic PROBRETRIEVAL may have learned, from all the treebank data, that German, for example, tends to like to perform `Right-Arc` attachments when words are far away, reflecting verb-secondhood in German. In this case, the probabilistic PROBRETRIEVAL would give a low surprisal rating, whereas the retrieval metric, RETRIEVAL, would be relatively high. Whether this benefit plays out in the syntactic locality data, or even other psycholinguistically interesting cases, is a question for future research.

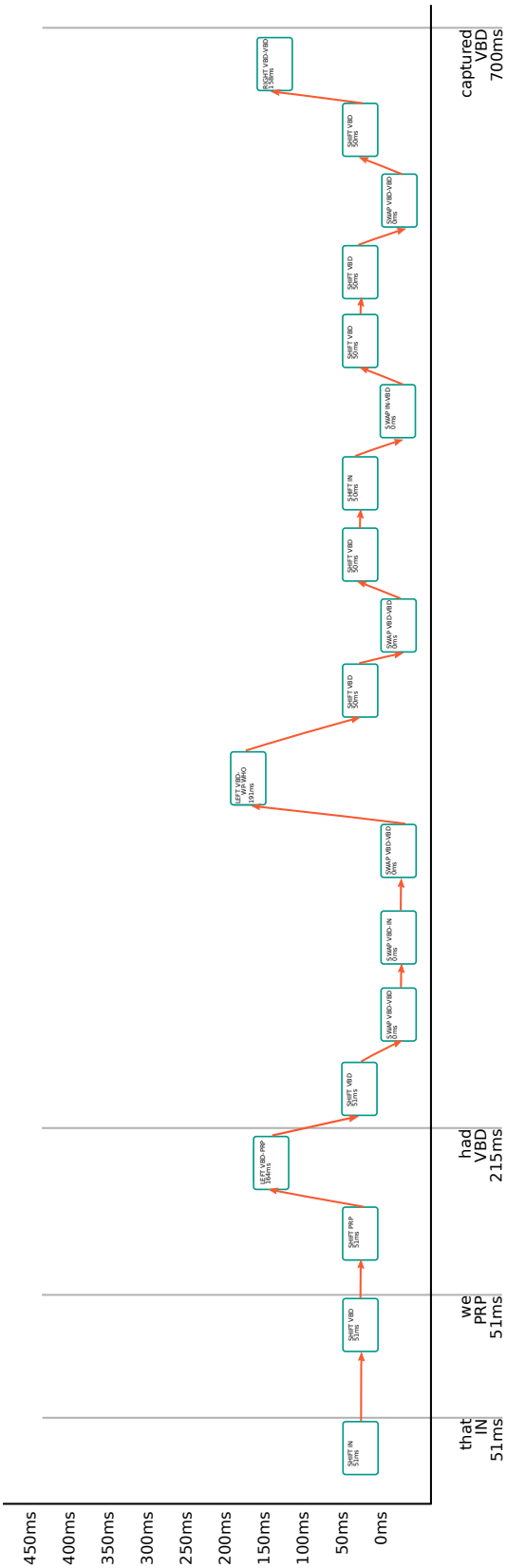


Figure 4.8: Retrieval calculation.

4.7 Dependency length and difficulty

This parser uses metrics that are sensitive to the number of parser actions required to parse a word. For surprisal, each parser action decreases the transition probability. If there are many parser actions between words, the prefix probability for that word will be lower than the previous word, and the surprisal value will be high. Likewise, each parser action has a fixed millisecond cost for retrieval; the more parser actions required to retrieve a word, the higher that word's retrieval time.

In the Nivre transition system, dependencies can only be constructed between words that are adjacent in the stack memory. The farther apart two words are, the more likely that other words intervene in the stack, and the more likely that many parser actions are required to bring the two words close to each other. Figure 4.7 and Figure 4.8 demonstrate this issue for the running CNP example. Notice that because the verb *captured* and its dependent *who* are far apart, many `Swap` transitions are required to get them next to each other, and then many *Swap-Shift-Shift* sequences are required to reorder the sentence. These transitions bring down the prefix-probability for surprisal, and drive up the production costs for retrieval, leading to high parser difficulty.

This tendency for longer dependencies to be harder in the parser works well for the cognitive model; generally, long-distance dependencies lead to more human difficulty. However, it is not the case that long-distance dependencies are always harder than shorter-distance dependencies, particularly if those shorter dependencies include nesting or other known memory burdens. When this is the case, the parser is not likely to be a good model of difficulty.

Although there is a tendency for the parser to have high difficulty with longer dependencies, this tendency can be over-ridden. It may be the case that the individual transition probabilities are high enough that they offset the probability decrease associated with multiplying probabilities. It may also be the case that the calculated activation for a word is high enough (because

it has been reactivated in memory) that its retrieval time would offset the high production costs for retrieving it. In the small number of cases where the experimental sentences have a dependency length mismatch, the cognitive constraints may be able to over-ride this tendency and correctly model the data.

These issues further support the use of associative memory for human models, as well as more human-like memory constraints. For this particular task, this issue is only problematic for a subset of the SUV data, as is discussed in Chapter 7.

4.8 Accuracy

One of the difficulties in developing a human cognitive model is that parser accuracy can be difficult to determine. The development set is often the same as the test set, which can lead to data overfitting, and sometimes even questionable cognitive models (Keller, 2010). The focus of this research is to discover how well cognitive features model syntactic locality; it is not to develop an accurate model of the phenomena, but rather to test how well a theory can handle the phenomena. Because of that, the simulations employ a methodology that factors the traditional notion of parser accuracy out of the model and focuses entirely on the surprisal and retrieval predictions for the gold parse.

This work departs from previous work not only for this parser but for many human sentence processing models in that accuracy will not be determined by the resulting dependency analysis. In broad-coverage parsing models, accuracy is measured by an f-score, or the precision and recall the parser has in building the correct analysis, or gold parse (Manning & Schütze, 1999, p.269). This particular parsing model has often had high accuracy for psycholinguistic sentences, particularly sentences from eye-tracking corpora (Boston et al., 2008, 2011). But it used a sequence of probabilistic features that had been tuned to work well with broad-

coverage data; it is unclear how accurate the cognitive features considered here would be. Further, the model had been tested on relatively "easy" sentences; CNP and WHI sentences are known to be difficult for humans, and the fact that they involve unbounded dependencies indicate that parser accuracy is likely to be low for even well-tuned implementations like the MALTParser (Nivre et al., 2010). Therefore, the parser is run on the gold parse itself, such as the dependency analysis in Figure 4.6.

This has a variety of advantages. The first, and most important, is that it allows for a clear comparison of exactly the cognitive theories considered here. They are not obscured by, for example, additional features required to ensure relatively good parsing accuracy. Instead, the surprisal and retrieval values will directly tell how difficult it is to make a CNP, WHI, or superiority-violating attachment. Further, it avoids the debate over whether the violating dependencies are created and simply difficult, or if they are not created at all. The answer to this question is unclear, and the different cognitive hypotheses would likely make different predictions. Although it would be interesting to do more in-depth research into this question, this model is kept as simple as possible and the question is avoided. It is assumed that all the cognitive hypotheses make the attachment, and in fact build the same exact dependency analysis. Accuracy is determined by how much difficulty they assign.

Unfortunately, this mode of running the parser also has a disadvantage: the parser must be run in serial mode. This particular parsing model has the ability to be run as either serial or parallel, and previous work has demonstrated that parallel models are most accurate (Boston et al., 2011). Further, preliminary work on SUVs in this parsing model found that using the increased memory of a parallel model offered a simple explanation of the gradience patterns in SUV data (Boston, 2010). Although the gold-parse mode of running the parser eliminates these options, they are still available for future work as this parsing model can easily be run for attachment accuracy.

Despite these issues, the parser will be run on gold parses as this provides the cleanest test

of the contributions of the various cognitive theories to a reductionist explanation of syntactic locality.

4.9 Conclusion

This chapter discusses how a dependency parser is implemented as a cognitive model for syntactic locality. The transition system used has the advantages of being probabilistic, incremental, and broad-coverage, allowing for a simple and accurate cross-linguistic sentence processing model that has been independently successful. It also allows for a simple encoding of cognitive theories, both as probabilistic features and as complexity metrics. The next chapter discusses how the results from this parser are compared to experimental data.

CHAPTER 5

EXPERIMENTS

5.1 Introduction

As with any controversial data set, the work on syntactic locality is rife with examples, counterexamples, and typological challenges. This work incorporates each aspect of this controversy, organized into the following three groups:

- Classic
- Gradience
- Challenges

Classic examples demonstrate the typical linguistic intuition behind a syntactic locality phenomenon. Often, these are syntactic judgments, but when possible they are conditions taken from experimental work. **Gradience** examples are counterexamples to the classic cases. For syntactic locality, they demonstrate gradience, or degrees of acceptability, which are often problematic for grammatical explanations. Cognitive factors should model gradience well, though, and these examples are often cited as evidence for reductionist accounts. **Challenges** are examples that demonstrate CNPs, WHIs, and SUVs going against to cognitive predictions; these are the examples often cited as evidence against reductionsist accounts.

The organizational chart in Figure 5.1 diagrams the experiments considered in this work. Bold boxes indicate that the experiment is used to support either the grammatical or reductionist account. As the results in Chapter 7 demonstrate, the cognitive model can test these claims.

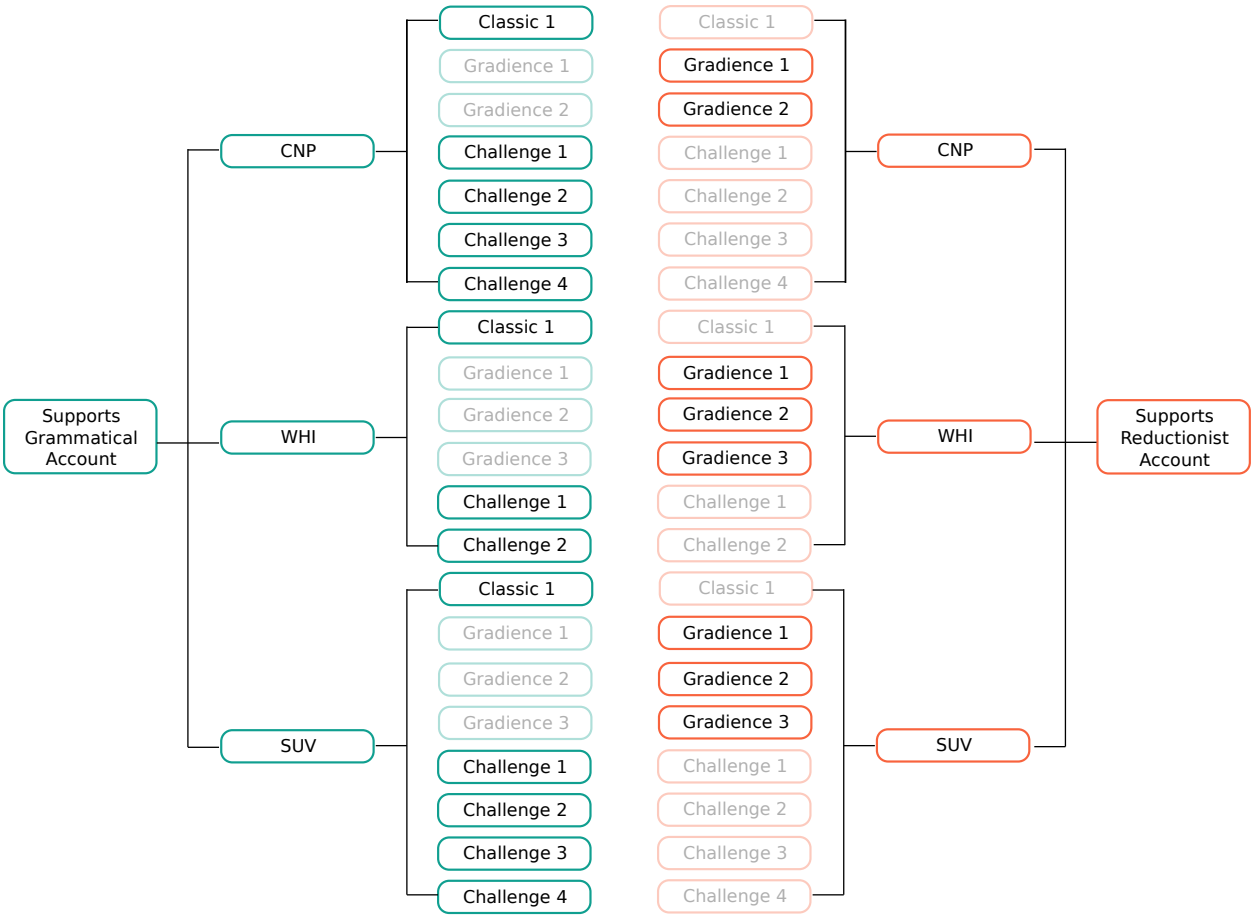


Figure 5.1: An organizational chart of syntactic locality experiments.

The majority of the experimental work on syntactic locality is in English, but a benefit of the broad-coverage model is its ability to handle cross-linguistic data. Further, the broad-coverage model can test the same implementation of a cognitive factor across many languages: one DECAY probabilistic feature is encoded, for example, and then run on multiple treebanks. This research therefore includes cross-linguistic data as often as possible: German data is available for all three phenomena, and it sometimes patterns with and sometimes against English results. Swedish data is considered because of arguments that Swedish speakers do not find CNP and WHI violations difficult. And Russian experimental data for SUVs is used as a challenge to reductionist accounts, since the data suggests Russian does not have SUVs.

This chapter is organized as follows: Section 5.2 discusses various experimental measures. This helps inform the discussion of the experimental data itself, organized by phenomenon in Section 5.3, Section 5.4, and Section 5.5. Section 5.6 details experimental data that was not considered due to a variety of factors.

5.2 Measures

5.2.1 Syntactic judgments

The advantage of the computational model designed here is that its complexity metrics can be compared to fine-grained difficulty measures, such as reading times (Boston et al., 2011). But, it can also model simpler syntactic judgments. Syntactic judgments are the most coarse-grained measure of difficulty: sentences are marked as either acceptable or unacceptable (or, more controversially, grammatical and ungrammatical). Haider (2004) provides the following judgments, where the SUV in the second sentence is classified unacceptable by the asterisk preceding it.

(37) **Who** did you persuade her to visit?

(38) *__Who__ did you persuade ***who*** to visit?

For the studies considered here that include syntactic judgments, the star designation is switched to a numerical one: acceptable sentences are marked 0 and unacceptable sentences marked 1. This aids in comparison to other studies, as discussed in Section 6.3.

5.2.2 Acceptability and Magnitude Estimation

Syntactic studies are only used when experimental results are unavailable. For syntactic locality, the majority of this experimental work uses acceptability judgments. Speakers rate sentences on a scale that goes beyond the binary classification of syntactic judgments. The scale can be ordinal or continuous. Ordinal scales have subjects rate a sentence from, for example, 1 to 7, with 1 being "completely unacceptable" and 7 "completely acceptable". The ordinal scales allow for gradient results, which is a distinct advantage over syntactic studies. But, they also don't allow for the level of accuracy that a continous scale has.

Therefore, there are other studies that use Magnitude Estimation (**ME**) (Stevens, 1975). ME studies operate as follows: subjects are asked to rate a perfectly normal sentence some number. They are then asked to compare subsequent test sentences to this modulus, assigning number ratings such as "twice as difficult". Performing a log-transformation on these judgments allows different subjects' scales to be compared. This technique has the benefit of being directly comparable across speakers while also giving a granularity that can yield statistically significant results (Bard, Robertson, & Sorace, 1996; Cowart, 1997).

5.2.3 Reading Time and Residual Reading Time (RRT)

Some of the studies make use of reading time data from self-paced reading studies. The amount of time it takes to read a word is correlated to processing difficulty, such that the longer it takes to read a word, the more processing difficulty at that word. Self-paced reading experiments generally operate as follows: a subject is shown a sentence on a computer screen one word at a time, with subsequent words revealed by a key press. The lag time between key presses is the reading time for that word.

Experimenters can either use the exact reading time, or use a calculation of reading time

that takes into account word length, Residual Reading Time (RRT). RRT is calculated by first using a linear regression to calculate a prediction of the amount of time it takes for each participant to read a word of a particular length. This prediction is subtracted from the actual reading time to get the RRT. Positive RRTs mean that the word is read slower than predicted, and that subjects are encountering difficulty at that word.

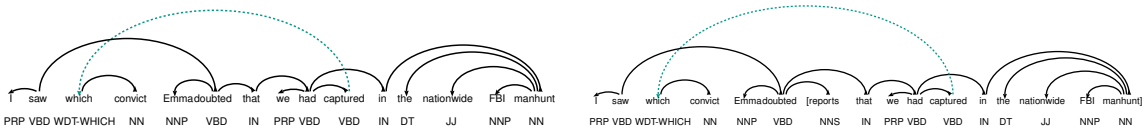
There is variation in these measures, both in terms of granularity and in terms of methodologies. But, even when two experiments use the same measure, their results can not be directly compared. This problem is addressed in Section 6.3; first, the experiments that provide classic, gradience, and challenging examples for each of the phenomena are discussed.

5.3 CNP

5.3.1 Classic

The classic CNP example comes from a larger experiment from Hofmeister and Sag (2010). This experiment tests the effect of both the filler-type and determinacy of the noun on gradience in CNP island violation acceptability. But, two of its conditions demonstrate the classic difficulty associated with CNP violations in English. The conditions are shown in Figure 5.2, along with their dependency analyses. The analyses will be discussed further in Section 6.2.1. In this example, the second condition includes a CNP island violation: *which convict* is extracted from the CNP island phrase [reports that we had captured...]. The island is marked by brackets, and the crucial dependency is marked by a dashed blue line.

This extraction from the island has a higher RRT, indicating that it is more difficult. One problem with this particular data set is that it doesn't demonstrate a strong CNP island effect: the RRT difference is not as high as it would be for a bare noun phrase extraction where



(a) -68.72 (b) -62.10

Figure 5.2: CNP classic example 1: English (measure: RRT).

who is extracted from the CNP. Unfortunately, this data is unavailable from this experiment. Although other experiments and syntactic judgments better demonstrate the island effect, their sentences often have a dependency-length mismatch: the extraction from the CNP island crosses more words than the non-island. As is detailed in Chapter 4.7, the parser has a strong tendency to find longer dependencies more difficult, which could unfairly skew results in favor of the cognitive factors. It is difficult to create experimental conditions where the words are directly matched for the CNP island. In fact, there is a dependency length mismatch in this example: the CNP island is one word longer than the non-island. However, this example is used to demonstrate the classic case because the one-word difference is the smallest found in the English experimental data.

5.3.2 Gradience

The classic example is part of a larger study from Hofmeister and Sag that demonstrates filler-type gradience for CNPs in English. The conditions are shown in Figure 5.3, where *which convict* and *who* are extracted from the CNP island [report that we had captured...]. This experiment also tests the definiteness of the complex noun phrase itself (*reports...* vs. *the report...* vs. *a report...*). Although the study does not find significant differences in terms of noun definiteness, RRTs for the *which convict* conditions are significantly faster than the bare conditions. The *which convict* condition is also faster than the baseline condition, but this result is not statistically significant, nor confirmed by other studies. The overall result demonstrates

gradience in strong islands, which is evidence for a reductionist explanation.

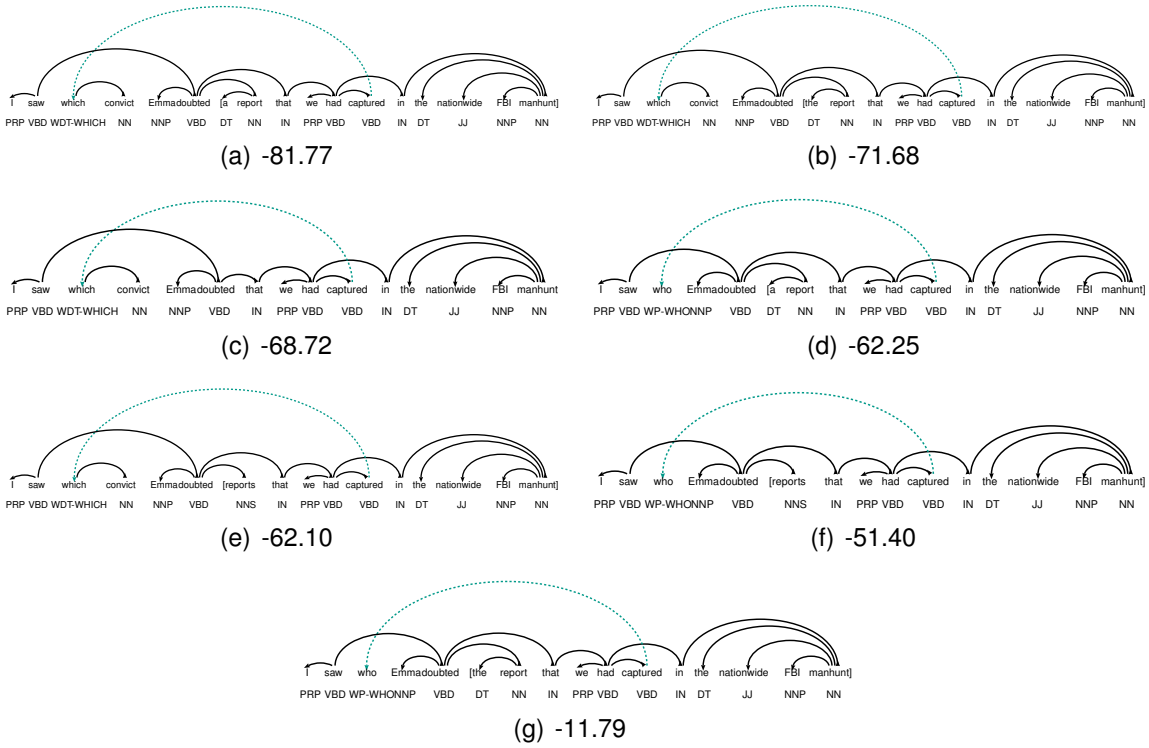


Figure 5.3: CNP gradience example 1: English (measure: RRT).

Keller (1996) also considers gradience in English CNPs with regards to filler-type. But, he considers more filler-types by testing Kluender's specificity hierarchy (1992). The results show a different pattern from the Hofmeister and Sag results. Kluender argues that the more specific the filler, the more acceptable the extraction from the CNP. Keller has subjects rate sentences along this hierarchy using ME. The results are shown in Figure 5.4. The most acceptable condition is the bare condition, *who*. This is followed by the *which theory* condition and the *what* condition, which have almost equal acceptability. The least acceptable condition is the *how many papers* condition. These results do not follow from Kluender's predictions: he predicted that the *which-N* and *how-many-N* conditions would be easier than the bare conditions. But the results confirm CNP gradience for fillers, and further support a reductionist explanation of strong islands.

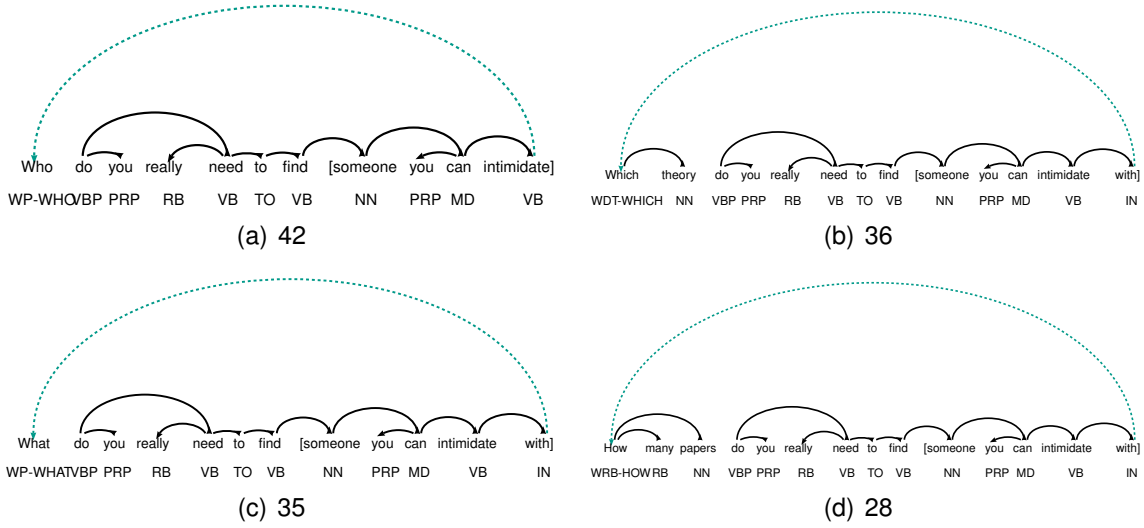


Figure 5.4: CNP gradience example 2: English (measure: acceptability).

5.3.3 Challenges

Reductionists argue that there are many processing factors at play in syntactic locality, and their combination yields a "perfect storm" of difficulty for islands. Sprouse and his colleagues try to tease apart two of these factors, length and syntactic structure, to determine whether the two types of processing difficulty are a) individually available and b) have strong interaction to yield islands. They therefore construct four conditions, shown in Figure 5.5, that alternate length and structure. The first condition, Figure 5.5(a), shows a short-distance dependency, between *who* and *claimed*, with no island syntactic structure. This is the short, non-island condition. The second shows a short-distance dependency which includes an island syntactic structure. Note that *who* is not extracted from the island in this case. The third shows a long-distance dependency, but once again it is not from an island. The final condition, and the one that is found most difficult, has both a long-distance dependency and a syntactic island.

The results are provided with z-scores from an ordinal (7 point) acceptability judgment task. The z-scores are the result of a procedure that corrects acceptability judgments for scale

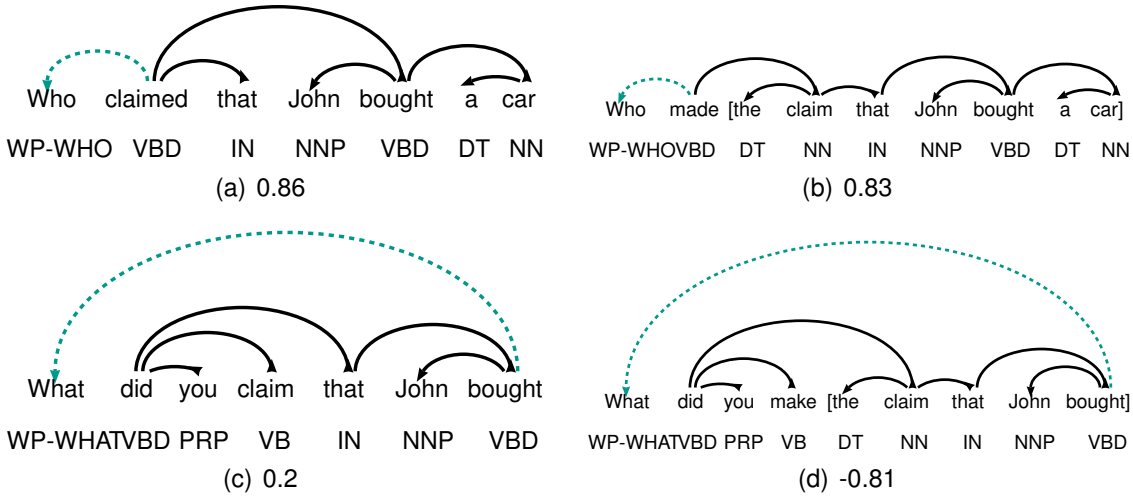


Figure 5.5: CNP challenge example 1: English (measure: acceptability).

bias when comparing to other participants' results[1], but they can be interpreted just as acceptability judgments are (higher z-scores means more acceptable). The authors find a definite interaction of length and structure, meaning that the island effect in the long, island condition holds. They also find a length effect, such that the longer dependencies are less acceptable than the short dependencies. But, they do not find an effect of structure: the short, island condition is not statistically-signifcantly less acceptable than the short, non-island. According to the authors, this result challenges reductionist accounts because if both the structure and the dependency length lead to difficulty in islands, one would expect the structure to be independently difficult.

One concern with this dataset is that the short island condition (Figure 5.5(b)) does not have extraction from within the island itself. The acceptability judgments test difficulty accrued from the whole sentence, so the CNP island should be difficult regardless of whether an extraction has been made or not. But, reductionist accounts are based on the difficulty of retrieving a word across a CNP boundary. As will be discussed below in Section 6.2, this cognitive model tests difficulty at the point of retrieval, which in this case is *claimed*. This means that

[1] See Sprouse, Wagers, and Phillips (To Appear, p.13) for more information on z-score transformations.

the parser's difficulty would be considered before the island is encountered, and would likely not be able to model an island effect. These results are still considered, though, because the structure effect was not statistically significant for the English-speakers.

Alexopoulou and Keller (2007) run a parallel study of CNPs and WHIs in English, Greek, and German. They not only test islandhood, but also how the level of embedding interacts with islands across the phenomena. Figure 5.6 and Figure 5.7 show the English and German experimental conditions along with the results, which are mean acceptability judgments from ME. The results are not as expected: for CNPs in both English and German, the doubly-embedded structure is more acceptable than the singly-embedded structure. This result seems to be at odds with a processing explanation, and leads the authors to conclude that CNPs are better-explained by grammatical approaches. One would predict that the cognitive factors will have difficulty modeling this behavior.

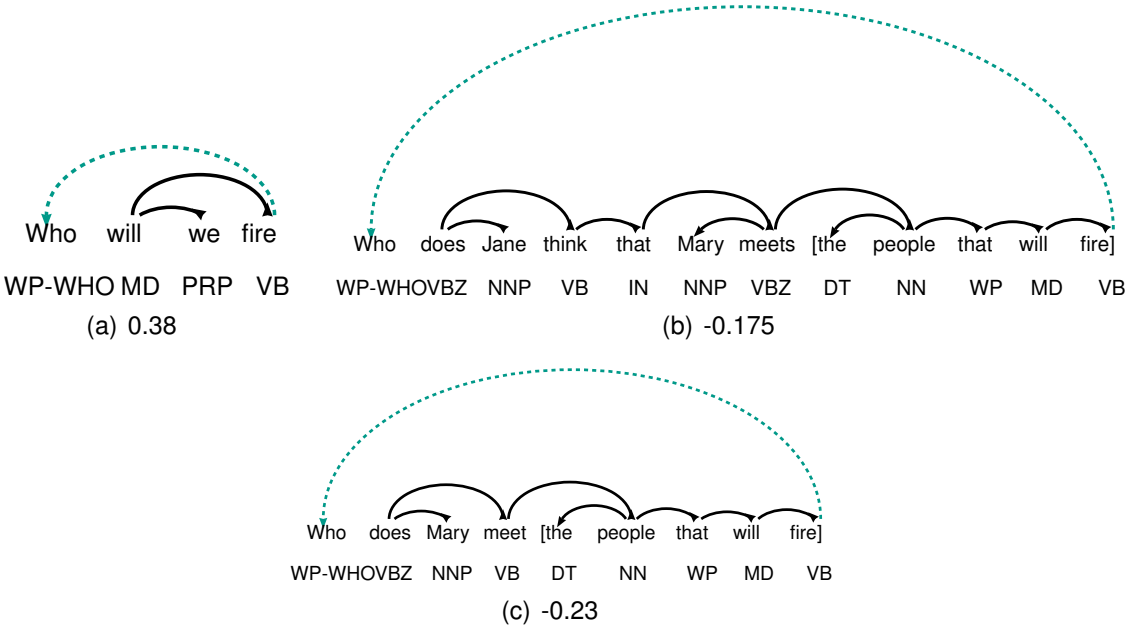


Figure 5.6: CNP challenge example 2: English (measure: acceptability).

Finally, perhaps the most well-known challenge to any account of CNPs is data from Swedish that indicates that extraction from CNP islands is acceptable. Kush and Lindahl

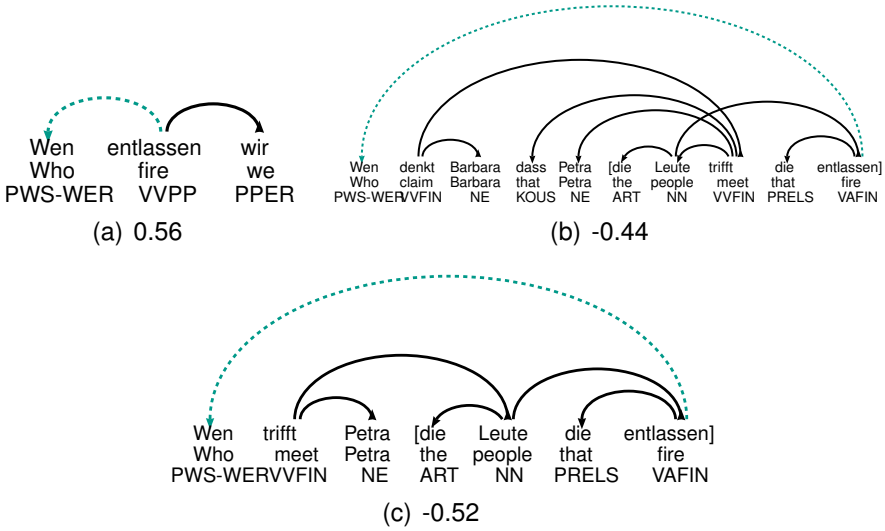


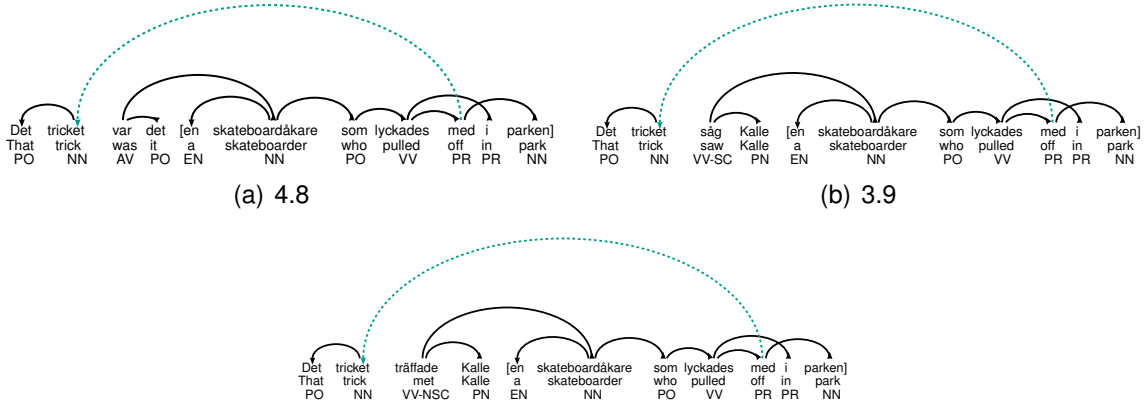
Figure 5.7: CNP challenge example 3: German (measure: acceptability).

(2011) provide acceptability judgments for non-island (Figure 5.8(a)) and island (Figure 5.8(b) and Figure 5.8(c)) conditions that have equal-length dependencies. The authors use acceptability judgments to find that the standard CNP island and the non-island have relatively equal acceptability.

The authors provide a grammatical explanation for what they consider to be this atypical behavior in Swedish. They claim that the apparent acceptability of CNPs is not evidence of the constraint not existing, but rather behavior typical of verbs that take small clauses. The word *såg*, "saw", is a verb that takes a small-clause, and Kush and Lindahl argue that extraction from islands within these verbs is easier than extraction from islands within verbs that do not take a small clause, such as *träffade*, "meet". Their hypothesis is born out by the acceptability results: the non-small clause verb condition in Figure 5.8(c) is less acceptable.

This experiment is considered for the typical Swedish CNP island-violating behavior represented by the first two conditions. To test the full Kush and Lindahl hypothesis, the treebank was also modified to be sensitive to both small-clause and non-small clause verbs. There were simply not enough data points to support the distinction and allow any of the cognitive

hypotheses to distinguish the two. However, the acceptability of Swedish CNPs still stands as a challenge to reductionist accounts: if there is something inherently difficult about retrieving a word across the CNP island, shouldn't this difficulty exist across languages? Or, if there is a structural difference between the two, could the cognitive constraints make the correct prediction?



(a) 4.8 (b) 3.9

(c) 2.8

Figure 5.8: CNP challenge example 4: Swedish (measure: acceptability).

5.4 WHI

5.4.1 Classic

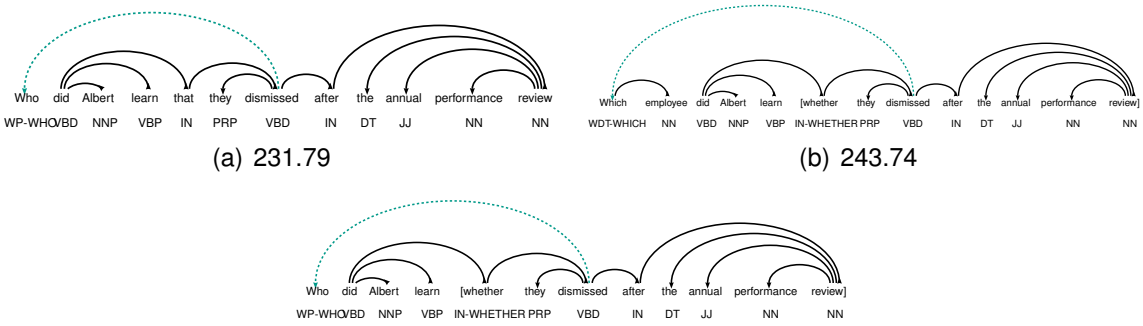
Like the classic example for CNPs, the classic example for WHIs is taken from a larger experiment on WHI gradience from Hofmeister and Sag (2010). This experiment includes the classic conditions demonstrating that extraction from a WHI is more difficult than extraction from a non-island (Figure 5.9). The results, shown in milliseconds of reading time, demonstrate that extracting *who* from the island *[whether they dismissed...]* is more difficult than extracting from the non-island condition. This is the classic WHI result common in the literature.

(a) 231.79 (b) 255.96

Figure 5.9: WHI classic example 1: English (measure: RT).

5.4.2 Gradience

Hofmeister and Sag (2010) also test filler-type gradience, shown in Figure 5.10. They add a condition, switching *which employee* for the bare *who* in Figure 5.10(b). The results are that the *which employee* extraction is more acceptable than the bare extraction, although it is not as acceptable as the non-island. This result lends support to the reductionist argument because the more informative filler leads to less difficulty.



(a) 231.79 (b) 243.74

(c) 255.96

Figure 5.10: WHI gradience example 1: English (measure: RT).

As was mentioned in Section 5.3.3, Alexopoulou and Keller (2007) tested how embedding interacts with islands in English, German, and Greek. The English and German conditions for WHIs are replicated in Figure 5.11 and Figure 5.12. The results, which are given as mean acceptability judgments from ME, are as one would predict: more embedding means less acceptability. This contrasts with the CNP results and lead the authors to argue that WHIs are more suitable for a reductionist explanation. For our purposes, this level of gradience should

be something that the cognitive factors can easily handle.

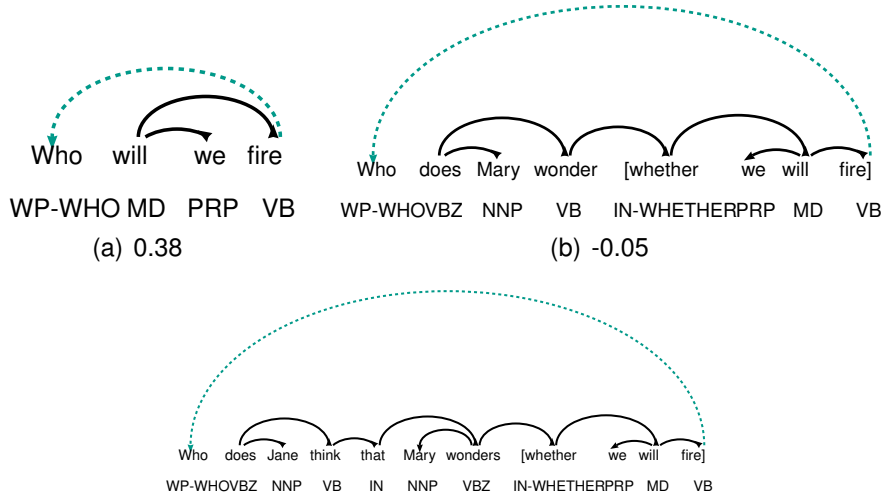


Figure 5.11: WHI gradience example 2: English (measure: acceptability).

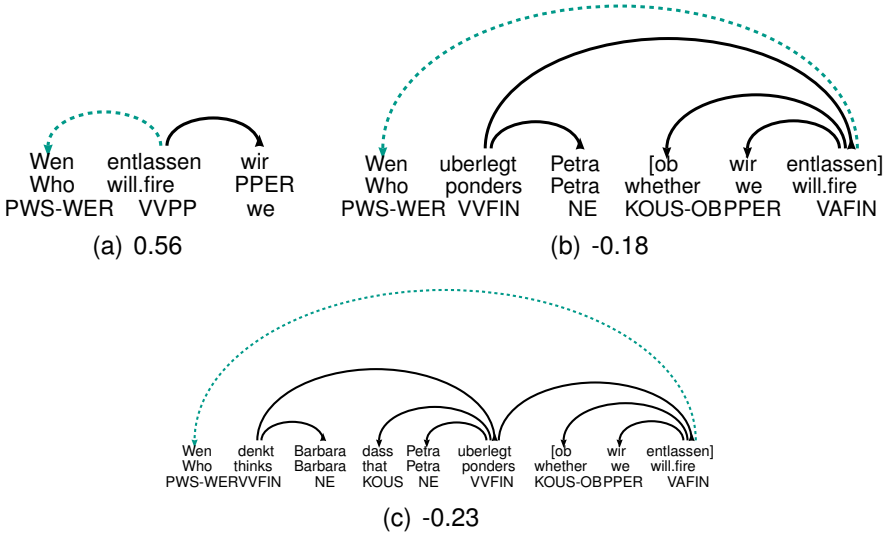


Figure 5.12: WHI gradience example 3: German (measure: acceptability).

5.4.3 Challenges

Sprouse et al. (To Appear) consider WHIs in addition to the CNPs discussed in Section 5.3.3. But, like the Alexopoulou and Keller (2007) data, the WHIs behave in keeping with reductionist assumptions. Namely, in their experiment that tests the interaction of both length and syntactic structure on WHIs, they find that there is an interaction between the two (an island effect) as well as individual effects for both length and syntactic structure. The conditions are provided in Figure 5.13, and like the CNP study include a short non-island (Figure 5.13(a)), a short island (Figure 5.13(b)), a long non-island (Figure 5.13(c)) and a long island (Figure 5.13(d)). Unlike CNPs, though, they find a significant effect on islands: acceptability is lower for the short island condition than it is for the short non-island. Note that this is despite the fact that there is no extraction from the short non-island. This experiment will likely pose a problem for the retrieval-based cognitive constraints, which will not be able to differentiate the first two conditions. This result highlights the possibility of storage-based difficulty for WHIs.

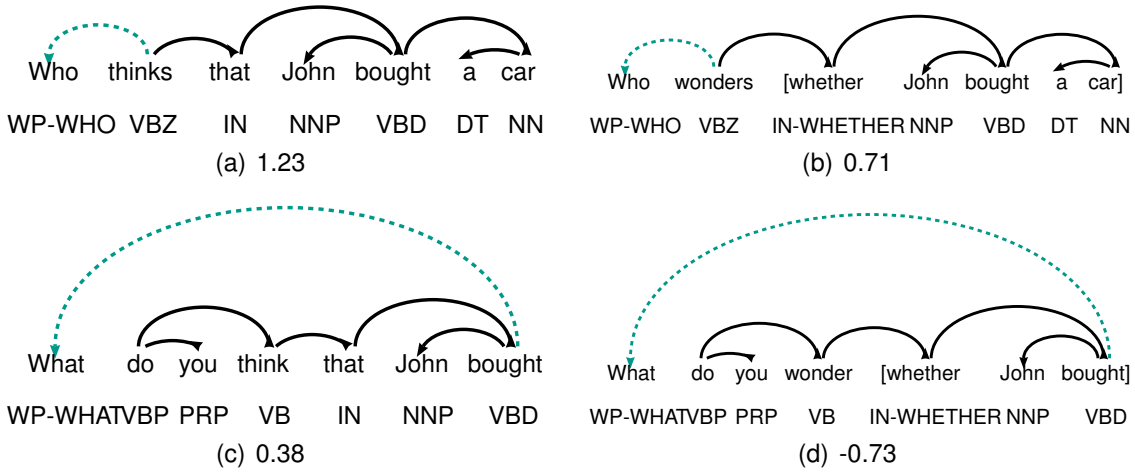


Figure 5.13: WHI challenge example 1: English (measure: acceptability).

Unlike the other types of syntactic locality, there are few examples that challenge reductionist accounts for WHIs, either experimentally or in the literature. One exception is Swedish, where Maling and Zaenen (1982) report that speakers do not have difficulty with WHI extrac-

tion. The conditions are provided in Figure 5.14, where the second sentence is an example of a WHI violation but is reported as acceptable via syntactic judgments. This data poses a challenge for reductionist accounts since the cognitive factors should behave similarly across languages. However, note that although the conditions are matched for dependency length, the syntactic structure is different. This difference could be a factor in the judgments reported by Maling and Zaenen (1982), and could aid the cognitive factors in modeling even this challenging result.

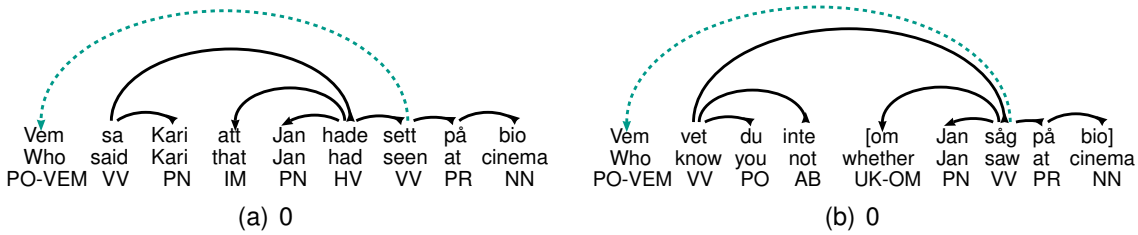


Figure 5.14: WHI challenge example 2: Swedish (measure: syntax).

5.5 SUV

5.5.1 Classic

The classic example for SUVs demonstrates the simple distinction that extracting a wh-word past another wh-word is more difficult than extracting a wh-word past a non wh-word. The conditons come from Haider (2004) and are provided in Figure 5.15, where Figure 5.15(a) shows a pronoun intervenor and Figure 5.15(b) shows a wh-intervenor. Note that all other aspects of the sentence, including the syntactic structure and the dependency lengths, are matched. But, syntactic judgments reveal the second sentence to be less acceptable.

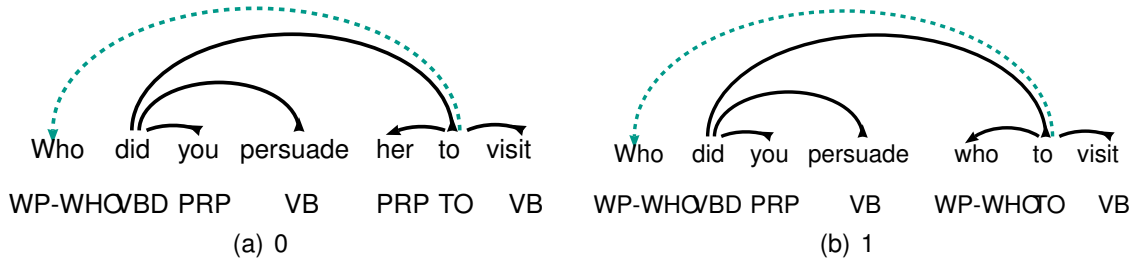


Figure 5.15: SUV classic example 1: English (measure: syntax).

5.5.2 Gradience

Gradience in SUVs has been reported for a long time (Karttunen, 1977). Arnon et al. (To Appear) demonstrate gradience of both filler and intervenor type in SUVs using RRT from a self-paced reading task. Their conditions and their results are provided in Figure 5.16, where the *which-N* filler and intervenor yields the lowest RRT and least difficulty (Figure 5.16(a)), and the bare filler and intervenor yield the highest RRT and difficulty (Figure 5.16(d)). The bare.which (Figure 5.16(b)) and the which.bare (Figure 5.16(c)) are intermediately difficult, as expected by a reductionist account.

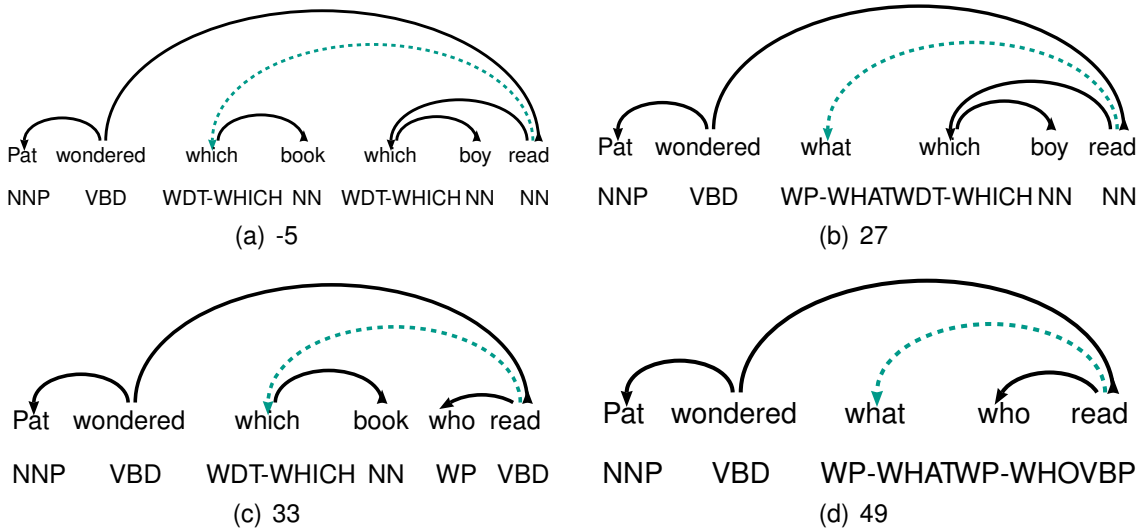


Figure 5.16: SUV gradience example 1: English (measure: RRT).

Although the RRTs reflect standard reductionist assumptions, the authors ran a simulta-

neous acceptability judgment task that yields slightly different results. Here, the bare.which condition is harder than the bare.bare condition. This result is precisely replicated by an acceptability judgment study from Fedorenko and Gibson (Submitted). Fedorenko and Gibson do a parallel study testing availability of SUVs and gradience in SUVs in both English and Russian. They find an SUV effect in English, but this work considers the gradience effect, replicated in Figure 5.17. The conditions are similar to Arnon et al. (To Appear), and the bare.which condition is considered more difficult than the bare.bare condition. Fedorenko and Gibson (Submitted) note that because these results are from an ordinal acceptability judgment task, the measure might be too coarse to give completely accurate results on this data set. Given the conflict, it will be interesting to find out which ordering the cognitive factors confirm.

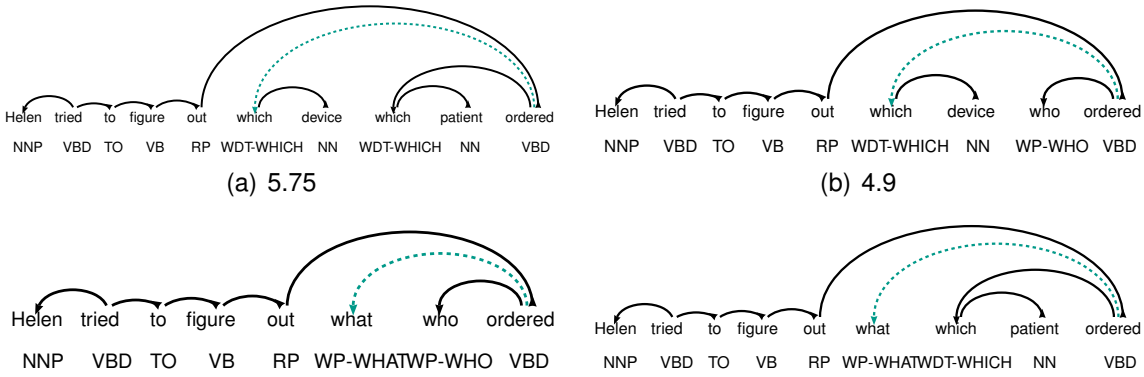
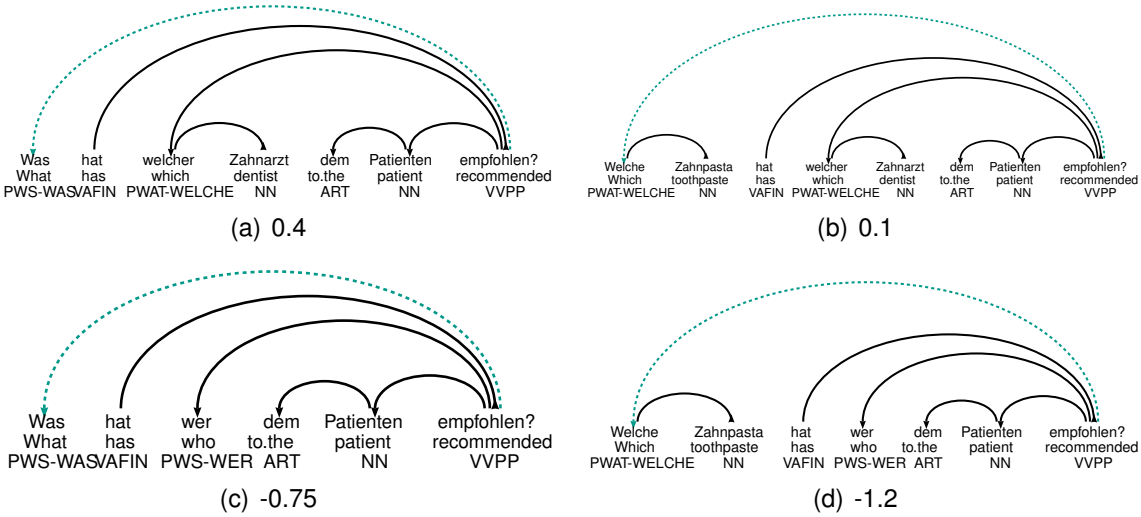


Figure 5.17: SUV gradience example 2: English (measure: acceptability).

Like the English studies, Featherston (2005) considers the effect of filler and intervenor type on German SUVs. But, he also considers three cases: nominative, accusative, and dative. Featherston's data suggests that the SUV condition holds in German. This work focuses on the experimental conditions that show evidence of gradience in German, provided in Figure 5.18. The German SUV ordering is markedly different from either of the English orderings, and also goes against what reductionist hypotheses would predict. Here the easiest condition is the bare.which (the most difficult condition for Fedorenko and Gibson (Submitted)). This is

followed by the which.which, the bare.bare, and the which.bare; the bare.bare condition is therefore not the most difficult. All told, this result runs counter to reductionist predictions, yet reveals gradience that could be handled by cognitive factors.



(a) 0.4 (b) 0.1

(c) -0.75 (d) -1.2

Figure 5.18: SUV gradience example 3: German (measure: acceptability).

5.5.3 Challenges

A challenge for not only the Featherston data but reductionist theories in general are the many linguistic studies that claim that German speakers do not find SUVs difficult. A classic example, using syntactic judgments, is provided by Fanselow and Féry (2007) and replicated in Figure 5.19. Here, both the SUV in the second sentence and the non-SUV in the first are rated as acceptable.

Fanselow and Féry (2007) provide a follow-up to the Featherston (2005) study that argues against the study's main claims. They argue that what is driving the supposed SUV effect in Featherston's results is case rather than SUVs. They mention that German behaves unpredictably when two animate dependents interact, and the additional dative dependent in the Featherston study, *dem Patienten*, "the patient", is causing problems.

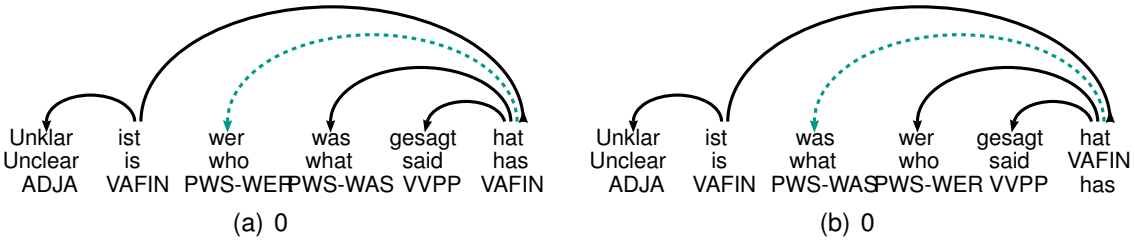


Figure 5.19: SUV challenge example 1: German (measure: syntax).

Fanselow and Féry suggest that SUVs should not be available in German because German is verb-second. Because many of its verbs occur at the end of the sentence, speakers are used to holding words in memory until the end of the sentence. Further, case-marking allows speakers to keep thematic roles separate. They therefore design an ordinal, 7-point scale acceptability judgment experiment that takes into account these factors; the conditions are provided in Figure 5.20. An accusative SUV (Figure 5.20(a)) is compared to an accusative non-SUV (Figure 5.20(b)), where the wh-intervenor *wem* "who" is switched to the personal pronoun *ihm* "him". And a dative SUV (Figure 5.20(c)) is compared to a dative non-SUV (Figure 5.20(d)). The results support the long-standing conclusion that German does not have SUVs: the SUV conditions (Figure 5.20(a) and Figure 5.20(c)) are considered more acceptable than the non-SUV conditions. They also find an effect of case: the dative conditions are more difficult than the accusative conditions. The authors conclude that the Featherston results are difficult to interpret because they do not take into account this effect of case on SUV behavior.

The Fanselow and Féry (2007) data poses a challenge to reductionist accounts because it demonstrates cross-linguistic variability. Similarly, evidence suggests that Russian speakers also do not find SUVs difficult. The Fedorenko and Gibson (Submitted) study has conditions translated directly from its English counterpart, described in Section 5.5.2. The conditions are provided in Figure 5.21, and include both the unextracted cases (non-SUVs) and the extracted cases described in the English study. The results are mean acceptability judgments, which were taken from a 7 point scale where 1 is considered unacceptable and 7 is considered

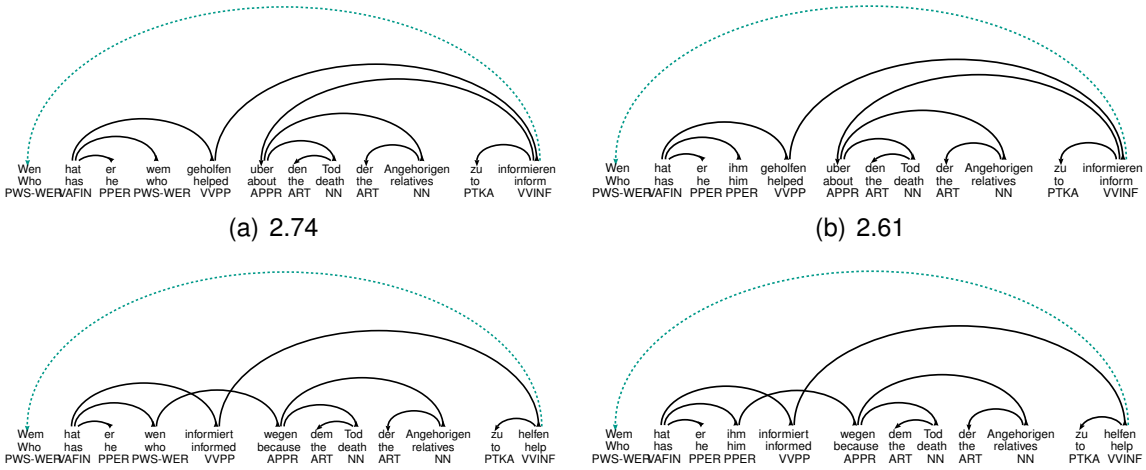


Figure 5.20: SUV challenge example 2: German (measure: acceptability).

acceptable. One thing to note is that the sentences have similarly high acceptability across conditions, demonstrating that SUVs are not difficult for Russian speakers. In fact, the most acceptable condition is the bare.bare SUV condition, which for English would be very difficult, if not the most difficult. Another surprising finding is that the filler type does not have an effect on Russian, even though it did have a statistically-significant effect in English. Reductionist accounts would predict that filler-type should always be a factor in difficulty, regardless of language. But, reductionists could also argue difficulty is obscured in easy sentences, particularly for a coarse-grained measure like ordinal acceptability.

All together, the Fanselow and Féry (2007) German data and the Fedorenko and Gibson (Submitted) Russian data pose a challenge for reductionist accounts. They demonstrate that SUVs and even SUV gradience patterns do not behave similarly acroos languages. Although these studies challenge reductionist accounts, they do not necessarily negate them. Rather, it could be that typological factors, like case and verb-final ordering, relieve processing difficulty for these languages. With a broad-coverage model sensitive to frequency effects, the cognitive factors could make the correct prediction.

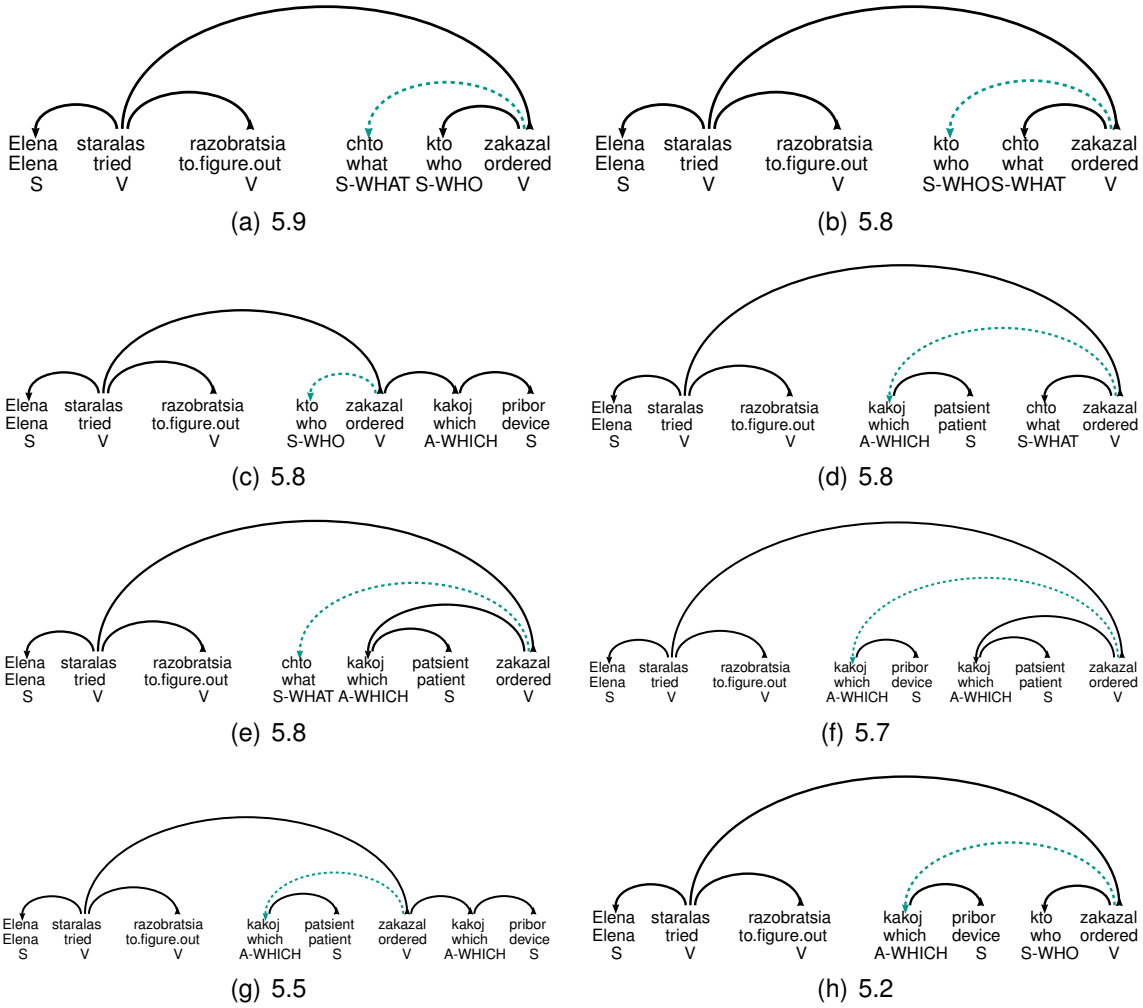


Figure 5.21: SUV challenge example 3: Russian (measure: acceptability).

5.6 Experiments that were not modeled

One final thing to note regarding the experiments selected above is that they do not represent every linguistic judgment or experimental data point ever found for syntactic locality. Instead, the focus is on experimental data that provides a mix of classic, gradient, and challenging data for reductionist accounts. Many other experiments were considered but ultimately left out of this discussion. Some were not considered because the dependency parser architecture could not easily provide accurate results. For example, the tendency for long dependencies to have higher difficulty than shorter dependencies, described in Chapter 4, would incorrectly provide

positive results for studies with grossly mismatching dependency lengths between conditions, such as syntactic judgments on Swedish CNPs and SUVs by Allwood (1982); Andersson (1982); Engdahl (1982); Maling (1978), and English SUVs (Bošković, 1998). For the Swedish CNP data, these results were subsumed by an experimental result by Kush and Lindahl (2011). Because Swedish SUVs behave similarly to German and Russian, the decision was made to forego these studies altogether. Similarly, the English results were subsumed by other experimental results.

Other experimental results were not considered because they were subsumed by larger studies that use finer-grained measures. For example, Keller (1996) studied the effects of definiteness on CNP islands using acceptability judgments. But, this study was subsumed by a larger study by Hofmeister and Sag (2010), who used reading time data to find no effect of definiteness. Also, Hofmeister (2007) reports results on SUV gradience that includes conditions that are very similar to the Arnon et al. (To Appear) results; because these results do not add to this discussion, this work focuses on the original Arnon results.

There is a challenge to reductionist accounts of SUVs not considered here. The argument has been made that adding a third wh-intervenor to SUVs causes more acceptable sentences (Kitahara, 1993; Reinhart, 1995). This result has since been overturned by Clifton et al. (2006), who find via an acceptability experiment that triple SUVs are, in fact, more difficult.

Finally, this study does not consider arguably the most famous experimental result for islands, from Stowe (1986). Stowe demonstrates that the parser does not consider attaching a wh-word within a strong island. Her reading time-based experiment demonstrates that not only is the attachment not made, it is not even considered by the parser. This result tests a reductionist hypothesis, but is difficult to model because the result argues against attachment. The current model is therefore not able to accurately model her conditions, which is in keeping with the argument that CNP may in fact be better explained by grammatical factors.

Despite the various experiments that were left out, the studies modeled in this work provide a suitable cross-section of the syntactic locality data. It is hoped that the addition of the categorical organization among the studies diminishes any potential loss in breadth.

5.7 Conclusion

This chapter discusses the experimental data used test this cognitive model. The data not only shows evidence for each of the syntactic locality phenomena, but it also shows experimental evidence both for and against reductionist accounts. Although there are a variety of studies tested, this work does not consider every experiment. Many were left out because they were subsumed by studies that were more conducive to modeling, either because they used finer-grained measures or because the experimental conditions were better-matched. The next chapter discusses how the experimental results are compared to parser difficulty.

CHAPTER 6

A METHODOLOGY FOR COGNITIVE MODELING

6.1 Introduction

This work compares a cognitive model's predicted difficulty to real human difficulty for syntactic locality data. The results quanitify how well a particular cognitive factor models CNPs, WHIs, or SUVs, graphed as in Figure 6.1.

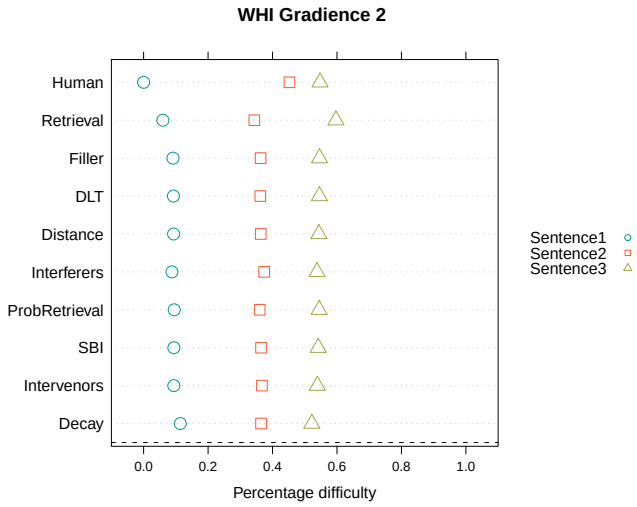


Figure 6.1: A sample graph comparing cognitive predictions to human difficulty.

Figure 6.1 provides explicitness previously unavailable for reductionist approaches of syntactic locality. However, the large-scale nature of this study, taking into account many cognitive factors, many phenomena, and many languages on broad-coverage model leads to complications. The following chapter outlines a methodology for cognitive modeling based on the techniques and approaches required for this variable data set. Figure 6.2 diagrams the methodology, which takes raw experimental results, translates them into dependency diagrams suitable

for the computational model, runs them through a parser to obtain difficulty predictions for each of the cognitive factors, and then allows for a comparison of these predictions to the human data.

Section 6.2 describes how the experiments are translated into a parsable format that allows for comparison to other experiments both within and across phenomena. Section 6.3 describes each of the steps highlighted in the evaluation process of the methodology diagram. Section 6.4 concludes.

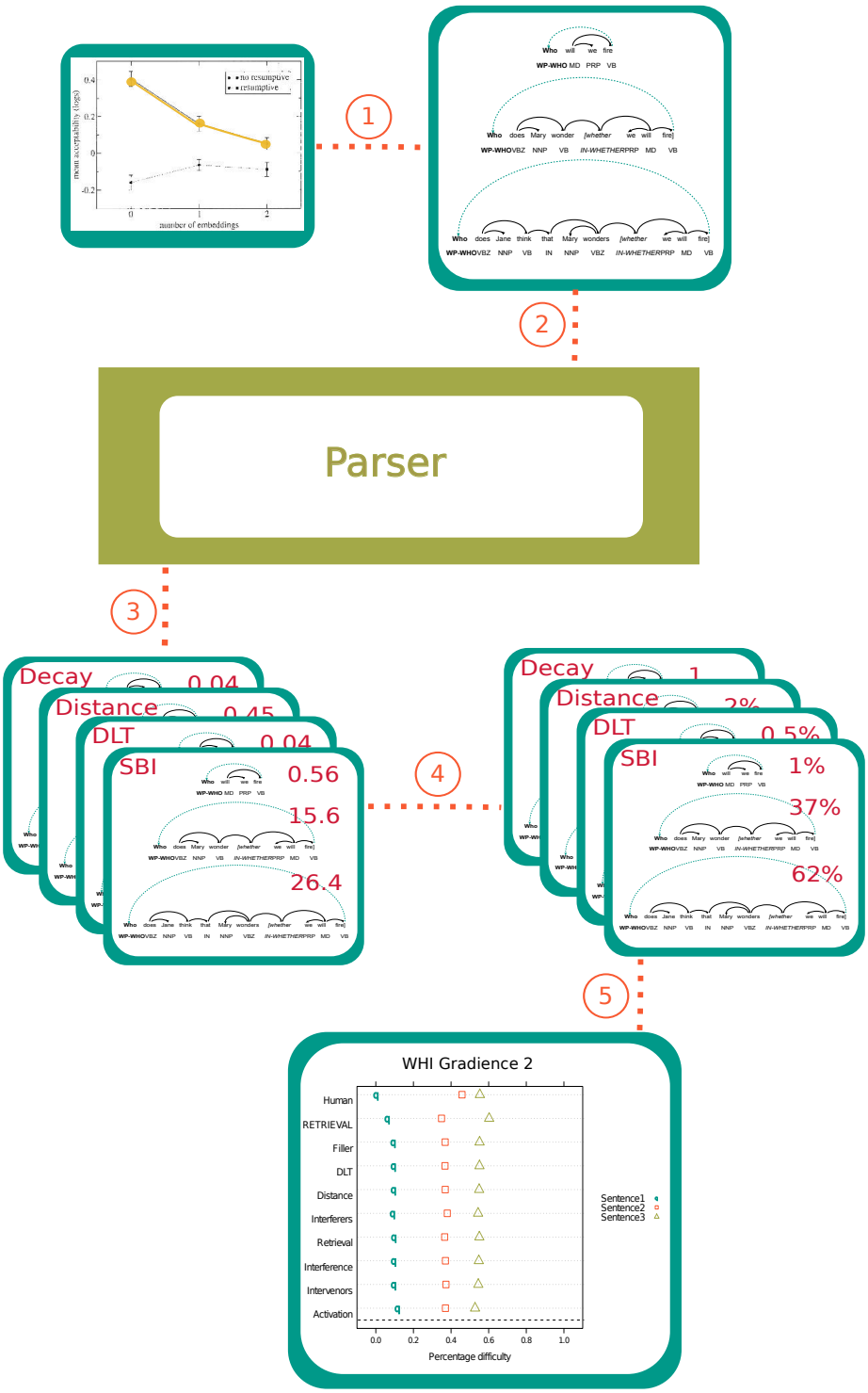


Figure 6.2: The evaluation process.

6.2 Encoding practice

The discussion so far has focused on human acceptability: regardless of the measure used to obtain the experimental results in Section 5.1, the focus of the discussion was on how acceptable human speakers found the conditions. The term acceptability glosses over an issue with comparing experiments, and this is the difference between what the experiments actually measure. For example, many syntactic judgments use binary acceptability, but the goal is to measure *grammaticality*. Acceptability judgments are slightly more fine-grained, and provide either an ordinal or continuous scale upon which to measure *acceptability*. And fine-grained measures like reading time and RRT can provide insight into *difficulty*. But, these terms do not always overlap in meaning: ungrammatical sentences are unacceptable, but not all unacceptable sentences are ungrammatical. A classic example, from Miller and Chomsky (1963), is the sentence with multiple levels of embedding in (39). This sentence is unacceptable for English speakers because it taxes working memory and is difficult to process. Yet it is grammatical.

(39) The boy the dog the cat chewed chased laughed.

Complicating matters further, sentences can be difficult without being either unacceptable or ungrammatical. For example, object relative clauses are more difficult than subject relative clauses, but these sentence are both grammatical and acceptable for English speakers. (40) and (41) show a classic example, from Grodner and Gibson (2005) where the object relative clause in (41) is more difficult than the subject relative clause in (40).

(40) The reporter who sent the photographer to the editor hoped for a story.

(41) The reporter who the photographer sent to the editor hoped for a story.

Although grammaticality is a factor in acceptability, it is not the whole story: a multitude of processing factors can also be at play. And although unacceptability and difficulty usually co-occur, they do not always do so. The experimental measures considered can not reliably differentiate between these factors, and although the experiments themselves are usually designed to do so, comparing across experiments leads to problems. How can one compare a syntactic study that argues a sentence is ungrammatical from a reading time study arguing that a particular region is difficult? This is not a new problem (Miller & Chomsky, 1963; Cowart, 1997; Featherston, 2005; Clifton et al., 2006; Featherston, 2009).

The focus of this research is measuring difficulty, as that is the quantity that the complexity measures from the parser give. It is also reliably available in the syntactic locality data considered here: islands and SUVs are difficult; whether this arises from acceptability or grammaticality is a controversial matter. Gradience, be it in acceptability judgments or reading times, can not draw the line between acceptable sentences and unacceptable sentences, or grammatical sentences or ungrammatical sentences. But, gradience can provide information on which sentences are more difficult than others.

This remainder of this section discusses how this work encodes the experimental data to bring into focus the difficulty associated with syntactic locality, and turn it into a format that can be compared to the parser's difficulty. Section 6.2.1 describes how to assign the sentences dependency analyses, and Section 6.2.2 discusses how the central point of difficulty is identified as one particular dependency arc for each sentence.

6.2.1 Assigning dependency analyses to experimental sentences

The parser builds dependency analyses of sentences. But, it is also possible to assign a dependency analysis to sentences by hand, and then use the parser to simply provide the com-

plexity measure results for that particular dependency analysis. This assigned dependency analysis is the gold-parse, or the analysis that is correct from a grammatical perspective. Consider the Alexopoulou and Keller (2007) conditions for WHIs in English, which are repeated below in Figure 6.3. The dependency analyses assigned follow from standard dependency grammar rules: verbs are the root of the dependency analysis. For islands and SUVs, the extracted element is a dependent on the embedded verb, marked by the dashed blue dependency.

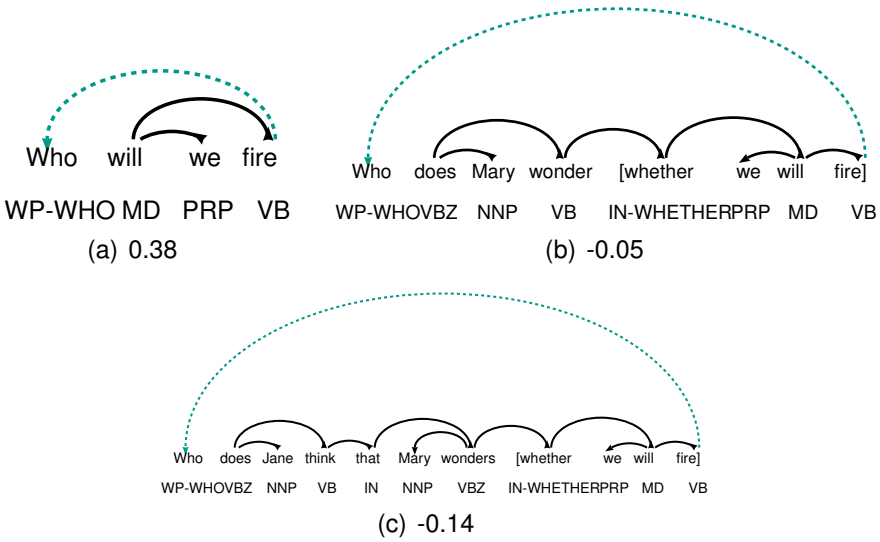


Figure 6.3: A WHI gradience example. Measure: acceptability.

For ambiguous or controversial decisions, such as whether to follow determiner phrase rules or noun phrase rules, the practices in the treebanks were followed. Although this allows the parser to get probabilities from the standards that are set by the treebank, a complication arises when treebanks make different decisions. For example, the Brown corpus for English uses embedded verbs as the heads of relative clauses, whereas the Russian Treebank uses the complementizer as the head of the relative clause. Because this variation did not directly effect syntactic locality processing, treebank procedures were used rather than attempt to standardize across languages.

An issue that does effect the modeling of syntactic locality is the NP analysis of noun phrases common to the treebanks, where the noun is the head of the determiner. Many of the experiments test *which*-noun filler gradience, and in this case it is the wh-phrase that should be retrieved from memory, not the noun itself. All *which*-noun phrases were switched so that *which* was the head of the noun, and the dependent of the verb. This had the added benefit of helping interference features, since retrieving *which* interferes with other wh-words, but retrieving its noun does not. One negative to this transformation is that it made dependency lengths for *which*-noun conditions longer than bare conditions. More details on this transformation are available in Chapter 4.4.1. More details on why longer dependency lengths can be a problem are available in Chapter 4.7.

The dependency analysis assigned to a sentence is based on the standards set by the treebanks and their head-finders. This provides a standard grammar across experiments; the next section discusses how to compare them.

6.2.2 Determining the crucial arcs

This research considers two parser-based complexity metrics, surprisal and retrieval, for comparison with human difficulty. One issue is that these metrics are not full-sentence measures: both give difficulty measures that correspond to words within a sentence. Although they could be converted to sentence-long measures, this doesn't capture the essence of syntactic locality. Whether from a grammatical perspective or from a processing perspective, the problem that syntactic locality addresses is assigning the fronted wh-word to its proper place, which for dependency grammar would be as the dependent of its verb. Creating this attachment is what is difficult: grammatical approaches argue that it is because of barriers to movement, such as islands or superiority constraints, and reductionist approaches argue that it has to do with correctly retrieving the word and attaching it.

For this reason, this crucial dependency between the extracted wh-phrase and its head verb is identified as the focal point of difficulty for the sentence. In particular, the region of interest will be at the head verb after the dependency is formed. For experimental measures like reading time, this makes sense: the region of interest for the reading time spike is at the head verb as well. For measures like acceptability and syntactic judgments, an argument can be made that the unacceptability and/or difficulty is also being driven by difficulty at the integration region. The reading time studies that have been done confirm this. Sentence-wide difficulty measures for acceptability and syntactic judgments are therefore assigned to the head verb of the crucial arc. This would be *fire* in the Alexopoulou & Keller examples in Figure 6.3.

This crucial dependency is by necessity created after attaching other dependents of the verb that occur after the wh-word. The verb therefore often has a high surprisal or retrieval time that could be due to other attachments. But this helps the parser be a more accurate model–syntactic locality difficulty does not occur in a vacuum, and the experimental measures are likely taking this difficulty into account as well.

Another issue is that any difficulty that is coming from structure built after this attachment can not be considered by the parser. The Sprouse et al. (To Appear) data in Chapter 5.3.3 and Chapter 5.4.2 demonstrates why this could be problematic, as the island structure appears after the dependency for some conditions. However, because this model tests integration-based cognitive constraints, this is not considered a shortcoming of the model, but rather of the particular cognitive constraints.

6.3 Evaluation process

Evaluating a cognitive model is a complicated task; not only are there the multiple experiments and experimental methods to contend with, but also the multiple phenomena, the multiple languages, and the various cognitive factors that will be tested. But at the larger scale, there is simply no consistent method for evaluating cognitive models as there is in other areas of computational linguistics research. Keller (2010) argues that this can be rectified in the future by developing a standard test set for comparison against psycholinguistic data. This will allow for better, quantitative evaluation methods against psycholinguistic data. Because this is not yet available, particularly for a study such as this one, this work relies on standard qualitative evaluation: cognitive factors that assign difficulty in an order similar to the human order are considered accurate models.

This section focuses on the procedure that allows the parser's difficulty measures to be compared to the human difficulty measures. Figure 6.4 provides a diagram of this evaluation process; this section steps through the diagram to demonstrate how cognitive factors can be compared to human difficulty for syntactic locality.

6.3.1 ① Transform into dependency analyses

The process of turning the experimental data into dependency analyses is detailed in Section 5.1 and Section 6.2. In brief, the sentences are assigned dependency analyses using standards set by the treebanks and the headfinder rules described in Chapter 4.4.1. For each sentence, the crucial arc is found between the extracted wh-word and its verbal head, and the sentence difficulty is assigned to the verbal head.

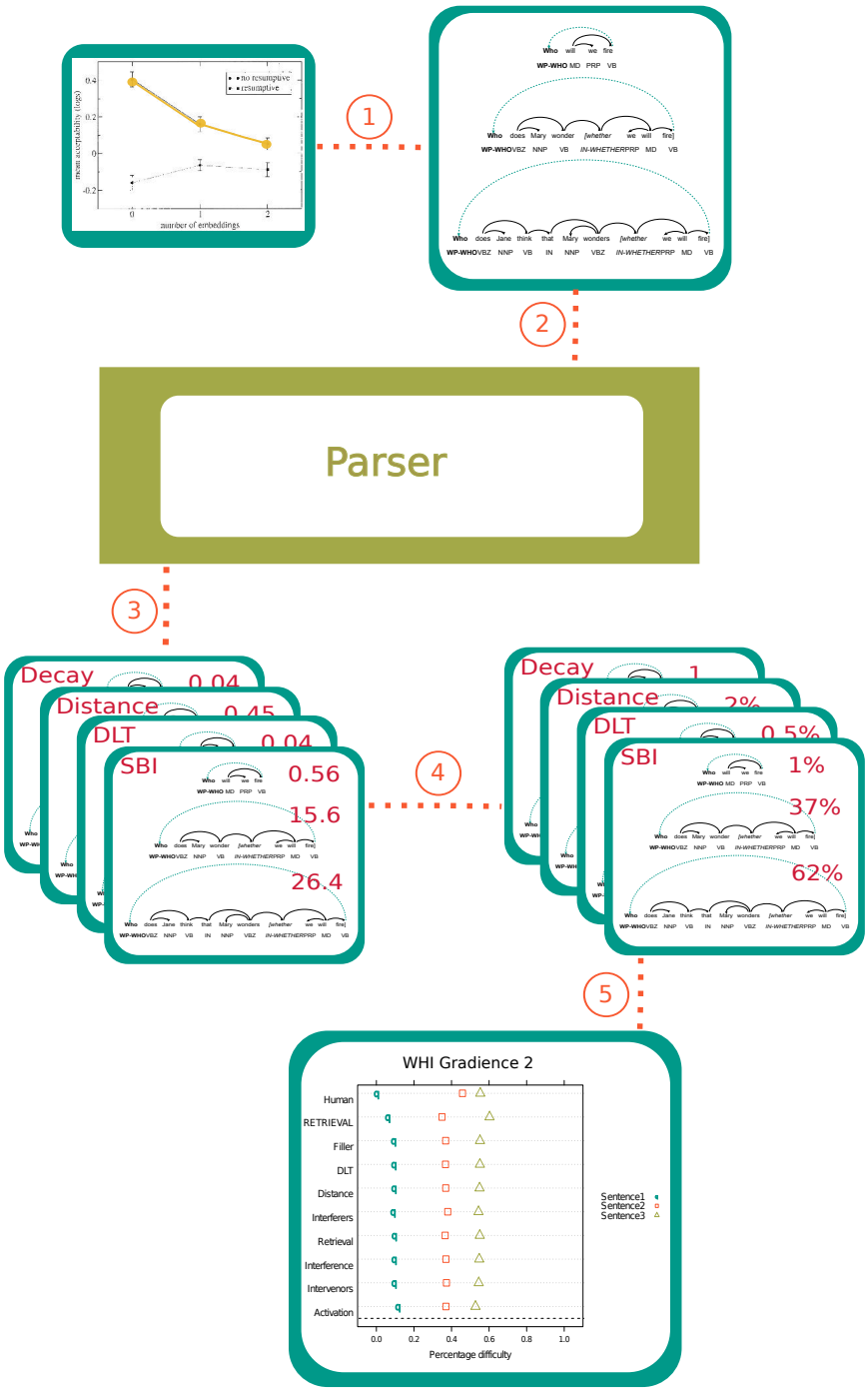


Figure 6.4: (1) Transform into dependency analysis. (2) Run through the parser. (3) Obtain difficulty measures for each cognitive factor. (4) Transform into percentage difficulty by condition. (5) Compare to human difficulty.

6.3.2 ② Run through the parser

The parser has two modes: in one, the cognitive factors are tested on how accurately they create the dependency analysis that is most like the one that we think humans are creating. Or, the parser can be given the correct dependency analysis, and the cognitive factors are tested on the difficulty they assign to that analysis at a particular region, like the verb. This work considers this second mode[1]. The result of the second mode is a series of difficulty measures for each of the words in the experimental sentences. These are described further in the next section.

6.3.3 ③ Obtain difficulty measures for each cognitive factor

The result of running the dependency analyses through the parser is a series of surprisal and retrieval values for each word in each of the sentences in each of the experiments. Each of the cognitive factors is translated into a probabilistic model. Surprisal values are then generated for these factors. Surprisal takes into account how the probabilities change as the sentence is parsed: if the probability space goes down once a word is parsed, like the head verb of a wh-extraction, it indicates an area of high surprisal, or high difficulty. If the probability space remains relatively equal or even goes higher, its indicates low surprisal and therefore low difficulty. Because the probability space is different for each of the cognitive factors, their surprisal values will also be different. The goal is to determine which cognitive factors make the parser sensitive to difficulty that is similar to the difficulty evident in the experimental data.

The other complexity metric is itself a cognitive hypothesis. Retrieval (Lewis & Vasishth, 2005) does not depend on the probability space of the parser; rather, it factors in the difficulty of retrieving an item from memory, like the extracted wh-element, and then attaching it to its

[1] See Chapter 4.8 for details.

head verb. The output is a time, in milliseconds, of how long it takes to process a word.

As has been discussed throughout this section, the focus of difficulty is the head verb of the extracted wh-element. Therefore, for each cognitive factor, the difficulty that it assigns to an experimental condition will be the surprisal or retrieval value it assigns to the head verb of the extracted wh-element. As can be seen in the diagram, the surprisal values can vary greatly across conditions.

This is not the only way to calculate surprisal and retrieval for the experimental conditions. For example, one could localize difficulty to the particular action of making the arc between the wh-element and its head verb. But, none of the experimental measures are nearly this local; in fact, reading time is the finest-grained experimental measure we consider, and it can only find difficulty at the word, at best. Therefore, the focus will be on difficulty at the verb, including any other difficulty associated with building its other dependents and processing it. Further details on the surprisal and retrieval implementations are available in Chapter 4.6.

6.3.4 ④ Transform into percentage difficulty by condition

The surprisal or retrieval scores for each of the conditions and each of the cognitive factors vary greatly, both across conditions (some factors assign lower surprisals than others) and across measures (surprisal values, retrieval values, and the various experimental measures have different scales). These differences can obscure the major experimental findings: syntactic locality violations are difficult. These numbers are therefore converted into percentages of difficulty for each experiment. This percentage is calculated by adding up the difficulty assigned by a particular cognitive factor to each of an experiment's conditions. This total difficulty is then divided by each condition's difficulty, giving what percentage of the total difficulty that cognitive factor assigns to that sentence. Taking the SBI data for the Alexopoulou and

Keller experiment, shown in ④ of Figure 6.4, the surprisal values show that SBI assigns 62% difficulty to the most difficult sentence, whereas it only assigns 1% difficulty to the baseline.

6.3.5 ⑤ Compare to human difficulty

Now that the percentage of difficulty that each of the cogntive factors assign to the experimental conditions is calculated, this quantity is compared to the percentage difficulty that humans assign to the experimental conditions. The same technique is used to get this percentage as for the cognitive factors: the difficulty across conditions is added to get the experiment total, and then each individual condition is divided from the total to get a percentage difficulty.

For acceptability and RRT some of the data is negative. For example, the two most difficult conditions in Alexopoulou & Keller have acceptability judgments of -0.05 and -0.14, repeated in Figure 6.5.

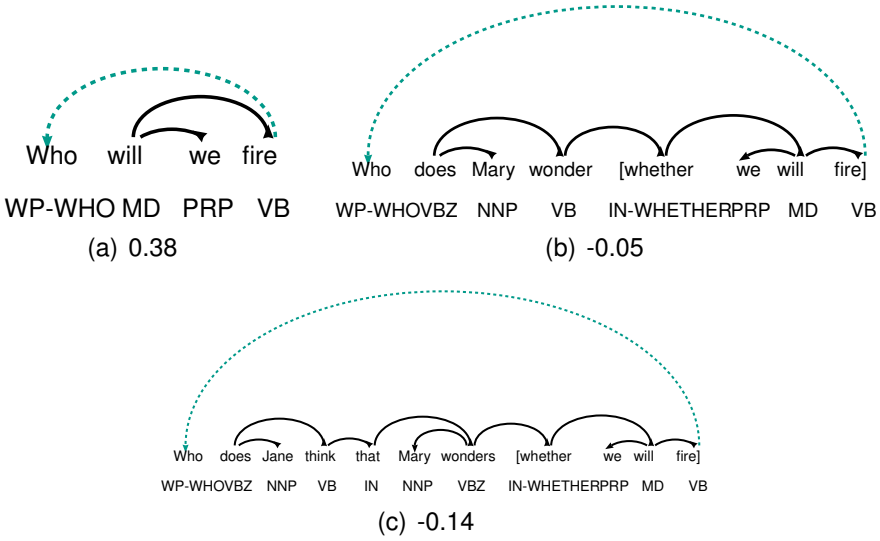


Figure 6.5: A WHI gradience example. Measure: acceptability.

One possible solution is to take the lowest number, in this case -0.14, and add its absolute value to all conditions. This makes the most difficult condition 0, and the best condition 0.52.

It is preferable to have the least difficult condition the lower number, and the most difficult conditions the higher number, to match the surprisal and retrieval predictions. Therefore, the highest number is added to the negative value of each of the conditions, as in Equation 6.1. Equations 6.2 through 6.4 demonstrate how this works for the Alexopoulou & Keller sentences. Now the data is arranged so that the most acceptable condition has the lowest percentage of difficulty, and the least acceptable condition the highest.

$$\text{Sentence difficulty} = -\text{sentence difficulty} + \text{experimental lowest value} \tag{6.1}$$

$$\begin{aligned} \text{Alexopoulou \& Keller Sentence 1} &= -0.38 + 0.38 \\ &= 0.0 \end{aligned} \tag{6.2}$$

$$\begin{aligned} \text{Alexopoulou \& Keller Sentence 2} &= -(-0.05) + 0.38 \\ &= 0.43 \end{aligned} \tag{6.3}$$

$$\begin{aligned} \text{Alexopoulou \& Keller Sentence 3} &= -(-0.14) + 0.38| \\ &= 0.52 \end{aligned} \tag{6.4}$$

One argument against this transformation is that it could be obscuring some aspect of the results. For example, the most acceptable condition will alway have 0% difficulty, even if this isn't the case given the acceptability judgments. However this is likely not a problem because the percentages are not directly compared. Instead, the order of the sentences, in terms of difficulty, is compared.

The graph shown after step (5), in larger scale in Figure 6.6, demonstrates this comparison. Percentage difficulty is provided along the x axis, and each of the measures are along the y axis, with the human difficulty on the top line, and the cognitive factors ordered by how well they predict the ordering. Each of the conditions are featured with a symbol along the x

axis. In this case, the easiest condition, Sentence 1, is represented by a blue circle. Sentence 3, the most difficult condition, is a yellow triangle. RETRIEVAL performs the best because it finds the most difference between the most difficult and least difficult conditions. In fact, for this data set, all of the cognitive factors get the correct order, and are therefore able to model the empirical difficulty ordering that was observed in the experiment. This is evident because the dashed line is at the bottom of the graph; normally, this dashed line appears somewhere in the middle, and separates those factors that get the correct ordering from those that do not.

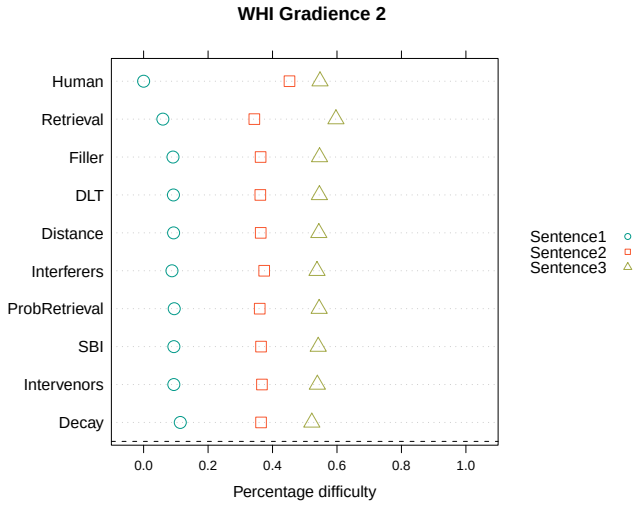


Figure 6.6: The result of the evaluation process is a graph comparing cognitive predictions to human difficulty.

One issue that comes up with determining a correct model by this ordering is that some cognitive factors could assign relatively equal percentage difficulty to conditions, but do so in the correct order. For example, in a syntactic judgment, the acceptable sentence is rated as 0% difficulty and the unacceptable sentence is rated as 100% difficulty. A cognitive factor could assign difficulty such that the acceptable sentence has 49% difficulty and the unacceptable sentence has 51% difficulty, and still get the correct ordering and be said to model the sentence. Although an argument can be made that this does not correctly model the sentence,

for the purposes here even these close calls are considered correct models. The reason is that it is unclear where to draw the line, and how this would relate to actual human difficulty. Given the coarse-grained syntactic judgment, it is difficult to assign a percentage difficulty that accurately reflects speaker intuitions. Is 60% difficulty okay? Is 40% difficulty not okay? Because this area is ambiguous, relative ordering is used. It will be noted in the results section when this issue comes up.

This evaluation procedure allows for a comparison of the difficulty that cognitive factors assign and the difficulty that humans assign to syntactic locality. Cognitive factors that assign difficulty in an order similar to the human order are considered accurate models. Cognitive factors that don't are not.

6.4 Conclusion

This chapter outlines a methodology for comparing a cognitive model to a variety of experimental measures for syntactic locality. The comparison is qualitative; there simply are not enough results in one study to allow for quantitative analysis, and the methods for quantitative analysis remain undefined. But, this evaluation procedure focuses on determining how to do a comparison given the current available methods. The next chapter discusses the results.

CHAPTER 7

RESULTS

7.1 Introduction

This chapter presents an analysis of cognitive theory predictions for syntactic locality on a working sentence processing model. Each phenomenon was tested against the cognitive theory predictions, which are encoded as probabilistic features within the dependency parser. The results demonstrate that cognitive factors model weak and non-island locality. However, they do not model the strong island data. This encourages a rethinking of syntactic locality, and the competence-performance divide.

The chapter is organized as follows: Section 7.2 discusses the parser's performance on each individual experiment. Section 7.3 aggregates the individual results by phenomenon to investigate differences between strong, weak, and non-island locality. Section 7.4 discusses the broad implications of these results, including insights on weaknesses and strengths of the methodology. Section 7.5 concludes.

7.2 Results by study

7.2.1 Introduction

As discussed in Chapter 6, a difficulty in modeling cross-linguistic data is the amount of experimental variation, in terms of methodology, practices, and even measures. This section details the parser's performance on the individual experiments, which will then be aggregated across experiments and phenomena in Section 7.3. The results are organized by phenomenon, with

discussion of CNP studies in Section 7.2.2, WHI studies in Section 7.2.3, and SUV studies in Section 7.2.4. Further details on the individual studies are provided in Chapter 5.

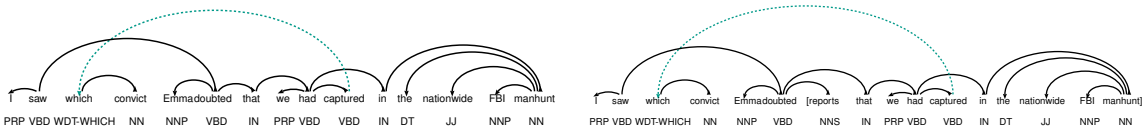
7.2.2 CNP

CNP strong islands offer little variation and gradience within English, and challenge reductionist accounts with evidence from Swedish. They represent the most difficult syntactic locality phenomenon for reductionist theories to explain, and this is further supported by the results reported here. Although some features can distinguish the classic island cases in English, the features perform poorly on the gradient and challenging data. This result does not support an account for syntactic locality based on these integration-based reductionist hypotheses, leading to the possibility that strong islands are indeed grammatical phenomena. Section 7.4 discusses the implications of this result. The following sections provide details on the parser's performance for classic, gradient, and challenging experiments.

Classic

Although there are many syntactic examples that demonstrate the difficulty of extraction from CNP islands in English, the classic example used here comes from an experiment by Hofmeister and Sag (2010), repeated in Figure 7.1. This experiment includes the most well-matched conditions for English CNPs. Here, the island-case in Figure 7.1(b) is only one word longer than the non-island, and all other words are the same[1]. Even though the conditions nearly match, the residual reading time (RRT) at *captured* is lower for the CNP island in Figure 7.1(b), indicating that the sentence is less acceptable than the base condition in Figure 7.1(a)[2].

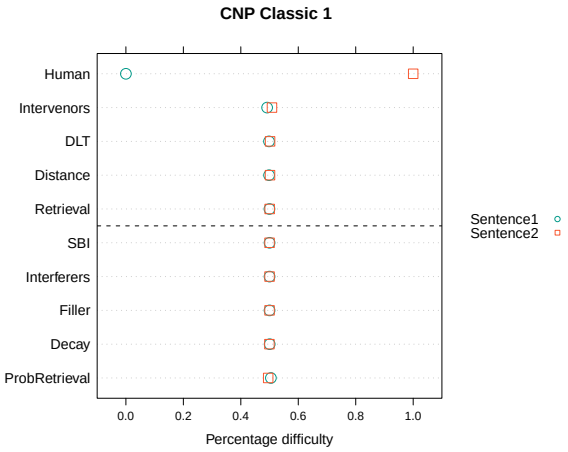
[1] See Chapter 4.7 for a full description of why dependency distance can sometimes affect modeling results.
[2] For more information on RRT and the other experimental measures, see Chapter 5.2.



(a) -68.72 (b) -62.10

Figure 7.1: CNP classic example 1: English (measure: RRT).

Several features do in fact get the distinction between the island and non-island condition, as shown in Figure 7.2(a). This graph shows the percentage of difficulty that a particular feature assigns to each of the sentential conditions in the experiment[3]. The top line shows the human result: in this case, the second sentence (corresponding to the second sentence in Figure 7.1) is more difficult, and is given a higher percentage of the difficulty. The following lines list each of the cognitive factors, ordered by degree of difference between the conditions. The dashed line separates those features that find Sentence 2 more difficult than Sentence 1 from those that do not.



(a)

Figure 7.2: Full CNP classic results.

[3]Further details on how the percentage difficulty is calculated are provided in Chapter 6.3.4.

The best-performing memory factor is Intervenors because it rates the difficulty for Sentence 2 higher than the other features. But, note that the difference between Sentence 1 and Sentence 2 is not large. The experimental difference in RRT between the two sentences is similarly not large, or at least not as large as one would expect for a grammatical distinction. Although this could follow from the experimental settings, the relatively similar difficulty rating, hovering at 50% for all factors, indicates that this experiment is not easily modeled by these cognitive factors. However, for the purposes of this work and as discussed in Chapter 6.3, any constraint that predicts the correct ordering (i.e., is listed above the line) is considered an accurate model of the experiment.

Considering only these cognitive constraints, it is not surprising that Intervenors and DLT gets this distinction. After all, the second sentence introduces another noun, or intervenor. Similarly, Distance gets the distinction because the extra noun makes the dependency length longer, leading to more difficulty in the second sentence. Retrieval only has a few milliseconds difference between the two conditions, but it is in the right direction. This follows from the activation equation for the retrieval theory.

The other features either do not distinguish between the two sentences or find the second sentence easier than the first. This may be because the sentences are too well-matched, or because the human acceptability differences are not pronounced enough. But, given the other experiments for CNPs considered here, the most likely cause is that strong islands are not easily modeled by these performance factors.

Although some cognitive constraints can model this classic example, this may not be because CNP island violations follow from the cognitive hypotheses themselves. Rather, this may be a result of the tendency for longer dependencies to be more difficult, as discussed in Chapter 4.7. There is a length difference between the conditions that is likely aiding the higher difficulty for Sentence 2. This explanation is further supported by the remaining results.

Gradience

The Hofmeister and Sag (2010) data discussed in the previous section is part of a larger experiment that tests whether CNP islands are subject to gradience. The experimental conditions are replicated in Figure 7.3, where conditions vary based on the filler type (bare or *which-Noun*) and the definiteness of the head noun (indefinite, definite, or plural). The findings indicate that there is a sensitivity to filler type, but not head-noun definiteness, with *which-N* type fillers being easier than bare fillers like *who*. Note that Figure 7.3(c) and Figure 7.3(e) are the classic examples from the previous section; compared to the other experimental conditions, they are not the easiest and most difficult conditions. However, the overall result indicates that there is gradience due to cognitive factors, which the authors use to support a reductionist account of CNP islands.

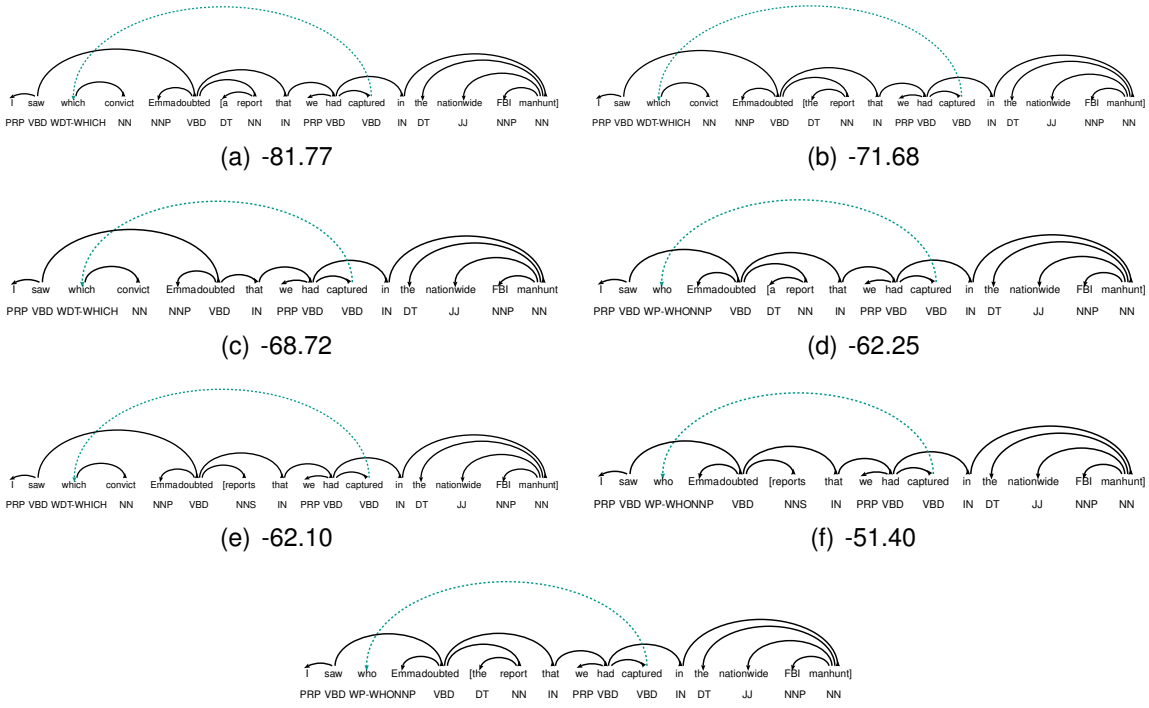
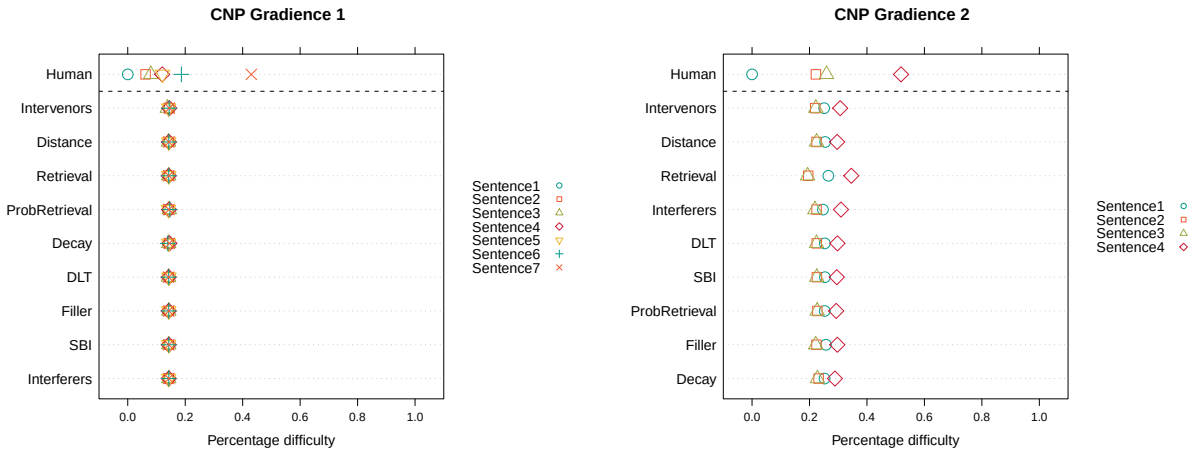


Figure 7.3: CNP gradience example 1: English (measure: RRT).

Figure 7.4(a), however, suggests otherwise. No cognitve factor models the human-like

acceptability patterns. Further, the seven conditions have a relatively even difficulty distribution across all factors, particularly when compared to the human data. This could be because the definiteness part of the experiment is obscuring the syntactic locality issue. To test this, a second experiment is considered.



(a) (b)

Figure 7.4: Full CNP gradience results.

This experiment, from Keller (1996), only considers filler-type gradience. The experiment provides acceptability ratings for various extracted filler-types from CNP islands, shown in Figure 7.5. Unlike Hofmeister and Sag, the *which-N* condition is considered less acceptable than the bare condition, Figure 7.5(a).

Unfortunately, the cognitive constraints now find the pattern from the Hofmeister and Sag experiment, and the bare condition is harder than the *which-N* condition in Figure 7.4(b). The cognitive constraints perform well on the other sentences: the *which-N* and *what* conditions, Sentences 2 and 3, have similar difficulty, and the *how* condition is hardest. However, the bare *who* condition (Sentence 1) is not ordered correctly, and the cognitive constraints are not able to model this particular experiment.

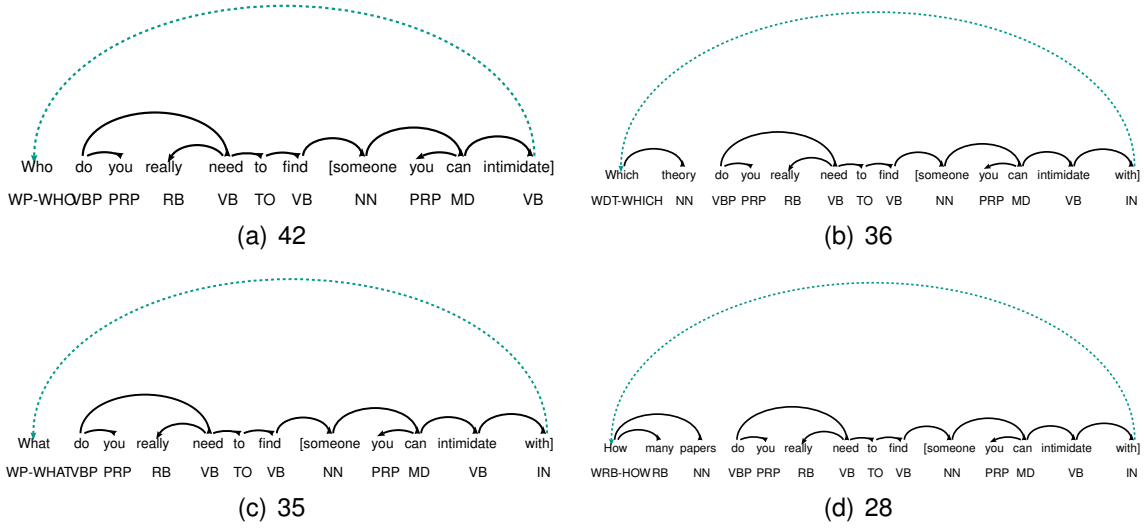


Figure 7.5: CNP gradience example 2: English (measure: acceptability).

There are two possible explanations for the different orderings of the which-N and *who* conditions in the experiments. One explanation is that the varying gradience is a result of different experimental measures: RRT is more sensitive than acceptability, and acceptability may be too coarse a measure to provide accurate results. In this case, the Hofmeister and Sag ordering should be considered more accurate, and it is possible that an RRT experiment of the Keller conditions would support the ordering from the cognitive constraints. However, it could also be the case that these sentences are difficult for English speakers, and any variation in acceptability does not reflect true gradience. This is further discussed in Section 7.4. For the time being, no cognitive factor can model either of the gradience experiments.

Challenges

The CNP gradience studies are not modeled by the cognitive constraints, indicating that an integration-based reductionist hypothesis for this type of syntactic locality is not possible. The following experiments, which are meant to challenge the cognitive constraints, provide further evidence against a reductionist explanation. The cognitive constraint do not model these

results either.

The first experiment, from Sprouse et al. (To Appear), provides acceptability ratings from an experiment that tests both distance and islandhood in a 2 x 2 design. The sentences are shown in Figure 7.6, with Figure 7.6(a) showing a short non-island, Figure 7.6(b) a short island, Figure 7.6(c) a long-distance non-island and Figure 7.6(d) a long-distance island.

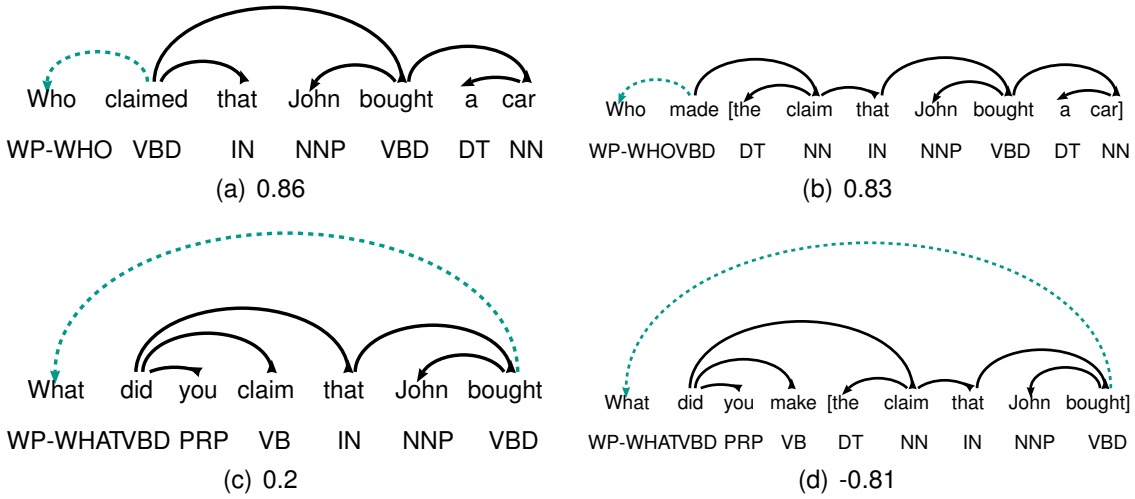


Figure 7.6: CNP challenge example 1: English (measure: acceptability).

Each of the features gets the distinction between the long-distance island and the long-distance non-island. However, it should be noted that this is not surprising since in terms of number of words, the second sentence is longer. The distinction between the short island and non-island is small but statistically-significant for English speakers, yet the parser does not find any difference. The cognitive constraints get the distinction between the short and long dependency lengths, but this falls out from the parser architecture.

The cognitive constraints also perform poorly on the other challenging experiments. Alexopoulou and Keller (2007) test how the level of embedding effects CNPs in English and German. The sentences are provided in Figure 7.8 and Figure 7.9 respectively. The acceptability judgments indicate that there is a difference between islands and non-islands in both languages. But, contra cognitive predictions, doubly-embedded island structures are more acceptable than

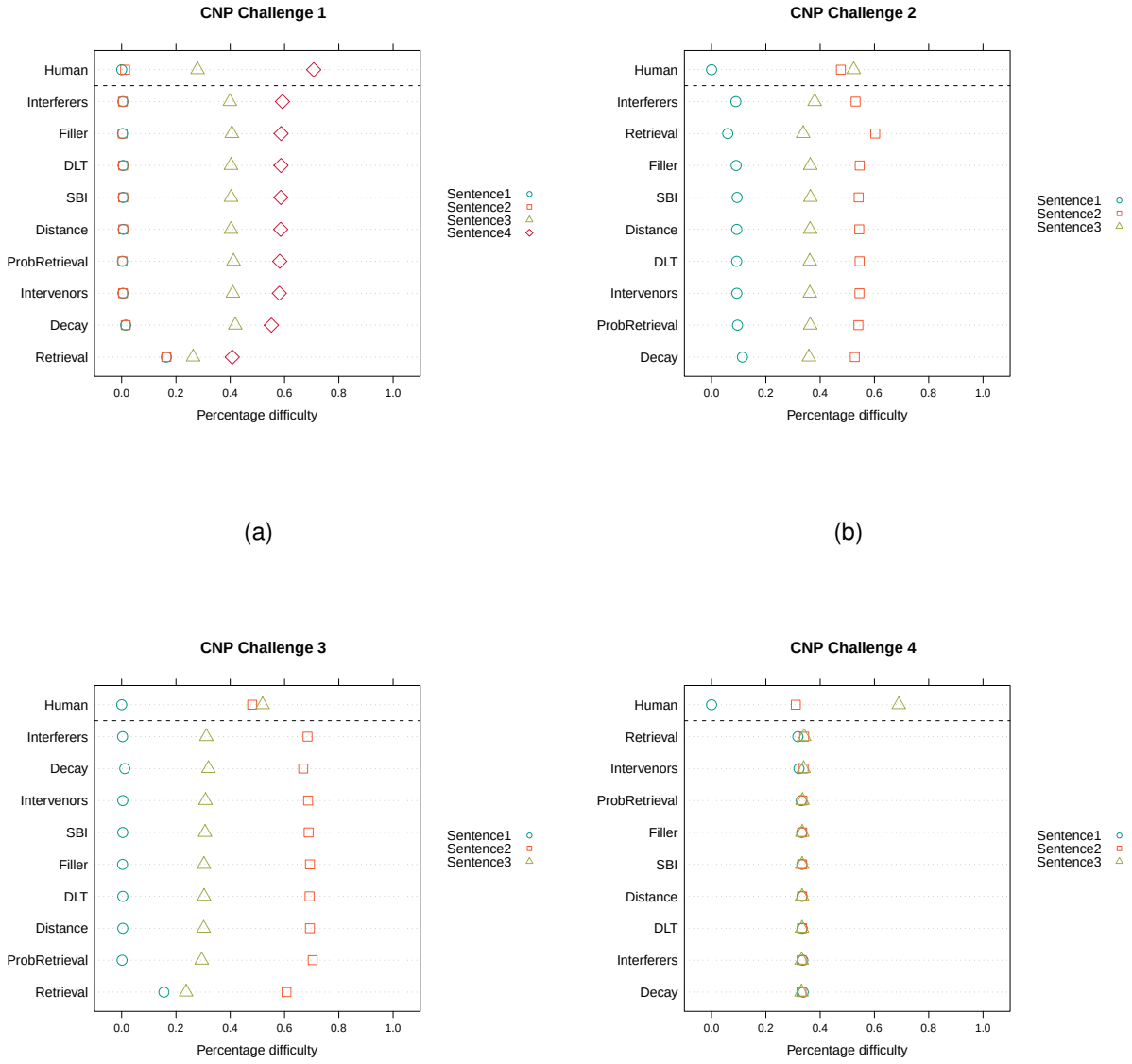


Figure 7.7: Full CNP challenge results.

singly-embedded islands.

As Figure 7.7(b) and Figure 7.7(c) demonstrate, the parsing model performs according to cognitive predictions. For all cognitive constraints, the doubly-embedded structure is more

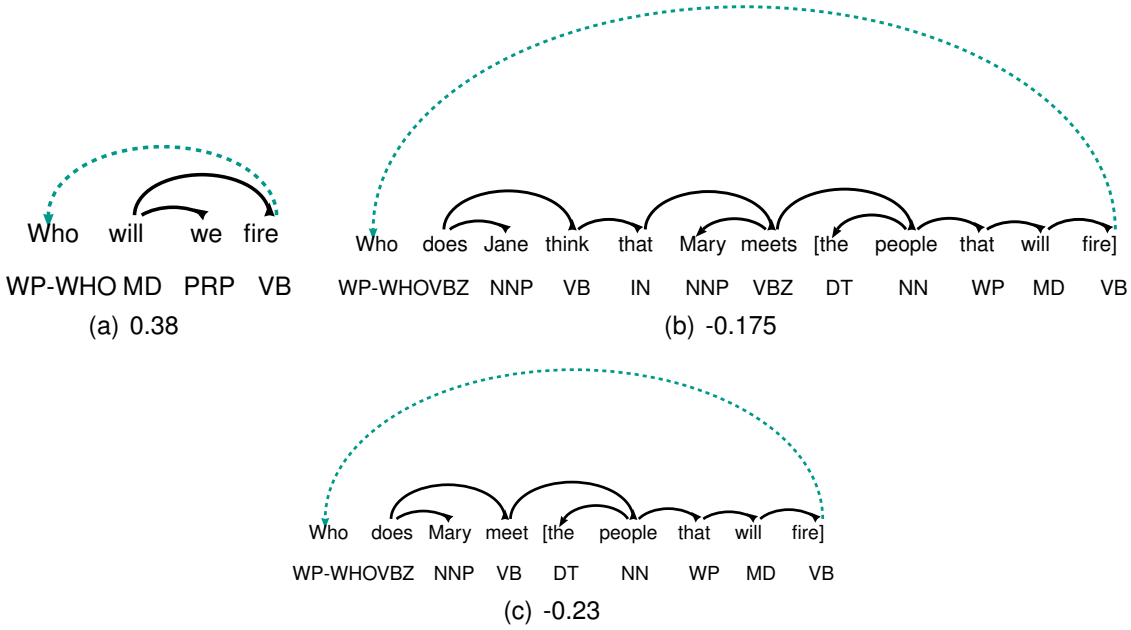


Figure 7.8: CNP challenge example 2: English (measure: acceptability).

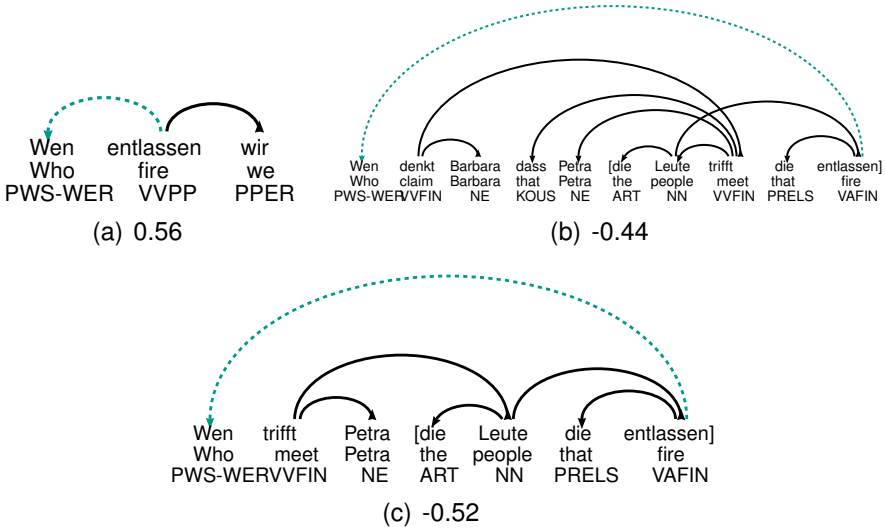
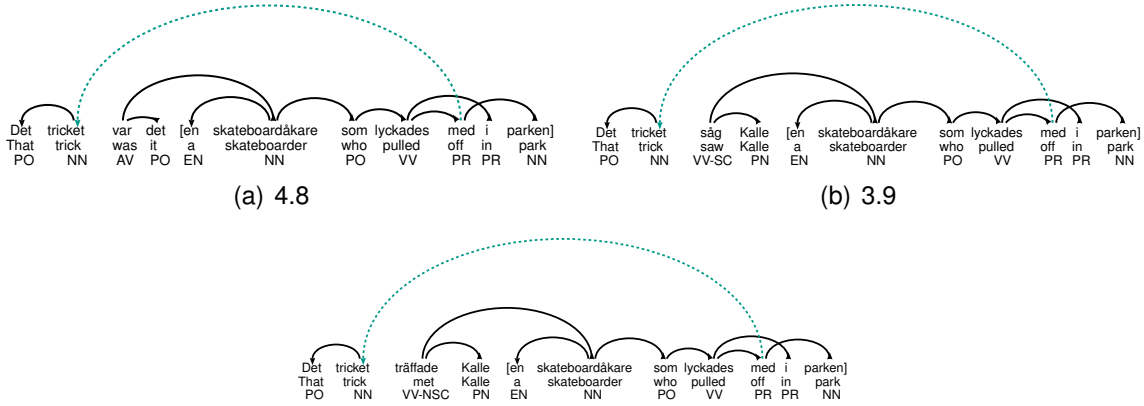


Figure 7.9: CNP challenge example 3: German (measure: acceptability).

difficult. Therefore, none of the constraints accurately model this data.

Data from Swedish has long caused problems for both grammatical and cognitive accounts of CNP islands because Swedish speakers do not find this extraction unacceptable. The Kush

and Lindahl (2011) study provides acceptability judgments demonstrating extraction from CNP islands is not as difficult as it is for English speakers. The sentences in Figure 7.10 show three cases: the first has extraction from a non-island, the second has extraction from an island, and the third has extraction from a complex-NP headed by a "non-small-clause" verb. Kush and Lindahl's grammatical explanation centers on this distinction: extraction from CNP islands is acceptable when the CNP is headed by a verb that takes a small clause. Their acceptability judgments support this claim, as Figure 7.10(b) has a high acceptability rating. However, CNPs headed by verbs that do not take a small-clauses do have lower acceptability, as in Figure 7.10(c).



(a) 4.8 (b) 3.9

(c) 2.8

Figure 7.10: CNP challenge example 4: Swedish (measure: acceptability).

The results from the parser can not provide evidence against Kush and Lindahl's grammatical argument. Figure 7.7(d) demonstrates that the parser not only finds a difference between the first and second cases, but it is unable to tell the difference between the small clause and non-small clause verbs. This is despite the fact that the treebank was modified to distinguish small-clause taking and non-small clause taking verbs[4]. Overall, this result indicates that the integration-based cognitive cognitive constraints can not handle the Swedish data.

The CNP island results demonstrate that although cognitive features can model the human

[4]See Chapter 4.4.1 for further details on treebank preparation.

pattern of difficulty in the classic case, this is likely driven by a word-number mismatch in the experimental conditions. Further, the difficulty values are so slight that they do not adequately model the well-documented patterns of unacceptability for these sentences. The inability of cognitive constraints to model the CNP island data is particularly noticeable when compared to their much better performance with WHIs, detailed in the next section.

7.2.3 WHI

The cognitive factors are able to consistently model a series of experiments for WHIs, including a range of classic, gradient, and challenging examples. The WHI experimental conditions are correctly matched in terms of words, so that any apparent modeling is not coming from a distance factor, as was the case in the last section. The classic case in Figure 7.11, provided by Hofmeister and Sag (2010), demonstrates. Reading time for the non-island condition, Figure 7.11(a), is lower than for the island condition in Figure 7.11(b).

(a) 231.79 (b) 255.96

Figure 7.11: WHI classic example 1: English (measure: RT).

Retrieval performs best on this experiment, which is in keeping with Hofmeister and Sag's argument that extraction from WHIs is difficult specifically because of a memory retrieval issue. Interferers and Filler also perform well; for Interferers, this is expected because *whether* interferes with the retrieval of *who*. For Filler, it is less understandable why it can distinguish between the two. It may be that the addition of *whether* provides some perturbation in the probability space, making the overall probability lower than it is for *that*.

As expected, the activation features perform poorly, as does DLT because it doesn't take

into account wh-intervenors like *whether*. It is surprising that the probabilistic ProbRetrieval feature does not perform well, although this could simply be due to some probabilistic anomolies rather than the theory itself. Further, SBI is not strong enough to get the distinction well–although the second sentence is harder, it's not harder by much and is more equal. Overall, however, the cognitive features model this experiment, as expected.

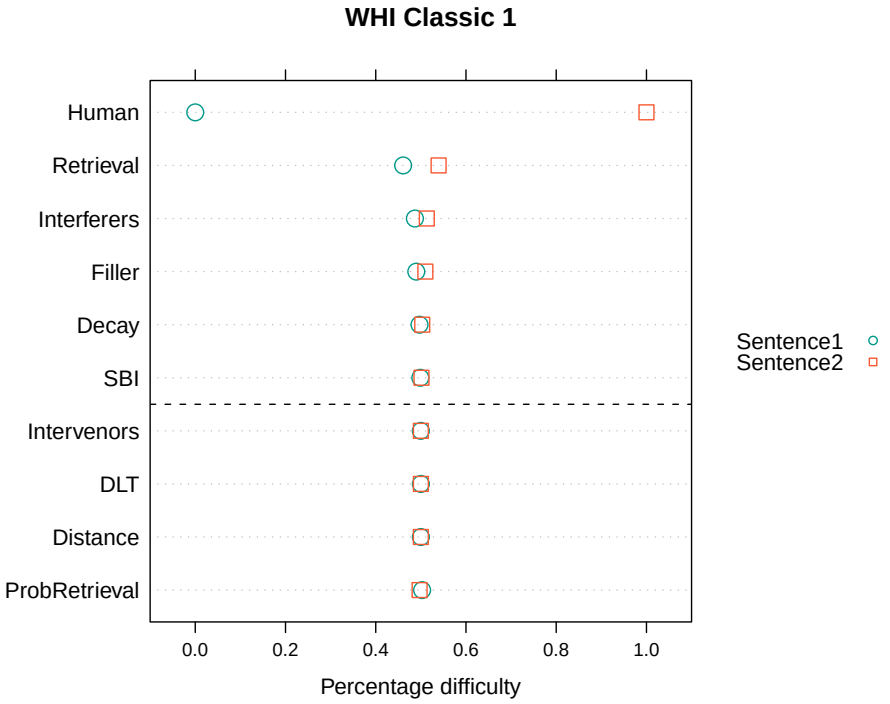
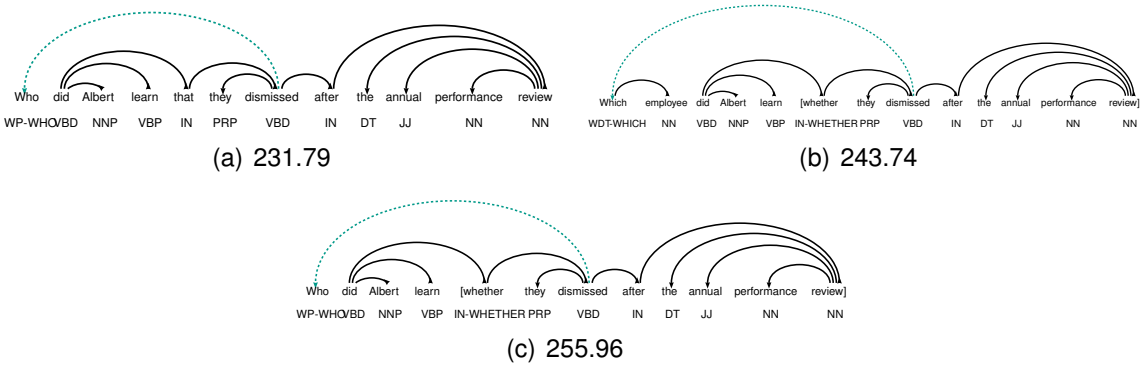


Figure 7.12: Full WHI classic results.

Gradience

Evidence that demonstrates gradience in WHIs is crucial for supporting cognitive explanations of weak islands. This gradience can be in terms of filler-type, as Hofmeister and Sag (2010)

demonstrate with their experiment in Figure 7.13. Here, the *which theory* example is easier to extract across the island than the bare *who*. The non-island baseline is easiest.



(a) 231.79 (b) 243.74

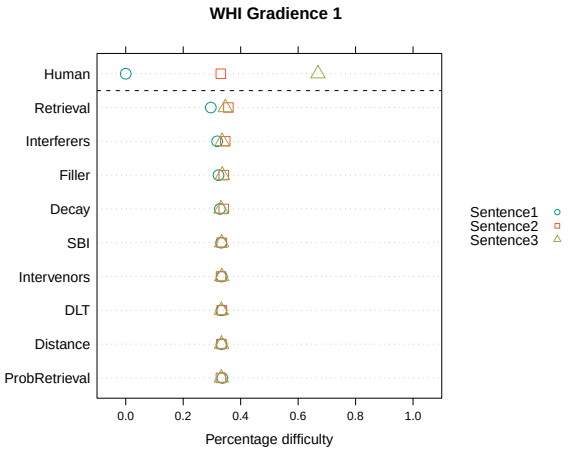
(c) 255.96

Figure 7.13: WHI gradience example 1: English (measure: RT).

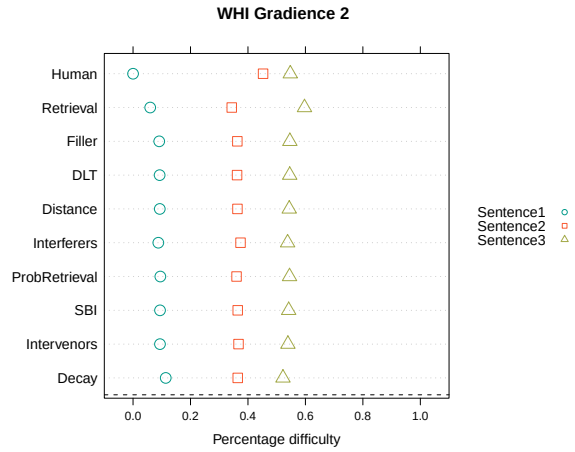
Unfortunately, the parser is unable to get this result. It finds the which-N condition harder than the bare condition for all of the cognitive theories. The baseline condition is easiest, as predicted. This problem may be a result of the tendency for the parser to find longer dependencies more difficult, which would be the case for the which-N conditions as they have an extra noun. This possibility is further discussed in relation to SUV gradience.

Other evidence for gradience in WHIs comes from Alexopoulou and Keller (2007). Like the CNP experiments, they test how embedding affects acceptability in English and German islands. In this case, the doubly embedded island is more difficult than the singly embedded island, which is in turn more difficult than the baseline non-island for both languages (Figure 7.15 and Figure 7.16).

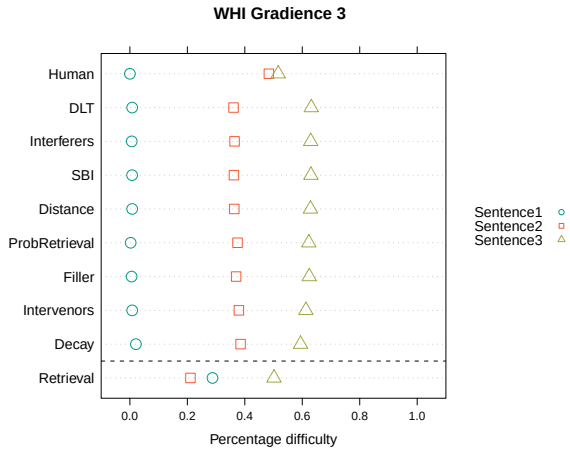
This result follows from cognitive explanations, and as expected, all of the factors perform well on the English and German examples. The doubly-embedded structure is hardest, and the baseline is easiest. In combination with the Alexopoulou and Keller result for CNPs, this result supports the authors' claims that reductionist accounts are better models of weak islands than strong islands. Further, it demonstrates that some gradience in WHIs can be modeled by cognitive constraints.



(a)



(b)



(c)

Figure 7.14: Full WHI gradience results.

Challenges

As with the CNP island example, the Sprouse et al. experiment for WHIs is a challenge for reductionist accounts. This challenge comes from the way that the experiment is constructed:

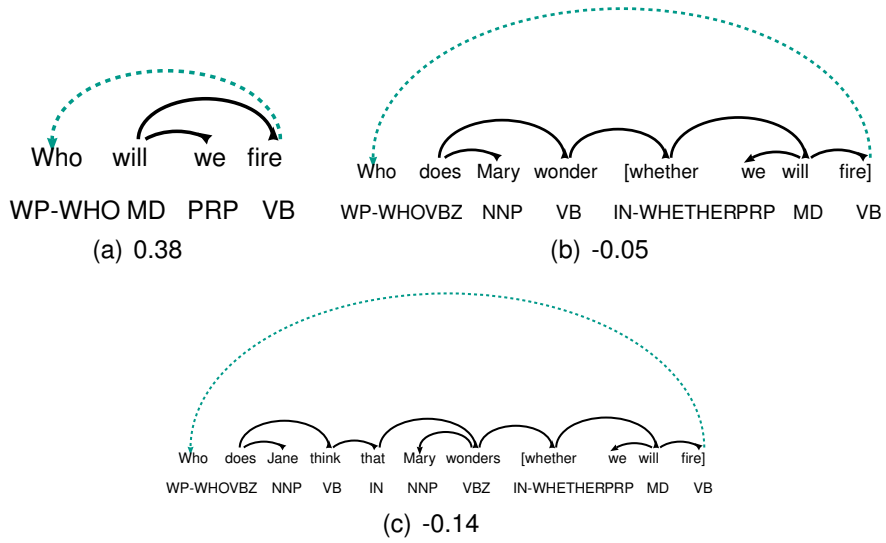


Figure 7.15: WHI gradience example 2: English (measure: acceptability).

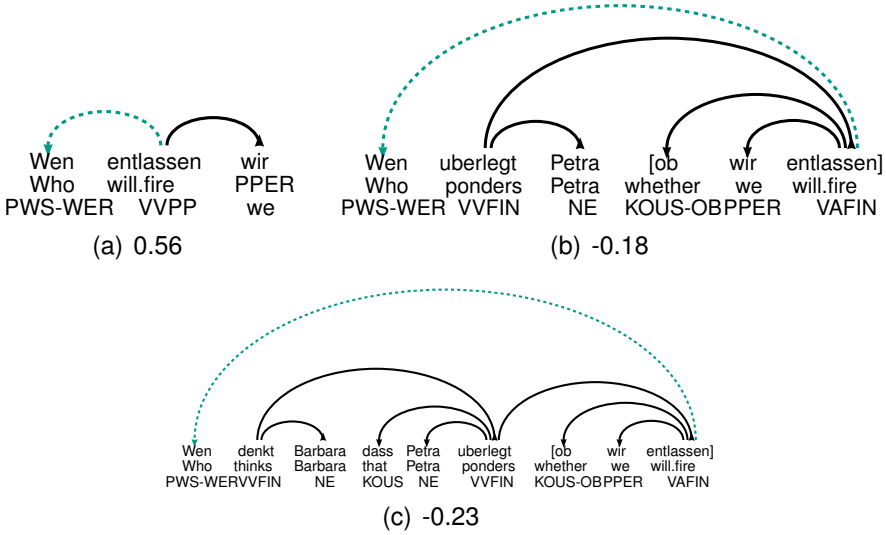


Figure 7.16: WHI gradience example 3: German (measure: acceptability).

it includes a distinction between short islands (Figure 7.17(b)) and short non-islands (Figure 7.17(b)) that can not be accounted for by integration-based theories. Although the parser is able to model the distinction between long islands and non-islands, shown in Figure 7.18(a), its inability to get this distinction demonstrates a potential problem with these reductionist hypotheses of WHIs. Although integration affects are sufficient for the majority of the results,

they can not model this distinction.

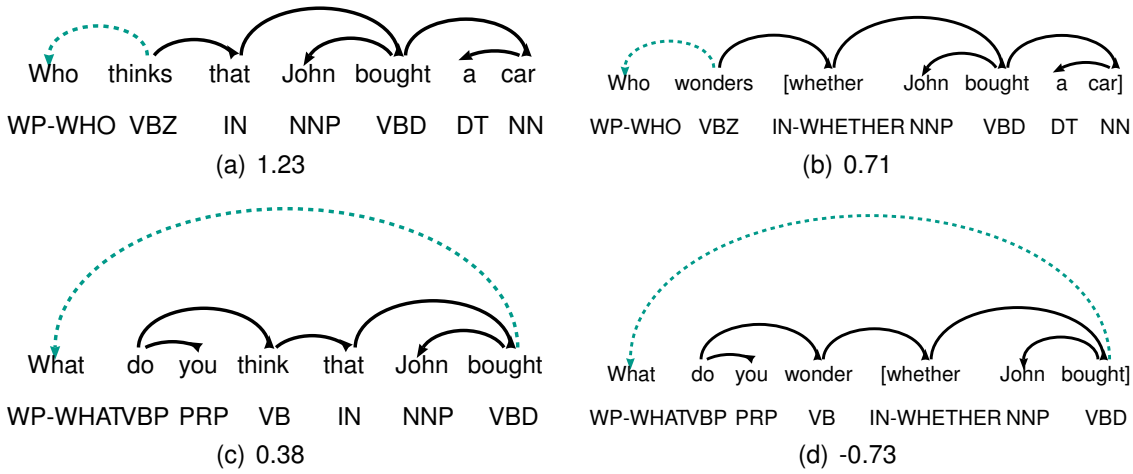
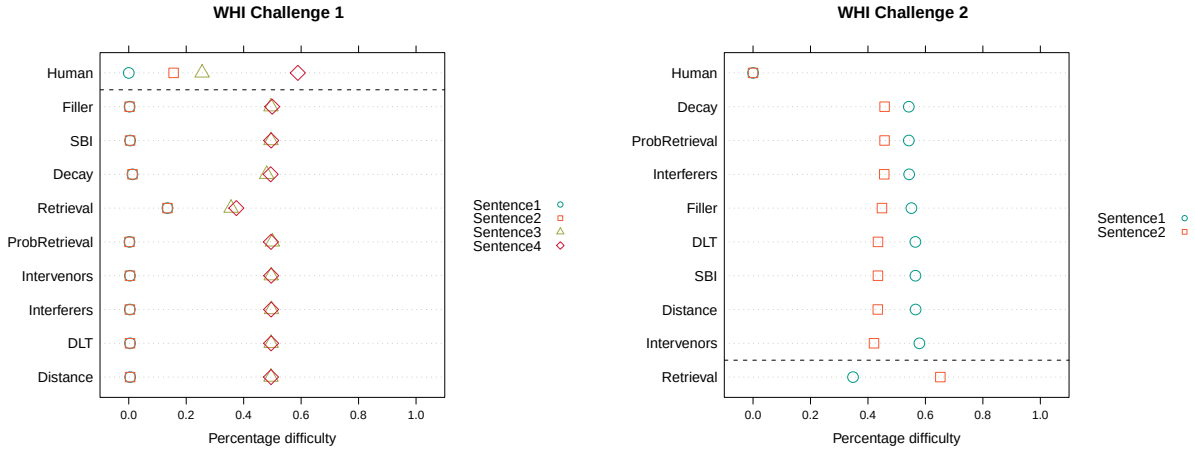


Figure 7.17: WHI challenge example 1: English (measure: acceptability).



(a) (b)

Figure 7.18: Full WHI challenge results.

Another piece of evidence that should cause problems for cognitive accounts is from Swedish. Like CNPs, Swedish seems to have relatively even acceptability judgments for extraction from island and non-island contexts, as shown in Figure 7.19. Here, Figure 7.19(b) is an island, but given syntactic judgments, it is perfectly acceptable. The reason this would

be a problem for cognitive accounts is because the same cognitive factors (interference, retrieval difficulty, activation distance) should be at play in Swedish as they are in English, and one would not want to argue that English and Swedish speakers have different memory constraints.

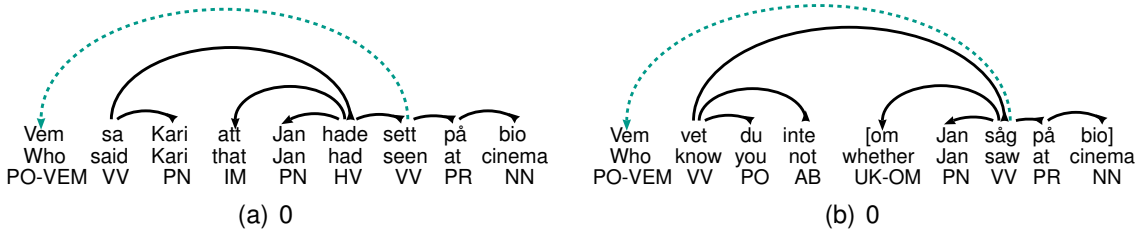


Figure 7.19: WHI challenge example 2: Swedish (measure: syntax).

Yet, one of the more suprising results is that the cognitive factors in fact find the Swedish island context *easier* than the non-island context, as shown in Figure 7.18(b). The only factor that does not is RETRIEVAL, which is likely due to the interference issue. However, in the cases of the probabilistic constraints, the interference appears to be over-ridden by frequency effects. This result indicates this Swedish example is either not an island, or there is something in the treebank frequencies that makes it easier to maintain the dependency across this island context.

These results demonstrate that cognitive factors model the classic, gradient, and challenging data. There are some anomolies, like the Hofmeister and Sag filler gradience experiment, but this could be explained by architectural issues. The cognitive factors are robust enough to model the Swedish experimental data, which does not directly follow from the cognitive hypotheses. This demonstrates the usefulness of an explicit computational model for these cases.

7.2.4 SUV

Unlike CNPs and WHIs, grammar-based accounts do not attribute SUV ungrammaticality to a phrase-marker barrier. Instead, the ungrammaticality is thought to arise from semantic or discourse issues (Szabolcsi & Zwarts, 1993). As discussed in Chapter 2.6, many of these theories incorporate aspects of interference without explicitly naming performance issues. Therefore, of the three constructs, cognitive theories should perform best on SUVs. In fact, factors that incorporate interference model the majority of the studies, whereas activation-based factors perform worse.

Classic

Consider the classic example, taken from Haider (2004) and displayed in Figure 7.20. Haider's syntactic judgments demonstrate that in the case where an intervening wh-word *who* replaces the pronoun *her*, the sentence becomes unacceptable.

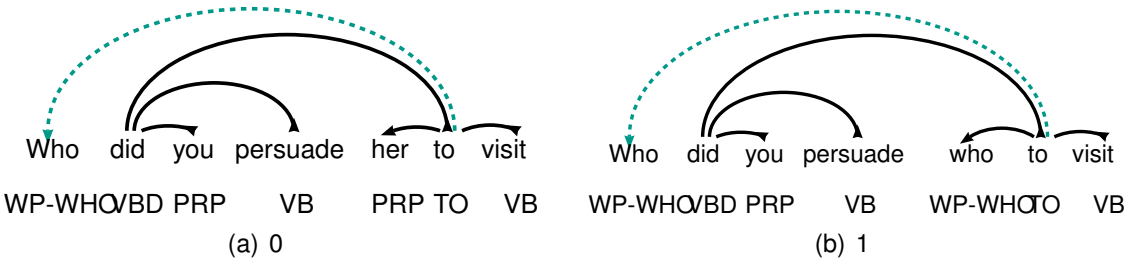
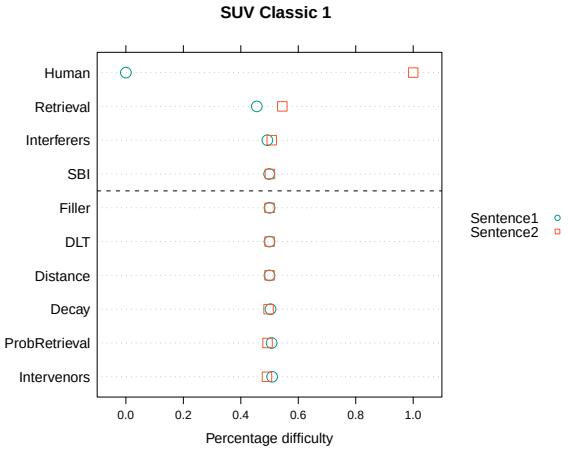


Figure 7.20: SUV classic example 1: English (measure: syntax).

The results, in Figure 7.21(a), demonstrate that RETRIEVAL, INTERFERERS, and SBI all find the second sentence more difficult than the first. The activation-based features, DISTANCE and DECAY, are not able to distinguish the two sentences, and some features, including probabilistic PROBRETRIEVAL, find the SUV to be easier than the non-SUV. It may seem surprising that the two retrieval factors make different predictions, but this is likely due to the addition of frequency

information to the ProbRetrieval constraint. These differences between the retrieval implementations are discussed further in Chapter 4.6.2.



(a)

Figure 7.21: Full SUV classic results.

Gradience

The Haider data is syntactic, and it only demonstrates that SUVs are more difficult than non-SUVs. One of the most interesting things about SUVs is the often gradient behavior they exhibit. Just as with CNPs and WHIs, experimental evidence from Arnon et al. (To Appear) suggests that informative fillers are more acceptable than bare fillers. Figure 7.22 shows the experimental conditions. Notice that in this experiment, the bare.bare condition (Figure 7.22(d)) is harder than the bare.which condition; both more informative fillers and intervenors are more acceptable than bare.

Figure 7.24(a) shows the cognitive predictions. The cognitive constraints have difficulty modeling the ordering due to dependency length: the which-noun phrases are longer, there-

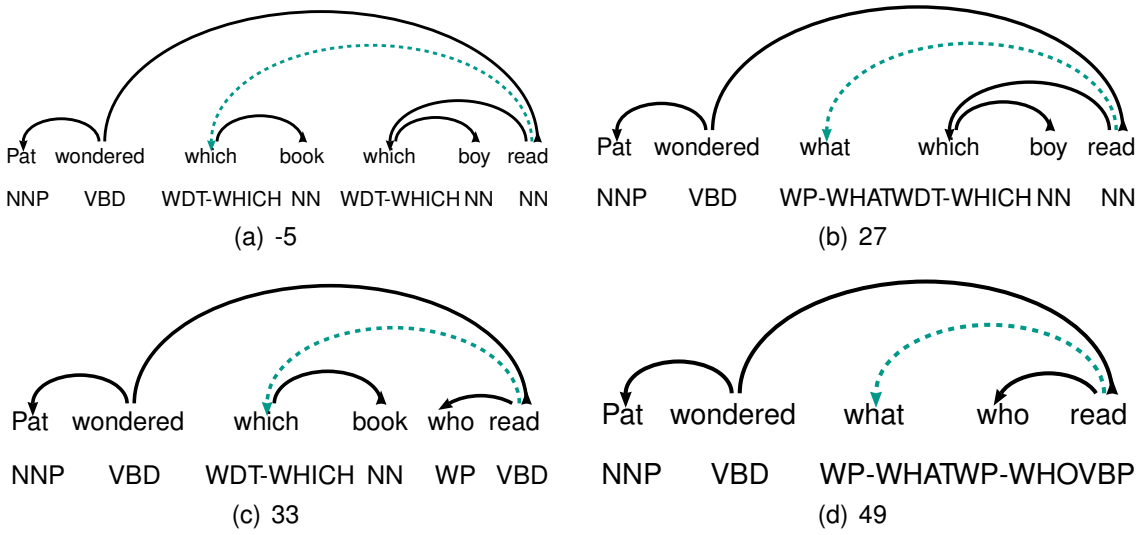


Figure 7.22: SUV gradience example 1: English (measure: RRT).

fore there will be more words between *which* and its verbal head. For many of these features, Sentence 1 is in fact the most difficult. A possible solution to this problem is to condense which-noun phrases into one word, so that *which device*, for example, would be headed by a single POS tag "WDT-WHICH-NOUN". But, this engineering fix could compromise the integrity of the cognitive model, and requires more testing. Further details on the link between dependency length and surprisal are available in Chapter 4. For now, this issue causes serious problems for a set of data that should be relatively easy for integration-based reductionist hypotheses.

The Fedorenko and Gibson (Submitted) experiment tests the same gradience as Arnon et al. (To Appear), but gets a different acceptability ordering. The sentences in Figure 7.23 show that *which patient* in Figure 7.23(a) and Figure 7.23(b) is more acceptable than *who* in Figure 7.23(c) and Figure 7.23(d).

Unfortunately, the cognitive theories are unable to replicate this ordering as well, as shown in Figure 7.24(b). Sentence 1, the condition with the longest dependency, is considered easier than the bare.bare and the bare.which conditions for all features in both experiments. Neither of these English gradience experiments is correctly modeled, and the results can not provide

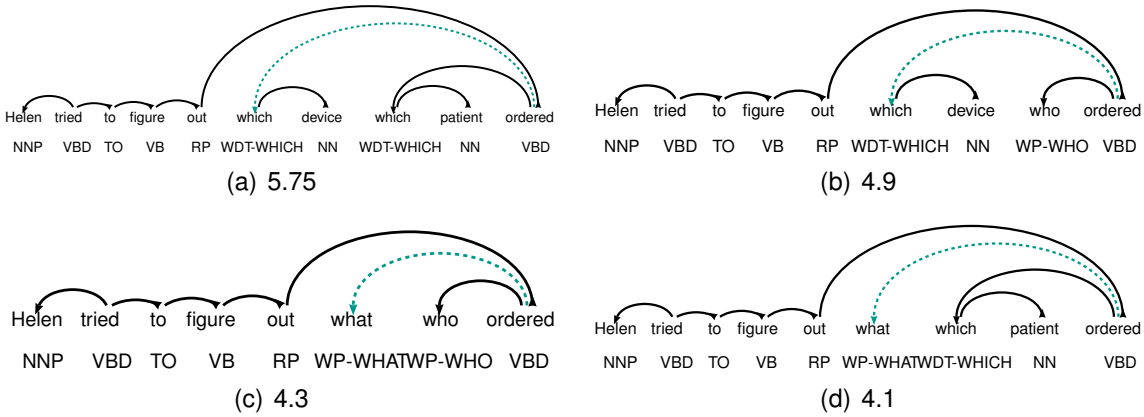
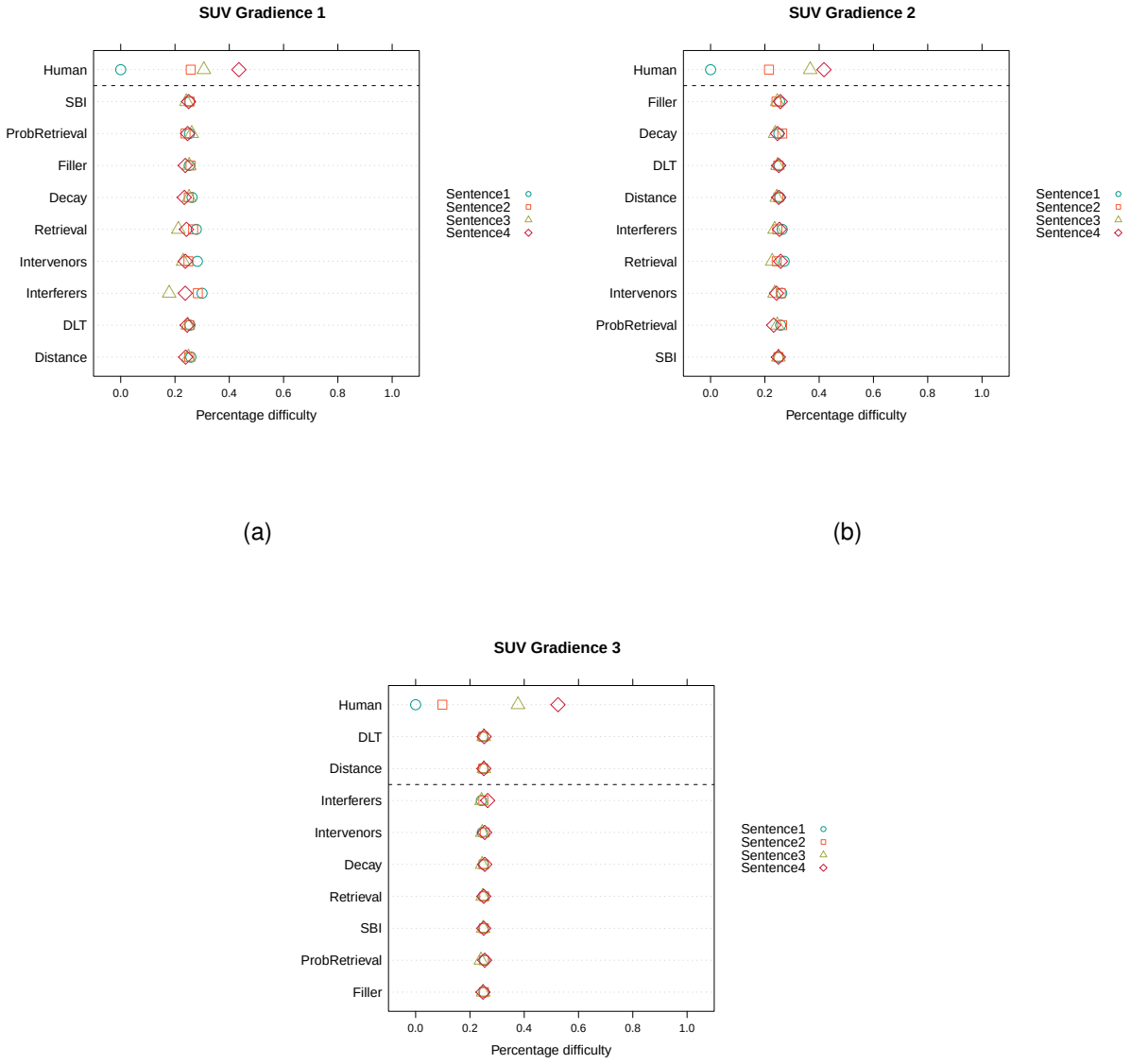


Figure 7.23: SUV gradience example 2: English (measure: acceptability).

insight into which ordering is best from the cognitive perspective. It should be noted, however, that this is caused by a shortcoming of the parser, and not of the cognitive constraints.

A final experiment that demonstrates SUV gradience is from Featherston (2005), who reports data demonstrating German gradience based on the informativeness of the filler (Figure 7.25). This experimental result, based on acceptability judgments, is controversial because it negates long-standing linguistic data suggesting that Germans do not find SUVs ungrammatical. However, this experiment not only found evidence for SUVs, but also found that SUV acceptability is highly variable depending on the case and the filler-type. For example, the least acceptable condition is when an accusative which-N filler is extracted past a nominative bare intervenor, Figure 7.25(h). The most acceptable condition is when a bare accusative is extracted past a nominative which-N intervenor. The reported acceptability ratings demonstrate that these differences are strong.

Like the English gradience data, the parser does not get the correct patterns here (Figure 7.24(c)). Unlike the English data, though, the parser does not find much variation among the conditions. For the features as well as the RETRIEVAL measure, these sentences have the same range of difficulty. A few features get the main pattern that the first condition is less difficult than the last condition, but no feature finds the correct pattern.



(a)

(b)

(c)

Figure 7.24: Full SUV gradience results.

Taken together, these results for gradience experiments are not promising for a cognitive account of SUVs. However, something curious happens with the next set of data points that requires a reconsideration of this view.

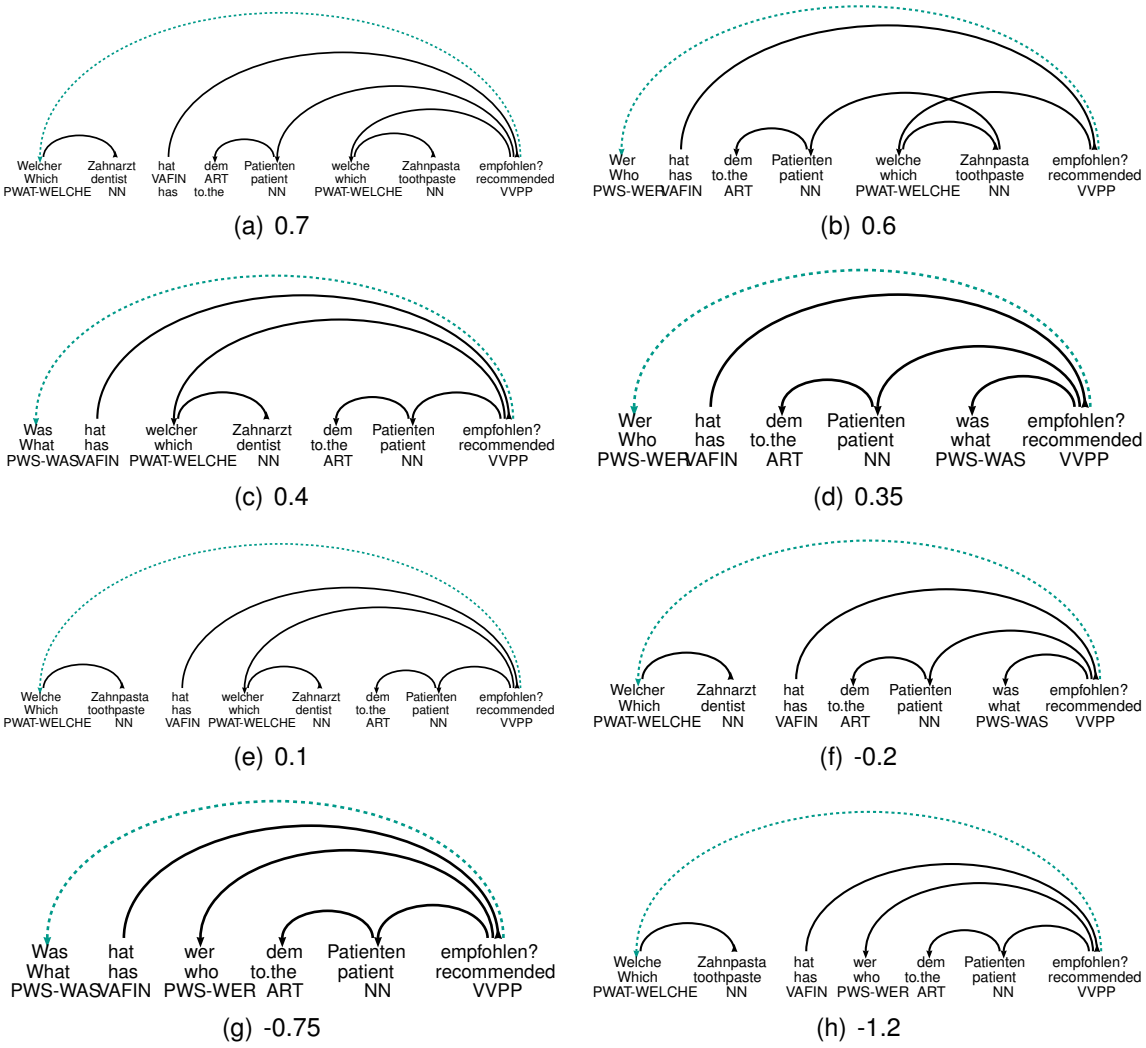


Figure 7.25: SUV gradience example 3: German (measure: acceptability).

Challenges

As discussed in the previous subsection, the Featherston (2005) experiment is considered controversial because it goes against long-standing evidence that German does not have SUVs. A typical example promoting this latter viewpoint is provided by Fanselow and Féry (2007), repeated below in Figure 7.26. Here, the SUV context in Figure 7.26(b) is considered of equal grammaticality as the non-SUV context.

Surprisingly, the cognitive factors support this result, and almost every factor finds the

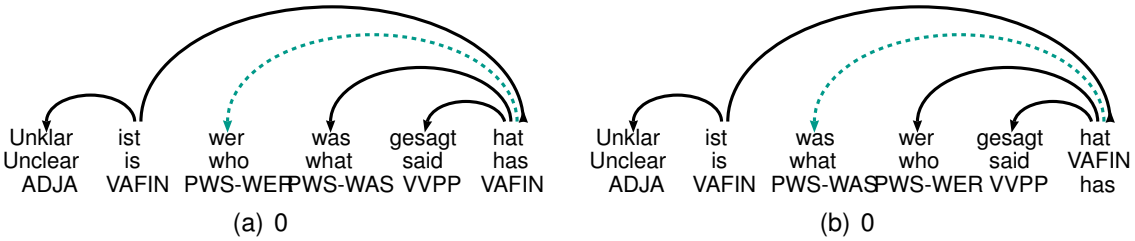


Figure 7.26: SUV challenge example 1: German (measure: syntax).

SUV context to be of equal or lesser difficulty to the non-SUV context (Figure 7.27(a)). Once again, the parser does not find much variation between SUVs and non-SUVs, supporting the linguistic position that there are no German SUVs. The only exception is INTERVENORS, which makes sense because there is an extra intervenor.

Fanselow and Féry (2007) suggest that the gradience in the Featherston (2005) experiment may be caused by case rather than actual SUV unacceptability and gradience. They design a 2 x 2 experiment wherein case and SUV are manipulated, but the conditions have equal dependency lengths. Their results demonstrate that when case is matched, the SUV conditions (Figure 7.28(a) and Figure 7.28(c)) are actually more acceptable than their case-matched counterparts (Figure 7.28(b) and Figure 7.28(d)) respectively.

And interestingly enough, this result is modeled by the SBI feature. Abstracting away from the slight acceptability differences between the SUV and non-SUV contexts, all of the features find the broad pattern of results for case. This overall pattern of results is interesting: the cognitive constraints should not be able to simultaneously predict difficulty for a language that has SUVs, and predict no difficulty for a language that does not. But they do.

This result is further substantiated by Russian, another language where SUVs are reportedly not difficult. Unlike the Featherston (2005) experiment, Fedorenko and Gibson (Submitted) find that filler informativeness does not make SUVs gradient in Russian. Figure 7.29 shows the experimental conditions and their acceptability judgments. Note that the SUV con-

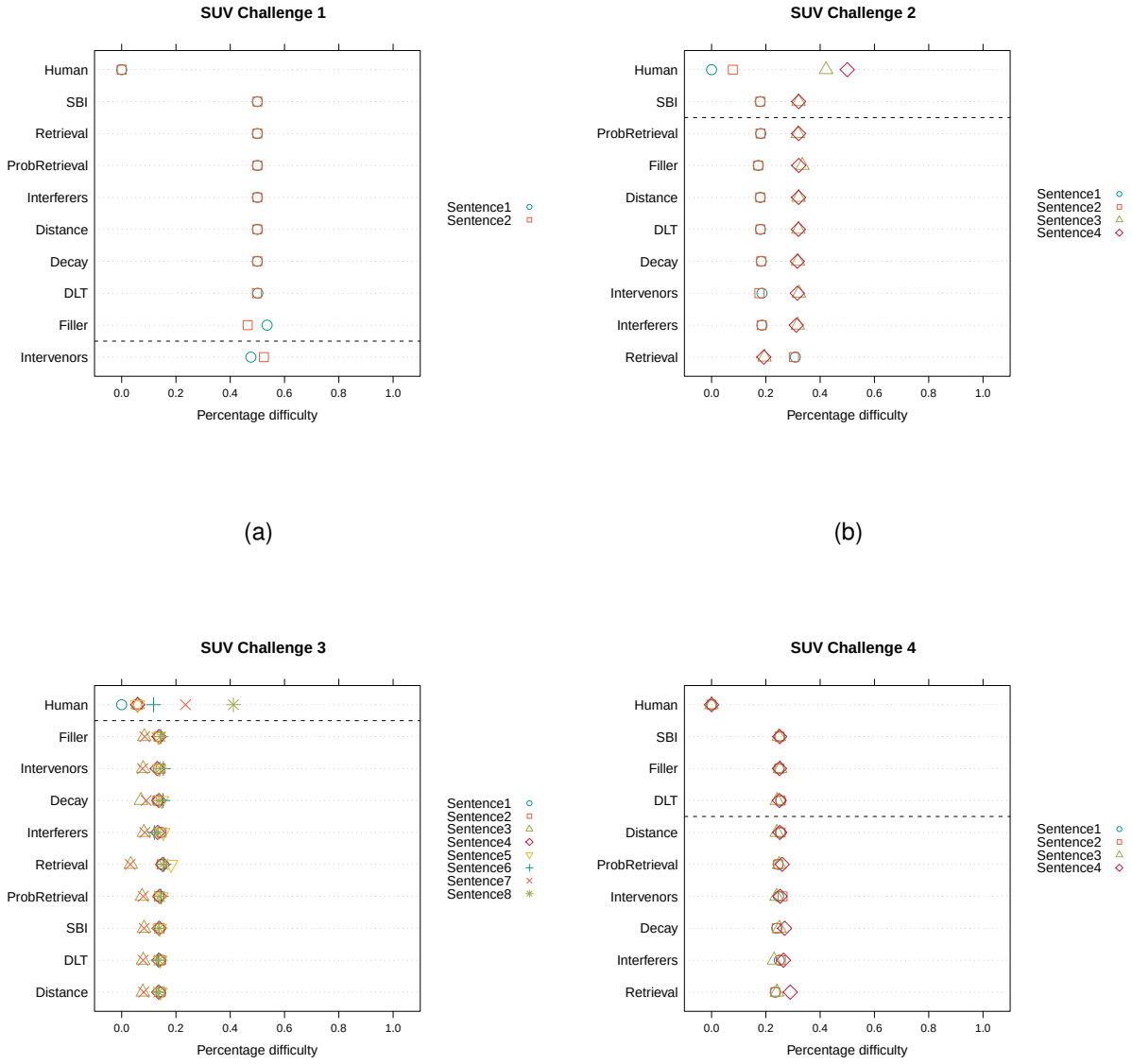


Figure 7.27: Full SUV challenge results.

texts (Figure 7.29(a), Figure 7.29(d), Figure 7.29(e), Figure 7.29(g)) are not the hardest, and that the acceptability judgments are relatively even.

Unfortunately, none of the cognitive constraints are able to exactly model these results

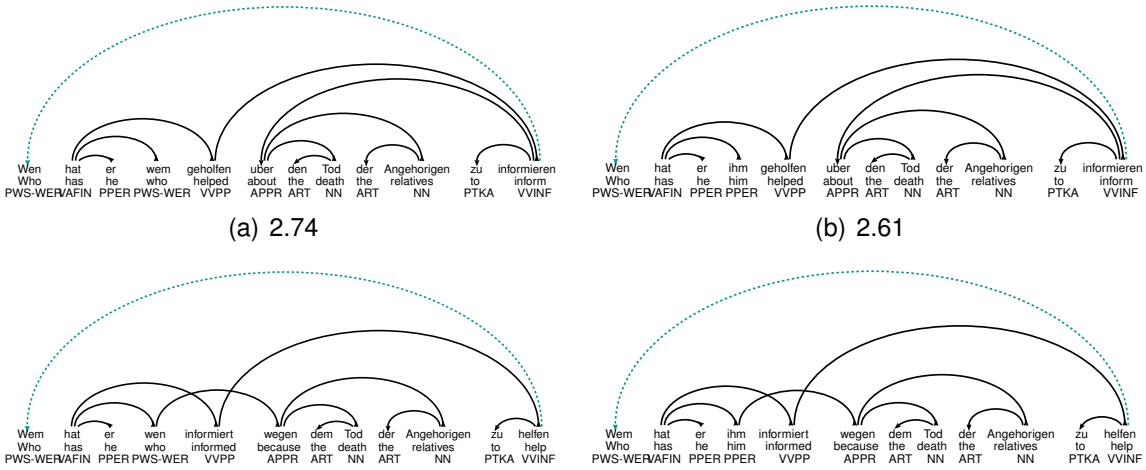


Figure 7.28: SUV challenge example 2: German (measure: acceptability).

(Figure 7.27(c)). But by abstracting away from the filler-type issue, it is possible to consider only the pairwise distinctions between SUVs and non-SUVs to see whether the cognitive constraints find SUVs to be difficult in Russian. This is shown in Figure 7.27(d). As with English and German SUVs, the SBI and FILLER features perform well. DLT does as well, but this could be due to the subject/object alternation also considered in the experiment.

The overall pattern of results for SUVs is not as clean-cut as it is for strong and weak islands. Whereas the memory theories performed poorly on the majority of the CNP sentences, and performed well on the majority of the WHI sentences, the memory theories only performed well on about half of the SUV experiments. Further, they performed worst on the experiments that they should have been doing the best on, those experiments that have been posited to support cognitive explanations of locality. However, this bad performance is explained by independent factors in the parser. Conversely, it may be these same independent factors that allow the cognitive theories to perform well on cross-linguistic data that has been used to argue against reductionist approaches. This topic, as well as further discussion of the comparison between the syntactic locality phenomena, is further discussed in the next section.

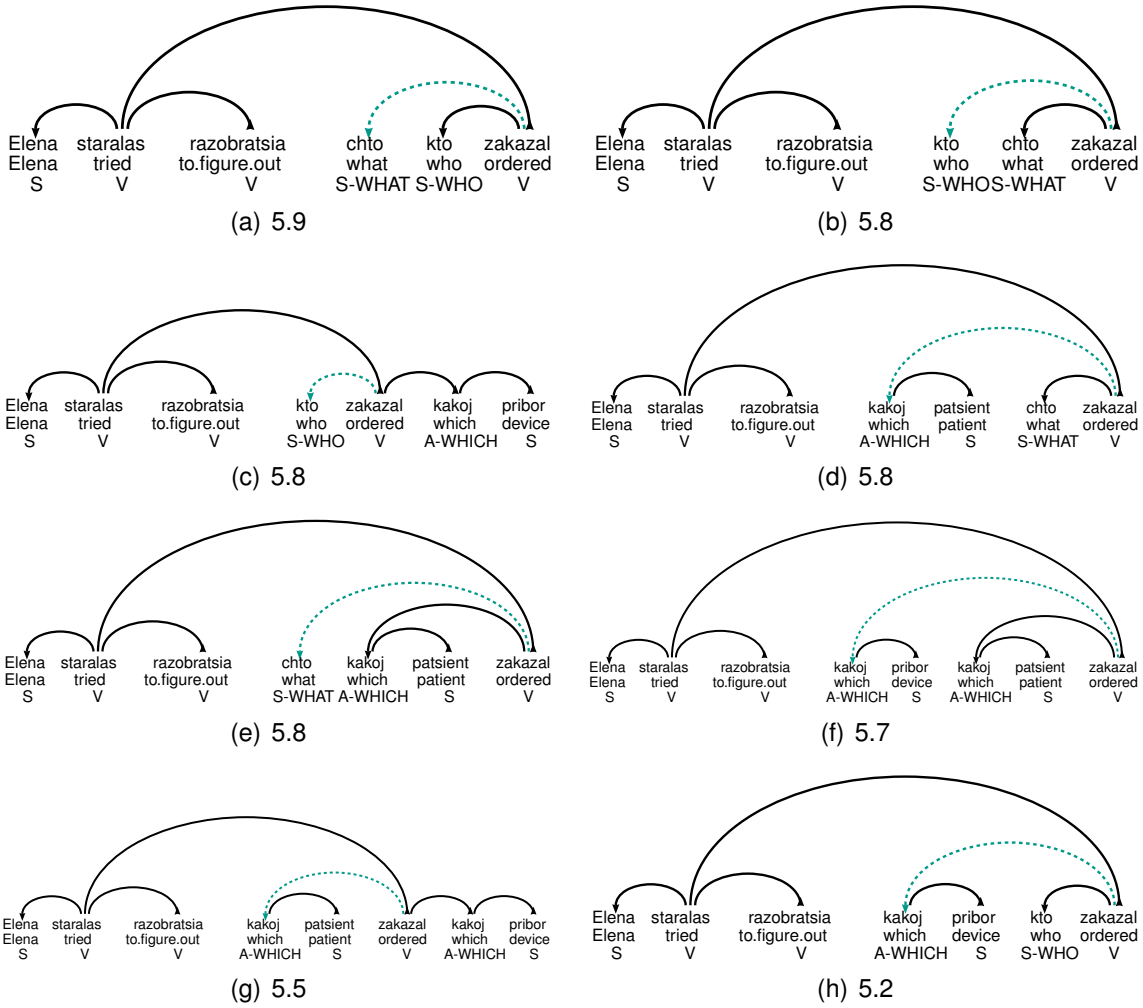


Figure 7.29: SUV challenge example 3: Russian (measure: acceptability).

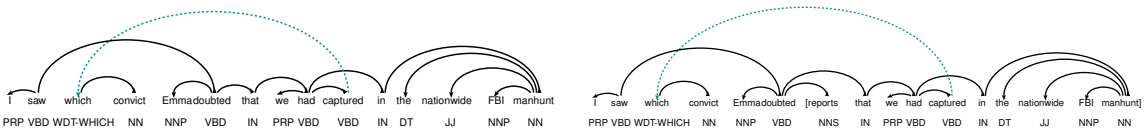
7.3 Results by phenomenon

To better understand how the cognitive constraints could explain the syntactic locality phenomena, it is helpful to take a step back and consider the broader results. Section 7.3.1 details CNPs, Section 7.3.2 details WHIs, and Section 7.3.3 details SUVs.

7.3.1 CNP

Table 7.1 provides a summary of the CNP study results ordered by cognitive constraint. Full circles represent experiments that are correctly modeled whereas empty circles represent experiments that are incorrectly modeled. The table demonstrates the relatively poor performance of the cognitive constraints on the CNP experiments. The cognitive factors only perform well on the classic cases, and as discussed in Section 7.2.2, these results do not indicate that the constraints explain the data. Rather, this result is likely because of the tendency within the parser for longer dependency lengths to have more difficulty.

In the classic experiment, replicated in Figure 7.30, the dependency length in the island is greater than in the non-island condition. Although features can override the tendency for longer dependencies to be harder (see Section 7.2.4), this is not common. It is therefore likely that the classic results are an effect of dependency length, particularly when compared to the other CNP results.



(a) -68.72 (b) -62.10

Figure 7.30: Dependency length is greater in CNP-violating contexts.

The gradience experiments from Hofmeister and Sag (2010) and Keller (1996) have length distinctions that should also work in the parser's favor. But the cognitive constraints perform poorly; many do not find the longest condition to be the most difficult, even though this is the case for the experimental data. The length-difficulty tendency was over-ridden, indicating a particularly strong prediction that goes against human difficulty.

It is not surprising, then, that the cognitive constraints also perform poorly on the challenging experiments. These are the experiments that demonstrate patterns that go against

Study	Distance	Decay	Filler	Intervenors	Interferers	SBI	ProbRetrieval	DLT	Retrieval
Classic 1: English	●	○	○	●	○	○	○	●	●
Gradience 1: English	○	○	○	○	○	○	○	○	○
Gradience 2: English	○	○	○	○	○	○	○	○	○
Challenge 1: English	○	○	○	○	○	○	○	○	○
Challenge 2: English	○	○	○	○	○	○	○	○	○
Challenge 3: German	○	○	○	○	○	○	○	○	○
Challenge 4: Swedish	○	○	○	○	○	○	○	○	○

Table 7.1: CNP Summary Table.

cognitive predictions. This includes the Alexopoulou and Keller (2007) data where doubly-embedded structures are easier than singly-embedded structures in English and German, the Sprouse et al. (To Appear) data that demonstrates a statistically-significant effect of structure and dependency length for CNP islands, and the Swedish data that demonstrates that extraction from CNP islands is not difficult for Swedish speakers. In each of these cases, the cognitive constraints behave as one would expect: they model difficulty for length, embedding, and extraction contra the experimental evidence. They are simply unable to handle this range of data.

By and large, the CNP results demonstrate that cognitive constraints are hard-pressed to explain the pattern of results from experimental data. However, they fare better for WHIs, as discussed in the next section.

7.3.2 WHI

There are fewer useable experimental results for WHIs, likely because WHIs tend to behave relatively uniformly and follow from cognitive predictions. For this latter reason, it is not surprising that the cognitive constraints perform well across the range of classic, gradient, and challenging data. The summary table, in Table 7.2, highlights some interesting exceptions.

Surprisingly, the classic case is not strongly modeled across phenomena, although a range of phenomena get it. There is both a Decay and SBI element in the classic case, and although the probabilistic ProbRetrieval performs poorly, the standard Retrieval performs well. Filler, an interference feature based on what kind of filler is being held in memory, also performs well. In fact, it is one of the best features for this data set.

This pattern, wherein activation, interference, and composite features perform well on the data set, extends through the gradience and challenging examples as well. Alexopoulou and

Study	Distance	Decay	Filler	Intervenors	Interferers	SBI	ProbRetrieval	DLT	Retrieval
Classic 1:English	○	●	●	○	●	●	○	○	●
Gradience 1: English	○	○	○	○	○	○	○	○	○
Gradience 2: English	●	●	●	●	●	●	●	●	●
Gradience 3: German	●	●	●	●	●	●	●	●	○
Challenge 1: English	○	○	○	○	○	○	○	○	○
Challenge 2: Swedish	●	●	●	●	●	●	●	●	○

Table 7.2: WHI Summary Table.

Keller (2007) test gradience with respect to embedding, and their experiments are modeled by nearly all the cognitive constraints. The Hofmeister and Sag (2010) and Sprouse et al. (To Appear) data is not easily modeled; however, as discussed in the previous sections, this has to do with experimental issues that can not be replicated by this methodology. A more suprising result comes from Swedish: in this case, the island example, which is matched for dependency length but adds an extra intervenor, is considered of equal difficulty to the non-island case. Although the extra intervenor should be harder, each of the probabilistic features finds the correct pattern. RETRIEVAL does not: the extra intervenor causes more difficulty.

This result could be problematic because it appears to go against the cognitive predictions. However, there is an explanation for this behavior considering the structure in the Swedish sentences. Figure 7.31 shows the dependency analysis for the two sentences. Note that although the dependency length is the same across the conditions, the dependency structure between the two words is markedly different. There are fewer parser actions required to make the dependency in the island than in the non-island, leading to an overall decrease in difficulty. Therefore, the cognitive theories are behaving as predicted given this dependency analysis. The relative ease of processing for the WHI in Swedish may be a result of the sentence structure itself rather than abnormal island behavior. More experimental data is needed to better understand Swedish WHIs, but for now they do not pose a problem for a cognitive explanation of weak island phenomena.

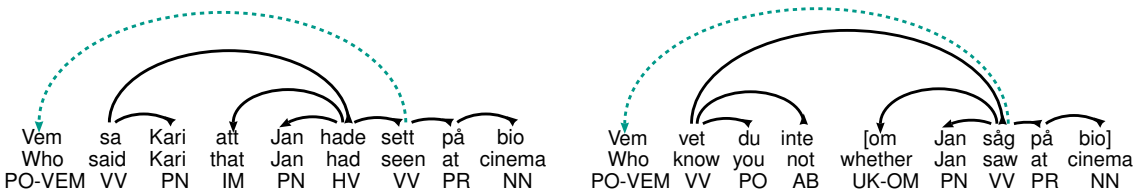


Figure 7.31: The dependency structure is different for WHI-violating Swedish sentences.

7.3.3 SUV

The SUV results are between the strong and weak island results: whereas strong islands are not modeled by cognitive theories, and weak islands are, SUVs are mostly modeled by *some* of the memory theories. Table 7.3 summarizes these results.

The classic case for SUVs, in the Haider (2004) experiment, is modeled by the interference-based features INTERFERERS and SBI, as well as RETRIEVAL, which includes an interference component. For the challenging data, which includes examples of German and Russian SUVs that do not incur difficulty, interference-based features handle the majority of the data. The only data not modeled comes from the gradience examples.

There is a simple explanation for this: experimental evidence for the filler-type gradience, which each of these studies tests, goes against the difficulty-length tendency. In these cases, *which-N* type fillers are easier to process than bare fillers. However, the *which-N* fillers include an extra word (the noun), which requires an additional parser action and brings down the overall difficulty. This is discussed more fully in Chapter 4 and Section 7.2.4. That this particular implementation is unable to model this sequence of results, particularly since it is so prevalent an issue across syntactic locality phenomena is a problem; however, getting around it on this architecture could compromise the overall model. It is therefore likely that the cognitive theories can handle this data, but it remains to be proven.

The cognitive constraints perform better on SUVs than they do on CNPs. They do not perform as well as they do on WHIs, both in terms of number of features as well as experiments. This may be due to the experimental evidence that is considered here. But, it may also shed light on the differences between weak islands and SUVs: whereas weak island difficulty has an interference and an activation component, SUV difficulty may simply have an interference component, as discussed in the next section.

Study	Distance	Decay	Filler	Intervenors	Interferers	SBI	ProbRetrieval	DLT	Retrieval
Classic 1: English	○	○	○	○	●	●	○	○	●
Gradience 1: English	○	○	○	○	○	○	○	○	○
Gradience 2: English	○	○	○	○	○	○	○	○	○
Gradience 3: German	●	○	○	○	○	○	○	●	○
Challenge 1: German	●	●	●	○	●	●	●	●	●
Challenge 2: German	○	○	○	○	○	●	○	○	○
Challenge 3: Russian	○	○	○	○	○	○	○	○	○
Challenge 4: Russian	○	○	●	○	○	●	○	●	○

Table 7.3: SUV Summary Table.

Finally, the results indicate a redrawing of the organizational chart for experimental studies provided in Chapter 5. The new chart is in Figure 7.32. Bolded boxes indicate experiments that can now be used to support reductionist claims, whereas dashed boxes indicate experiments that no longer support reductionist claims. Overall, it is evident that WHI and SUV experiments support reductionist claims, but CNP experiments do not.

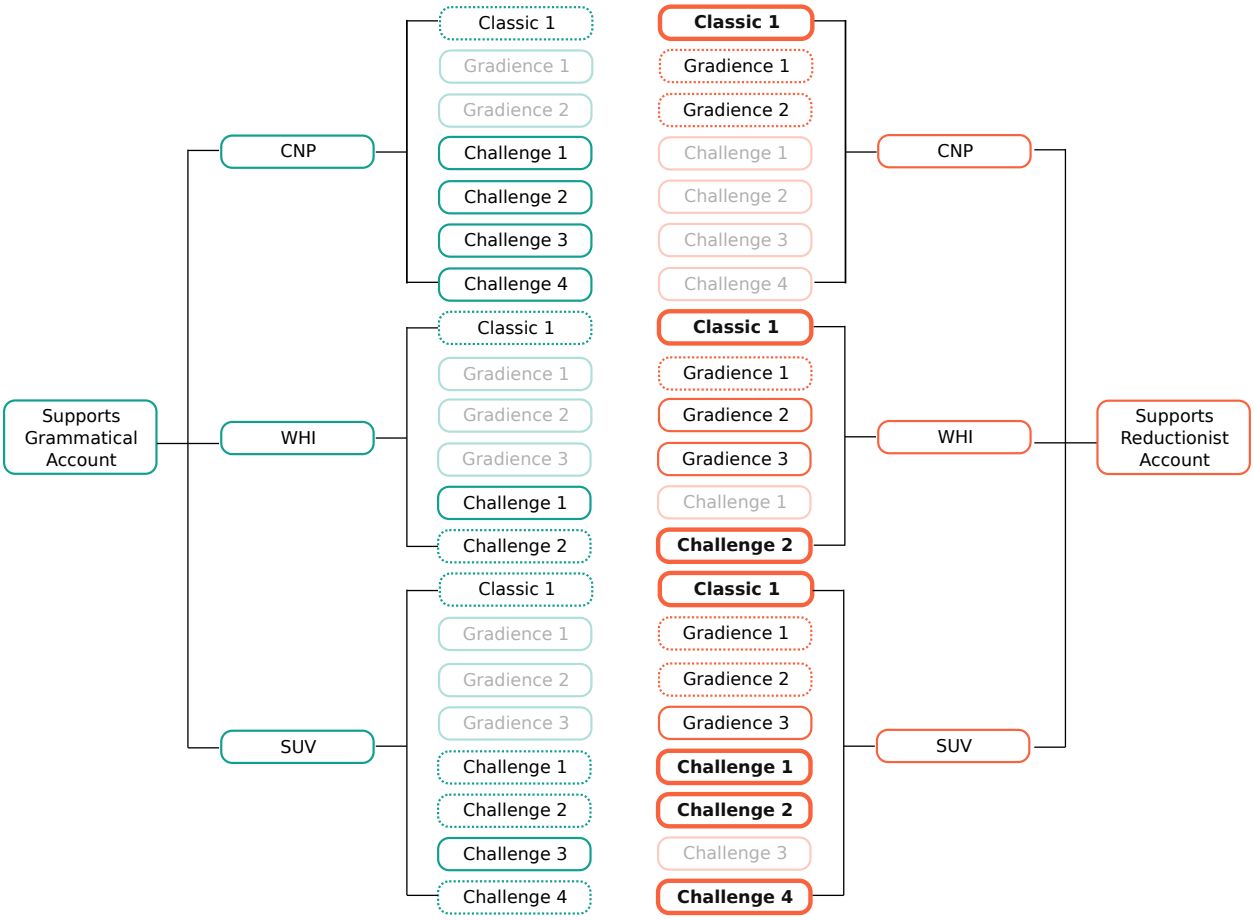


Figure 7.32: An updated organizational chart of syntactic locality experiments.

7.4 Discussion

7.4.1 Comparison to other work

The main results indicate that cognitive factors encoded as probabilistic features can model weak and non-island locality, but can not model strong islands. This supports a distinction between the three phenomena, which may in fact be on the boundary of competence and performance.

The distinction between strong and weak islands is as old as islands themselves. Ross (1967) notes that WHIs should allow more extractions than strong islands, and this prediction is born out and supported by a variety of syntactic and semantic studies (Rizzi, 1990; Manzini, 1992, 1994; Szabolcsi & Zwarts, 1993). For example, Rizzi (1990) argues that weak islands are subject to grammatical constraints like Relativized Minimality, whereas strong islands are subject to stronger, path-based constraints like the Extended Category Principle (ECP). Szabolcsi and Zwarts (1993) take a more semantics-based approach, wherein the gradient behavior of WHIs is explainable because wh-words that range over individuals can escape from weak islands, but not strong islands.

Although this work does not directly test either of these grammar-based distinctions for strong and weak islands, it can support a competence-performance explanation, such as that provided by Alexopoulou and Keller (2003) and Alexopoulou and Keller (2007). The authors demonstrate experimentally that weak islands, like non-islands, are subject to memory-derived difficulty like embedding, whereas strong islands are not. Their experiments are detailed in Chapter 6, Section 7.2.2, and Section 7.2.3. The modeling work done here supports their main conclusions: WHIs are modeled by memory constraints, whereas CNPs are not.

Alexopoulou and Keller (2007) provide an updated version of the SPLT (Gibson, 2000)

to handle this contrast, further detailed in Chapter 4. There are a variety of differences between our methods. First, the methodology used here does not require a modification to the theoretical basis of existing theories to model the data: Lewis and Vasishth (2005) Activation and Interference quantities, retrieval, the DLT, and ideas like filler-type are all pulled directly from the literature and implemented in the parser probabilistically. Like Alexopoulou and Keller (2007), the results here demonstrate that DLT performs poorly on the data, which indicates it may not be the best cognitive hypothesis for this data set. But, it's not necessary to create a new memory theory to model this data: several others do just as well modeling their results.

The second major difference between this work and Alexopoulou and Keller (2007) is that this is a working, broad-coverage model. Alexopoulou and Keller (2007) encode by hand the new SPLT and work it on a few examples, whereas this is tested across many languages, phenomena, and experiments. It is possible that encoding their SPLT would provide even better performance on this data set, which can be tested in future work. But as it stands, this model can achieve the strong and weak island distinction with established, independently-posited cognitive hypotheses.

Alexopoulou and Keller (2007) were not the first to attempt a cognitive explanation of islands. Kluender (1998) argues that the strong and weak island distinction is instead based on an interaction of processing factors. Activation causes difficulty in strong islands, and an additional interference effect holds with the wh-intervenor for weak islands. Thus, the strong and weak island distinction is simply due to different types of processing difficulty.

Although Kluender's hypothesis for WHIs is spot-on, the results do not support his claim for CNPs. Despite using a variety of established processing factors in an independently-motivated psycholinguistic model, the CNP data is not modeled. This could be a result of the type of strong island data modeled: the complement clause CNPs considered here are not the only kind of strong island. It may also be a result of the experiments chosen. However, these experiments were chosen because they demonstrated a broad range of behavior (classic,

gradient, and challenging) in different languages; one would expect the cognitive factors to perform well on at least a few.

It may also be the case that the cognitive factors do operate as Kluender suggests as long as they are combined. Kluender argues that it is a confluence of processing factors that predict strong islands. It is possible that the conjunction of filler and activation and DLT could model the CNP data, but the individual features do not. One of the advantages of this parsing model, and the research design, is that it is possible to combine features into larger features to get different predictions. Pursuing this question is a topic of future research, discussed in Chapter 8.

Sag and his colleagues have also argued for a processing-based account to syntactic locality. In addition to activation and interference, they argue that filler-type (the basis for the FILLER feature), frequency effects, and surprisal can explain gradience in CNP and WHI data. This model incorporates all of these factors, yet it is unable to model strong islands.

This may be because CNPs are just hard. In the experimental evidence, the reading time differences for CNPs aren't as great as they are for WHIs in the Hofmeister and Sag (2010) experiments for both phenomena. This is what the Alexopoulou and Keller (2007) data indicates as well. Perhaps the human subjects reach a certain level of difficulty after which any evidence of gradience is unreliable. The parser itself finds the CNPs to be around the same level of difficulty: all very difficult, and any gradience doesn't seem to make a difference. This points to a concern with experimental design: if sentences are difficult, how much does a statistically-significant difference in reading time mean?

This is similar to an argument against reductionist accounts by Phillips (In Press). He argues that memory factors, like filler-type, may cause slightly gradient behavior in even ungrammatical sentences. The problem is that the distinction between ungrammatical, unacceptable, and difficult sentences is unclear, as is further discussed in Chapter 5.1.

These results support Phillips and his colleagues' claims for strong islands, but do not support their claims for weak and non-island locality. The cognitive factors model these phenomena relatively effortlessly, without recourse to specific grammatical constraints like Relativized Minimality. In fact, as discussed in Boston (2010), a version of Relativized Minimality encoded in the parser performs badly on all but the classic examples. It may be that the probabilistic grammar is sensitive to a particular grammatical constraint across all the languages and the treebanks, helping along these memory constraints. But this grammatical factor appears to be probabilistic and frequency-based, not a hard constraint.

7.4.2 CNPs: Argument from a null result

To view these results as supporting a competence-performance explanation of the strong and weak island distinction, the negative result for CNPs is crucial. But, it is a null result, and therefore a weaker argument than could be made if the parser could model the behavior in some way. The parser's inability to model CNPs could be because there actually is a competence-performance explanation of strong and weak islands, and cognitive factors simply can not model them. This is argued by Alexopoulou and Keller (2007). But, it is also possible that these cognitive factors require integration as Kluender argues, or that a different type of cognitive constraint, such as integration, is required. And it may just be that this particular model has a crucial shortcoming that does not allow it to accurately model the cognitive factors in relation to CNPs.

If Alexopoulou and Keller (2007) are correct, an extra grammatical constraint is required to model the CNP results. One possibility from the linguistic literature is the ECP, Empty Category Principle (Aoun et al., 1982). The ECP requires traces, or gaps in psycholinguistic terminology, be properly governed by their antecedents, or fillers, as is detailed in Chapter 2. Unfortunately, this constraint can not be implemented in the parser because the parser can not encode c-

command. C-command can not be encoded in dependency grammar because it requires a notion of hierarchy: it is not possible to encode dominance such that x does not dominate y, y does not dominate x, and the maximal projection of x dominates y. Future work on a parser operating on a more constrained formalism, such as TAG (Joshi, Levy, & Takahashi, 1975) or Minimalist Grammars (Stabler, 1997) is required to test whether the addition of a grammatical constraint can allow the computational model to handle CNPs.

Alternatively, it could be that no grammatical constraint is necessary. As Kluender argues, an interaction of the cognitive factors that are considered here, such as combining DLT with DECAY, may provide the correct recipe for a reductionist account of strong islands. It's also possible that the winning recipe requires some other cognitive feature, or even a non-cognitive feature from the dependency parsing community that hasn't been tested yet. Examining these possibilities is possible and relatively easy to do given this parser design. Unfortunately, though, it requires a significant time commitment requiring testing and data analysis across all possible combinations of features. Given the eight features above, there are 246 combinations of feature pairs, triples, etc. This process can of course be automated to reduce analysis time, but it falls outside the scope of this work.

Finally, the third possible reason why CNPs are not modeled is because of a flaw within the parser. The nature of modeling is such that there can never be a perfect model. As was detailed in Chapter 4 and Chapter 6, computational models necessitate explicit decisions when there is no clear correct or well-supported answer. There are many decisions that may have prevented the parser from modeling these sentences, from the parsing algorithm, the grammar, the treebanks, the syntactic analyses, the complexity measures, the cognitive factors, their encoding as features, etc. But, these decisions were not made arbitrarily; they were made, as often as possible, thoughtfully and with the support of previous work. A complete story of syntactic locality modeling would consider each of these alternations and compare results to find the best possible model for syntactic locality. But, by necessity, this complete story is

beyond the scope of this dissertation.

7.4.3 WHIs: A reductionst phenomenon

Of the three phenomena, the cognitive factors perform best on WHIs. Nearly every experiment is modeled by a wide range of features, but the best-performing features across the experiments are DECAY, SBI, and FILLER. This result supports arguments by Kluender (1998) and Hofmeister and Sag (2010) that weak islands are explainable by a combination of processing factors. The result is also predicted because WHIs are susceptible to activation and interference effects, and behave as the cognitive theories would predict. The only experiment that could not be modeled is explained by a dependency-length mismatch, and the tendency for surprisals to be higher for longer dependencies. Although there are fewer WHI studies modeled than CNP and SUV studies, this is because there is relatively little variation, both within the English experimental evidence and cross-linguistically. Overall, the WHI result supports reductionist accounts of weak islands.

7.4.4 SUVs: Insight from a computational model

SUVs can be modeled by a specific set of cognitive factors that take into account interference. This result is clear if the gradience experiments are ignored because of dependency length mismatches. The well-modeled experiments provide insight into several on-going controversies. For example, linguists have long-argued that SUVs are acceptable in German, a position supported by Fanselow and Féry (2007). But, Featherston (2005) provides acceptability judgments that demonstrate strong SUV unacceptability for German speakers, questioning not only the role of superiority violations in German, but also the syntactic judgments used by linguists. Fanselow and Féry (2007) then conduct a follow-up acceptability experiment to demonstrate

that Featherston's results are most likely driven by case differences rather than SUVs within the sentence. Their results indicate that Germans find SUVs to be just as acceptable as non-SUVs.

This model supports the original linguistic intuitions and the Fanselow and Féry (2007) results; it does not find SUVs to be more difficult than non-SUVs in German. The parser's behavior is the same across all the German experimental data, including Featherston's. Note that the only difference between the parsing model for English, where the parser finds SUVs difficult, and German, where the parser does not find SUVs difficult, is the treebank. This different behavior should pose a problem for the cognitive factors, but it does not. This is likely due to some aspect of the probabilistic grammar and the dependency analysis itself that aids SUVs in German, but is unavailable in English. What it may be is a question for future research, but this result emphasizes the utility of explicit computational models in linguistic work: the German SUV data should be a challenge to reductionist accounts, yet factors like frequency information, surprisal, and architectural mechanisms can interact in surprising and complex ways to aid in modeling the data.

Finally, unlike WHIs, SUVs are best-modeled by features that take into account one particular cognitive theory: interference. Activation and filler-type are less useful for this type of syntactic locality. Within linguistics, SUVs and WHIs are handled by two different mechanisms, Relativized Minimality and Subjacency. But, this model indicates that the distinction between WHIs and SUVs could instead be based on different types of cognitive factors, precisely as is argued by Kluender (1998).

7.5 Conclusion

The original intent of this research was to build a computational model that could inform the competence-performance divide. This idea is not novel: in fact, many researchers have demonstrated the utility of computational work for explaining phenomena that straddle the divide between grammar and processing (Kimball, 1973; Wanner & Maratsos, 1978; Marcus, 1980; Berwick & Weinberg, 1984; Bresnan & Kaplan, 1982). This model is a small contribution towards the goal of understanding the complexities of syntactic locality. The results demonstrate that a computational model armed with cognitive constraints performs surprisingly well on a variety of syntactic locality phenomena. Although the model points out shortcomings in both grammatical and reductionist accounts, it also highlights the strengths in well-thought-out experimental data and supports arguments for a reductionist account of weak islands and superiority violations.

CHAPTER 8

CONCLUSIONS

8.1 Introduction

The objective of this research is to build a computational model that tests reductionist explanations of syntactic locality. This model tests a range of experimental data against cognitive constraints, seeking to illuminate the nature of syntactic locality and its role in the competence-performance debate. This chapter discusses how the results fulfill these objectives. Section 8.2 summarizes the findings and their implications, while Section 8.3 discusses what these findings contribute to the field. Section 8.4 discusses directions for future research, and Section 8.5 concludes.

8.2 Summary of findings

The findings support reductionist accounts for some syntactic locality phenomena, namely weak islands and superiority violations. In particular, reductionist accounts that focus on working memory retrieval are able to model the phenomena. Whereas difficulty in weak islands is associated with both activation and interference, difficulty for superiority violations appears to mainly be a result of interference difficulty.

Strong islands are not modeled by the retrieval-based cognitive constraints considered in this work. This indicates that the differences between strong and weak islands are real, and may even span the competence and performance divide. It is also possible that strong islands, like weak islands, are explained by processing factors, but by different factors than are considered here. In particular, storage-based accounts of working memory difficulty were not

tested, and could provide a reductionist explanation of difficulty.

These results support a methodology that uses broad-coverage statistical parsers to test a wide range of experimental data. They also support encoding cognitive factors as probabilistic features. This encoding practice has two benefits. First, it provides a standard for comparison across cognitive theories. Secondly, it limits the computational model's memory in a cognitively plausible way, providing a more realistic human sentence processing model.

8.3 Contributions

The methodology used here provides a variety of novel approaches to the often difficult and non-standard task of human sentence processing modeling. It provides a method for aggregating data from experiments that use many methodologies and often support opposing claims. The methodology organizes the experimental data into categories based on whether they confirm or contradict reductionist claims, and uses cross-linguistic data as often as possible. This classification of experimental data, and use of many experimental data points, allows for a broad-coverage approach that lessens the risk of overfitting the data. This methodology can be useful for researchers interested in specific phenomena rather than a range of difficulties, and can fill-in until standard test sets are available for human sentence processing models.

At a more theoretical level, this research combines the long tradition of computational locality models (Marcus, 1980; Berwick & Weinberg, 1984; Deane, 1991; Frank, 1992; Pritchett, 1992, 1993) with the long tradition of cognitive models of language (Gibson, 1991; Just & Carpenter, 1992; Lewis, 1993, 1996; Gibson, 1998, 2000; Lewis, 1999; Vosse & Kempen, 2000; Lewis & Vasishth, 2005; Just & Varma, 2007). The first computational models that specifically address syntactic locality focus on how grammatical locality constraints derive from a deterministic parsing architecture (Marcus, 1980; Berwick & Weinberg, 1984). These models

demonstrate that local attachments are preferred architecturally, thereby demonstrating the strong link between the grammar and parser. Subsequent work by Frank (1992) provides further support for an approach that maintains strong competence, or a direct correspondence between grammar rules and parser operations (Bresnan & Kaplan, 1982). Frank, however, argues that locality constraints reside in the grammar, and demonstrates how they help a sentence processing model predict island difficulty. Both of these contributions focus on grammatical constraints, and whether the grammatical constraints arise from the architecture or from the grammar. Although this model also instantiates the strong competence hypothesis, it focuses on cognitive rather than grammatical constraints.

Deane (1991) and Pritchett (1993) describe models that similarly test cognitive constraints against syntactic locality data. But unlike these implementations, this work does not argue for a specific cognitive hypothesis or explanation of the phenomena. Rather, it takes advantage of broad-coverage parsing to test a variety of cognitive principles, including those argued for by Deane and Pritchett. Pritchett argues that syntactic locality difficulty is caused by unrecoverable parsing errors when attempting to attach a filler to its gap. Like Pritchett, this work localizes the difficulty to this particular attachment, but this research does not consider how the cognitive constraints could prevent the attachment. Rather, the work considers the difficulty of the attachment once it has been made. Pritchett's model is able to account for Subjacency, particularly CNP difficulty, which is the phenomenon that this model struggles with. Therefore, it would be interesting in the future to run the parser to test attachment rather than difficulty, to see whether this would allow strong island difficulty to be handled by these cognitive constraints. However, this may be detrimental for WHI and SUV modeling.

Like Deane, this work argues that working memory can account for both the ease and the difficulty in processing certain long-distance dependencies. Deane details a model of syntactic locality based on spreading activation in memory, and how memory decay contributes to island difficulty. The DECAY and RETRIEVAL probabilistic features used here use specific cognitive-based

measures of spreading activation, and can be thought of as a variant of Deane's spreading activation hypothesis. However, this model also incorporates interference-based measures that are found to be helpful in modeling syntactic locality. Further, this model tests more phenomena and languages than the Deane model. Regardless, many of the ideas from Deane's work are applicable to this model as well.

In fact, this work builds on the many computational models of language that specifically address diverse cognitive constraints in sentence processing[1]. The current work in many ways simply applies the hypotheses put forth by these cognitive models to the syntactic locality data. Some implementations account for human sentence processing difficulty on the basis of activation, such as the Unification Space Model of Vosse and Kempen (2000), CC READER (Just & Carpenter, 1992), and its modern implementation, CAPS (Just & Varma, 2002). In each of these systems, the amount of memory affects the structure being built; in the Unification Space Model, the structures are lexicalized and very similar to the DG approach considered here.

Most DLT-based models consider the effect of storage cost of unsatisfied elements (Gibson, 1991, 1998, 2000; Alexopoulou & Keller, 2007; Demberg & Keller, 2008; Demberg-Winterfors, 2010). This work considers DLT in a different way: not only is the DLT used as a model of integration cost and less a model of storage cost, but it is also adapted to broad-coverage parsing. The second modification is similar to previous work by Demberg and Keller (2008) and Demberg-Winterfors (2010) that implements the DLT on broad-coverage model. In the Demberg-Winterfors (2010) model, though, the DLT is part of a new complexity metric that combines a DLT-like metric with a probabilistic metric, surprisal. In some ways, one could consider this current work as a combination of the DLT, and several other memory-based hypotheses, and surprisal. Whereas Demberg-Winterfors explicitly combines the two in a

[1] There is a large literature on connectionist models of human sentence processing (Elman, 1990, 1991; Tabor et al., 1997; Tabor & Tanenhaus, 1999; Christiansen & Chater, 1999), but symbolic and hybrid architectures are considered here.

new metric, in this work the two are combined in the probabilistic space: the DLT defines the probability space, and surprisal then measures difficulty based on this space. It is unclear how these methodologies differ; it would be necessary to apply them to the same data to better understand how the two relate.

This model also builds off of implementations that consider the effect of interference on sentence processing (Lewis, 1993, 1996, 1999; Lewis & Vasishth, 2005). In many ways, the notion of SBI used here, which comes from Lewis and Vasishth (2005), is a direct descendent of the interference implementations first coded in NL-SOAR by Lewis. However, this work only considers retrieval interference and does not consider how storage interference could lead to processing difficulty in syntactic locality sentences.

Finally, the DECAY, SBI, and PROBRETRIEVAL features, as well as the RETRIEVAL complexity metric, all directly come from the retrieval theory and computational implementation in Lewis and Vasishth (2005). This work simply adapts this system to the Nivre dependency transition system, as detailed in Chapter 4. The chapter also details several important differences between this implementation and the Lewis and Vasishth implementation, the most important of which is the nature of the memory in the system. In particular, this parser does not encode a difference between procedural and declarative memory, an important part of the Lewis and Vasishth system. It also uses a limitless stack-based memory, not the more human-like associative memory in the previous system.

This model undoubtedly owes much to the many implementations that have come before it. However, it is different because it combines computational modeling with many cognitive constraints to test a specific linguistic phenomenon. While it is most similar to human sentence processing models that directly encode memory difficulty (Lewis & Vasishth, 2005), its focus is a topic at the heart of grammar, reminiscent of classic locality models (Marcus, 1980; Berwick & Weinberg, 1984). Yet it is not the first computational model of reductionist claims of syntactic locality (Deane, 1991; Pritchett, 1993); unlike these predecessors, though, the focus of this

work is not to prove a specific reductionist account. Rather, it is to better understand the many theories implicated in reductionist accounts, and to test them against a wide range of data. This contribution, though small in the large literature of syntactic locality, has been missing in this crucial linguistic debate.

8.4 Future work

Despite the relative simplicity of this model, there are a variety of testable hypotheses available in the current implementation that, due to lack of time, are as yet unexplored. First, by encoding the constraints as probabilistic features, the framework supports a method for directly comparing multiple cognitive theories. An additional benefit, though, is that the probabilistic features can be combined to generate new predictions. For example, retrieval theory argues that sentence processing difficulty is a result of two quantities: word decay and similarity-based interference. The theory combines these two as simple addends. Yet, it is equally possible to have the combination mixed by a machine learning algorithm, and generate new predictions that can provide more information on working memory difficulty. Similarly, it is possible that while the similarity-based interference quantity that retrieval argues for is accurate, a more accurate measure of decay is something simpler, like string distance. Or perhaps both are needed. The advantage of the probabilistic model is that each of these combinations can be compared, directly testing Kluender's "confluence of factors" theory.

The current implementation has the added advantage of having multiple methods of predicting processing difficulty. This work considers how difficult it is to make the correct attachment given the correct parse. But, the architecture also supports k-best search, such that the model can take into account an arbitrary number of analyses to determine what amount of memory is required to make the attachment. For example, it may be the case that a serial parser can make relatively easy, non-violating attachments, but a highly parallel parser is

required to make attachments in islands. This provides another estimate of working memory difficulty that is quite different from the difficulty argued for in reductionist accounts. Yet, it can be combined with these accounts, as demonstrated in previous work on superiority violations (Boston, 2010).

This alternative method of testing difficulty, based on whether the parser creates an attachment in an island-violating condition or not, implements an accurate test for reductionist accounts that is not often considered (Phillips, In Press). Although this work considers complexity metrics like surprisal and retrieval, these metrics require the attachment be made. However, as Phillips points out, this is not a foregone conclusion. An evaluation of this approach is left to future work.

This work uses two distinct complexity metrics, surprisal and retrieval. In recent work, attempts have been made to create an over-arching complexity metric that combines both surprisal-based quantities and memory-based quantities in the form of the DLT (Demberg-Winterfors, 2010). Although the combination metric works well, in this work each is considered separately. Yet by calculating suprisal based on cognitive features like the DLT, the complexity metrics are being combined. In future work, it would be interesting to see how this method compares to the other.

The original goal was to create and evaluate a model that encodes both grammatical and cognitive constraints for syntactic locality. Unfortunately, that was not possible because many of the most promising grammatical constraints require more restrictive MCSG formalisms. Some, like Relativized Minimality (Rizzi, 1990; Cinque, 1990), were implemented, as discussed in previous work (Boston, 2010). It seems the most promising model for strong islands would incorporate both grammatical and cognitive constraints. A model that includes an ECP-like constraint, while at the same time using retrieval-based cognitive features like activation, may provide an accurate model of the gradience evident in strong island data. However, as has been discussed in previous chapters, it is impossible to accurately encode the ECP in a

dependency parser. A move to a more restrictive formalism, such as a Minimalist Grammar parser, would be necessary in future work.

8.5 Conclusion

This research satisfies the original research objective, which is to build a computational model that illuminates the nature of syntactic locality. Yet the laundry list of questions that this research has raised, as well as the many tasks left for future work, highlight the difficulty of syntactic locality data. In the larger debate on grammatical versus reductionist approaches, and the even greater debate on competence and performance in linguistics, this research provides a small computational perspective that affirms previously held beliefs: syntactic locality, and language, is explained by a combination of grammatical and cognitive factors. However, by combining the methodologies of linguistics, psycholinguistics, and computational linguistics, I hope to have further motivated the use of human sentence processing models to inform these large questions.

APPENDIX A

LIST OF ABBREVIATIONS

ACT-R Adaptive Control of Thought-Rational, 34

CFG Context Free Grammar, 43

CNPs Complex Noun Phrase Islands, 3

DG Dependency Grammar, 8

DLT Dependency Locality Theory, 5

ECP Empty Category Principle, 18

MCSG Mildly Context-Sensitive Grammar, 43

ME Magnitude Estimation, 85

NLP Natural Language Processing, 6

PCFG Probabilistic Context-Free Grammar, 44

POS Part-of-Speech, 54

RRT Residual Reading Time, 86

SBI Similarity-based Interference, 37

SUVs Superiority Violations, 3

WHIs Wh-Islands, 3

REFERENCES

Adger, D. (2003). *Core syntax: A minimalist approach*. New York, NY: Oxford University Press. 15

Alexopoulou, T., & Keller, F. (2003). Linguistic complexity, locality and resumption. In G. Garding & M. Tsujimura (Eds.), *WCCFL 22 proceedings* (pp. 15–28). Somerville, MA: Cascadilla Press. 158

Alexopoulou, T., & Keller, F. (2007). Locality, cyclicity, and resumption: At the interface between the grammar and the human sentence processor. *Language*, *83*(1), 110–160. 32, 91, 94, 96, 111, 113, 118, 119, 129, 135, 152, 154, 158, 159, 160, 161, 169

Allwood, J. (1982). The complex NP constraint in Swedish. In E. Engdahl & E. Ejerhed (Eds.), *Readings on unbounded dependencies in Scandinavian languages* (pp. 15–32). Stockholm: Almqvist & Wiksell International. 104

Anderson, J. R. (1976). *Language, memory and thought*. Hillsdale, NJ: Lawrence Erlbaum. 3, 34, 37

Anderson, J. R. (2002). ACT: A simple theory of complex cognition. In T. A. Polk & C. M. Seifert (Eds.), *Cognitive modeling* (pp. 49–68). Cambridge, MA: MIT Press. 31

Anderson, J. R. (2005). Human symbol manipulation within an integrated cognitive architecture. *Cognitive Science*, *29*, 313–341. 3, 34, 37, 73

Anderson, J. R., & Lebiere, C. (1998). *Atomic components of thought*. Hillsdale, NJ: Lawrence Erlbaum. 34, 73

Andersson, L.-G. (1982). What is Swedish an exception to? extractions and island constraints. In E. Engdahl & E. Ejerhed (Eds.), *Readings on unbounded dependencies in Scandinavian languages* (pp. 15–32). Stockholm: Almqvist & Wiksell International. 104

Aoun, J., Hornstein, N., & Sportiche, D. (1982). Some aspects of wide scope quantification. *Journal of Linguistic Research*, 69–95. 18, 161

Arnon, I., Snider, N., Hofmeister, P., Jaeger, T. F., & Sag, I. (To Appear). Cross-linguistic

variation in a processing account: The case of multiple wh-questions. In *Proceedings of Berkeley Linguistics Society* (Vol. 32). Berkelely, CA: Berkeley Linguistics Society. 2, 22, 29, 30, 56, 98, 99, 104, 141, 142

Attardi, G. (2006). Experiments with a multilanguage non-projective dependency parser. In *Proceedings of the 10th Conference on Computational Natural Language Learning (CoNLL)* (pp. 166–170). New York, NY. 45, 59

Bard, E. G., Robertson, D., & Sorace, A. (1996). Magnitude estimation of linguistic acceptability. *Language*, *72*, 32–68. 85

Berwick, R., & Weinberg, A. (1984). *The grammatical basis of linguistic performance*. Cambridge, MA: MIT Press. 45, 165, 167, 170

Boston, M. F. (2010). The role of memory in superiority violation gradience. In *Proceedings of the ACL Workshop on Cognitive Modeling and Computational Linguistics.* 80, 161, 172

Boston, M. F., Hale, J. T., Patil, U., Kliegl, R., & Vasishth, S. (2008). Parsing costs as predictors of reading difficulty: An evaluation using the Potsdam Sentence Corpus. *Journal of Eye Movement Research*, *2*(1), 1–12. 44, 46, 79

Boston, M. F., Hale, J. T., Vasishth, S., & Kliegl, R. (2011). Parallel processing and sentence comprehension difficulty. *Language and Cognitive Processes*, *26*, 301–349. 39, 44, 46, 69, 75, 79, 80, 84

Bourdages, J. (1992). Parsing complex NPs in French. In H. Goodluck & M. Rochemont (Eds.), *Island constraints: Theory, acquisition and processing* (pp. 61–87). Dordrecht: Kluwer. 32

Bošković Željko. (1998). On certain violations of the superiority condition, AgrO, and economy of derivation. *Journal of linguistics*, *33*, 227–254. 104

Brants, S., Dipper, S., Eisenberg, P., Hansen-Schirra, S., König, E., Lezius, W., et al. (2004). TIGER: Linguistic interpretation of a german corpus. *Research on Language and Computation*, *2*, 597–619. 54

Bresnan, J., & Kaplan, R. M. (1982). Introduction: Grammars as mental representations of

language. In J. Bresnan (Ed.), *The mental representation of grammatical relations* (pp. xvii–lii). Cambridge, MA: MIT Press. 165, 168

Brown, J. (1958). Some tests of the decay theory of immediate memory. *Quarterly Journal of Experimental Psychology*, *10*, 173–189. 31

Buch-Kromann, M. (2006). *Discontinuous grammar: A dependency-based model of human parsing and language learning*. Unpublished doctoral dissertation, Copenhagen Business School. 44

Caplan, D., & Waters, G. S. (1999). Verbal working memory and sentence comprehension. *Behavioral and Brain Sciences*, *22*, 77–94. 31, 73

Charles Clifton, J., & Frazier, L. (1986). The use of syntactic information in filling gaps. *Journal of Psycholinguistic Research*, *15*, 209–224. 65

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press. 4

Chomsky, N. (1970). Remarks on nominalization. In R. Jacobs & P. Rosenbaum (Eds.), *Readings in English transformational grammar* (pp. 184–221). Waltham, MA: Ginn and Company. 13

Chomsky, N. (1973). Conditions on transformations. In S. Anderson & P. Kiparsky (Eds.), *A Festschrift for Morris Halle* (pp. 232–286). New York: Holt, Reinhart and Winston. 2, 18, 20

Chomsky, N. (1977). On *Wh*-movement. In P. W. Cullicover, T. Wasow, & A. Akmajian (Eds.), *Formal syntax* (pp. 71–132). New York, NY: Academic Press. 18

Christiansen, M., & Chater, N. (1999). Toward a connectionist model of recursion in human linguistic performance. *Cognitive Science*, *23*(2), 157–205. 34, 43, 169

Chung, S. (1994). "Referentiality" in Chamorro. *Linguistic Inquiry*, *25*(1), 1–44. 27

Cinque, G. (1990). *Typse of A'-dependencies*. Cambridge, MA: MIT Press. 2, 5, 15, 20, 25, 27, 172

Clifton, C., Fanselow, G., & Frazier, L. (2006). Amnestying superiority violations: Processing multiple questions. *Linguistic Inquiry*, *37*(1), 51–68. 2, 104, 110

Clifton, C., & Frazier, L. (1989). Comprehending sentences with long-distance dependencies. In G. N. Carlson & M. K. Tanenhaus (Eds.), *Linguistic structure in language processing* (pp. 273–318). Kluwer. 2

Comorovski, I. (1989). Discourse-linking and the wh-island constraint. In J. Carter & R.-M. Déchaine (Eds.), *Proceedings of the Nineteenth Meeting of the North East Linguistic Society* (pp. 78–96). Amherst, MA: University of Massachusetts Department of Linguistics. 27

Covington, M. A. (2001). A fundamental algorithm for dependency parsing. In *Proceedings of the 39th Annual ACM Southeast Conference* (pp. 95–102). 46, 47

Cowart, W. (1997). *Experimental syntax: applying objective methods to sentence judgments*. Thousand Oaks, CA: Sage Publications. 85, 110

Crocker, M. (2005). Rational models of comprehension: Addressing the performance paradox. In A. Cutler (Ed.), *Twenty-first century psycholinguistics: Four cornerstones* (pp. 1–18). Hillsdale, NJ: Lawrence Erlbaum. 7, 61

Crocker, M., & Brants, T. (2000). Wide-coverage probabilistic sentence processing. *Journal of Psycholinguistic Research*, *29(6)*, 647–669. 61

de Swaart, H. (1992). Intervention effects, monotonicity and scope. In C. Barker & D. Dowty (Eds.), *SALT II: Proceedings of the Second Conference on Semantics and Linguistic Theory* (Vol. 40, pp. 387–406). Columbus, OH. 25

Deane, P. (1991). Limits to attention: A cognitive theory of island constraints. *Cognitive Linguistics*, *2*(1), 1–63. 29, 32, 167, 168, 169, 170

Demberg, V., & Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, *109*(2), 193–210. 32, 169

Demberg-Winterfors, V. (2010). *A broad-coverage model of prediction in human sentence processing*. Unpublished doctoral dissertation, University of Edinburgh. 32, 43, 169, 172

Dubey, A. (2004). *Statistical parsing for German: Modeling syntactic properties and annotation*

differences. Unpublished doctoral dissertation, Saarland University, Germany. 54

Earley, J. (1970). An efficient context-free parsing algorithm. *Communications of the ACM*, *13*, 94–102. 45

Eisner, J. (1996a). *An empirical comparison of probability models for dependency grammar* (Tech. Rep. No. IRCS-96-11). Institute for Research in Cognitive Science, University of Pennsylvania. 45

Eisner, J. (1996b). Three new probabilistic models for dependency parsing: An exploration. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING-96)* (pp. 340–345). Copenhagen. 45

Eisner, J. (2000). Bilexical grammars and their cubic-time parsing algorithms. In H. Bunt & A. Nijholt (Eds.), *Advances in probabilistic and other parsing technologies* (pp. 29–62). Dordrecht: Kluwer Academic Publishers. 45

Eisner, J., & Smith, N. A. (2005). Parsing with soft and hard constraints on dependency length. In *Proceedings of the International Workshop on Parsing Technologies (IWPT)* (pp. 30–41). Vancouver. 45, 76

Eisner, J., & Smith, N. A. (2010). Favor short dependencies: Parsing with soft and hard constraints on dependency length. In H. Bunt, P. Merlo, & J. Nivre (Eds.), *Trends in parsing technology: Dependency parsing, domain adaptation, and deep parsing* (pp. 121–150). New York, NY: Springer. 45, 76

Elman, J. (1990). Finding structure in time. *Cognitive Science*, *14*, 179–211. 33, 43, 169

Elman, J. (1991). Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning*, *7*(2-3), 195–225. 33, 43, 169

Engdahl, E. (1982). Restrictions on unbounded dependencies in Swedish. In E. Engdahl & E. Ejerhed (Eds.), *Readings on unbounded dependencies in Scandinavian languages* (pp. 151–174). Stockholm: Almqvist & Wiksell International. 104

Fanselow, G., & Féry, C. (2007). Missing superiority effects: Long movement in German (and other languages). In J. Witkòs & G. Fanselow (Eds.), *Proceedings of PLM 2006.*

Frankfurt: Lang. 100, 101, 102, 145, 146, 163, 164

Featherston, S. (2005). Universals and grammaticality: wh-constraints in German and English. *Linguistics*, *43*(4), 667–711. 99, 100, 110, 143, 145, 146, 163

Featherston, S. (2009). Relax, lean back, and be a linguist. *Zeitschrift für Sprachwissenschaft*, *28*(1), 127–132. 110

Fedorenko, E., & Gibson, E. (Submitted). *Syntactic parallelism as an account of superiority effects: Empirical investigations in English and Russian.* 56, 99, 101, 102, 142, 146

Fodor, J. (1979). 'Superstrategy'. In W. E. Cooper & E. C. T. Walker (Eds.), *Sentence processing.* Hillsdale, NJ: Erlbaum Press. 32

Fodor, J. (1992). Islands, learnability and the lexicon. In H. Goodluck & M. Rochemont (Eds.), *Island constraints: Theory, acquisition and processing* (pp. 109–180). Dordrecht: Kluwer. 15

Fodor, J. D. (1978). Parsing strategies and constraints on transformations. *Linguistic Inquiry*, *9*(3), 427–473. 29

Francis, W. N., & Kucera, H. (1979). *Brown corpus manual* (Tech. Rep.). Department of Linguistics, Brown University, Providence, RI. 54

Frank, R. (n.d.). *Phrase structure composition and syntactic dependencies*. Cambridge, MA: MIT Press. 15

Frank, R. E. (1992). *Syntactic locality and tree adjoining grammar: Grammatical, acquisition and processing perspectives*. Unpublished doctoral dissertation, University of Pennsylvania. 167, 168

Frazier, L. (1979). *On comprehending sentences: Syntactic parsing strategies*. Unpublished doctoral dissertation, University of Connecticut. 32

Gärtner, H.-M., & Michaelis, J. (2007). Some remarks on locality conditions and Minimalist Grammars. In U. Sauerland & H.-M. Gärtner (Eds.), *Interfaces + recursion = language?* (pp. 161–195). Berlin: Mouton de Gruyter. 4

Gibson, E. (1998). Linguistic complexity: locality of syntactic dependencies. *Cognition*, *68*,

1–76. 30, 32, 33, 35, 36, 37, 73, 167, 169

Gibson, E. (2000). Dependency locality theory: A distance-based theory of linguistic complexity. In A. Marantz, Y. Miyashita, & W. O'Neil (Eds.), *Image, language, brain: Papers from the First Mind Articulation Symposium* (pp. 95–126). Cambridge, MA: MIT Press. 3, 5, 30, 32, 33, 35, 37, 38, 61, 158, 167, 169

Gibson, E., Pearlmutter, N., Canseco-Gonzalez, E., & Hickok, G. (1996). Recency preference in the human sentence processing mechanism. *Cognition*, *59*, 23–59. 33

Gibson, E., & Pearlmutter, N. J. (1998). Constraints on sentence comprehension. *Trends in cognitive science*, *2*(7), 262–268. 33

Gibson, E., & Thomas, J. (1999). Memory limitations and structural forgetting: The perception of complex ungrammatical sentences as grammatical. *Language and Cognitive Processes*, *14*(3), 225–248. 33

Gibson, E. A. F. (1991). *A computational theory of human linguistic processing: memory limitations and processing breakdown*. Unpublished doctoral dissertation, Carnegie Mellon University. 30, 32, 167, 169

Gordon, P. C., Hendrick, R., & Johnson, M. (2001). Memory interference during language processing. *Journal of experimental psychology: Learning, memory, and cognition*, *27*(6), 1411–1423. 31, 32, 37

Gordon, P. C., Hendrick, R., & Johnson, M. (2004). Effects of noun phrase type on sentence complexity. *Journal of Memory and Language*, *51*, 97–114. 32, 37, 73

Gordon, P. C., Hendrick, R., Johnson, M., & Lee, Y. (2006). Similarity-based interference during language comprehension: Evidence from eye-tracking during reading. *Journal of experimental psychology: Learning, Memory, and Cognition*, *32*(6), 1304–1321. 32, 37

Gordon, P. C., Hendrick, R., & Levine, W. H. (2002). Memory-load interference in syntactic processing. *Psychological Science*, *13*(5), 425–430. 31, 32, 37

Grodner, D. J., & Gibson, E. A. F. (2005). Consequences of the serial nature of linguistic input for sentential complexity. *Cognitive Science*, *29*, 261-91. 38, 73, 109

Grove, K. (2011). Why unaccusatives have it easy: Reduced relative garden path effects and verb types. *University of Pennsylvania Working Papers in Linguistics*, *17*(1). 43

Gundel, J. K., Hedberg, N., & Zacharski, R. (1993). Cognitive status and the form of referring expressions in discourse. *Language*, *69*, 274–307. 35

Haider, H. (2004). The superiority conspiracy: four constraints and a processing effect. In A. Stepanov, G. Fanselow, & R. Vogel (Eds.), *Minimality effects in syntax* (pp. 147–176). Berlin: Mouton de Gruyter. 84, 97, 140, 155

Hakuta, K. (1981). Grammatical description versus configurational arrangement in language acquisition: the case of relative clauses in Japanese. *Cognition*, *9*, 197–236. 30

Hale, J. T. (2001). A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of NAACL 2001* (pp. 1–8). 31, 43, 62, 68

Hale, J. T. (2003). *Grammar, uncertainty, and sentence processing.* Unpublished doctoral dissertation, Johns Hopkins University. 43

Hall, K. (2007). k-best spanning tree parsing. In *Proceedings of the 45th annual meeting of the Association for Computational Linguistics (ACL)* (pp. 392–399). Prague, Czech Republic. 45

Hall, K., Havelka, J., & Smith, D. A. (2007). Log-linear models of non-projective trees, k-best MST parsing and tree-ranking. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)* (pp. 962–966). Prague, Czech Republic. 45

Hawkins, J. A. (1990). A parsing theory of word order universals. *Linguistic Inquiry*, *21*(2), 223–261. 32, 33

Hawkins, J. A. (1999). Processing complexity and filler-gap dependencies across grammars. *Language*, *75*(2), 244–285. 29, 33

Hays, D. G. (1964). Dependency Theory: A formalism and some observations. *Language*, *40*, 511-525. 11, 12, 45

Hofmeister, P. (2007). Retrievability and gradience in filler-gap dependencies. In *Proceedings*

of the 43rd Regional Meeting of the Chicago Linguistics Society. Chicago: University of Chicago Press. 2, 27, 28, 29, 32, 34, 36, 104

Hofmeister, P., Jaeger, T. F., Sag, I. A., Arnon, I., & Snider, N. (2007). Locality and accessibility in *wh*-questions. In S. Featherston & W. Sternefeld (Eds.), *Roots: Linguistics in search of its evidential base* (pp. 185–206). Berlin: Mouton de Gruyter. 2, 28, 29, 37

Hofmeister, P., & Sag, I. A. (2010). Cognitive constraints and island effects. *Language*, *86*(2), 366–415. 2, 3, 17, 18, 19, 23, 29, 30, 32, 39, 47, 86, 87, 88, 93, 94, 104, 123, 126, 127, 128, 133, 134, 139, 150, 154, 160, 163

Hudson, R. (n.d.). *Word grammar*. Oxford: Blackwell. 13

Johansson, R., & Nugues, P. (2007, May 25-26). Extended constituent-to-dependency conversion for English. In *Proceedings of NODALIDA 2007.* Tartu, Estonia. 55

Joshi, A. K. (1990). Processing crossing and nested dependencies: an automaton perspective on the psycholinguistic results. *Language and Cognitive Processes*, *5*, 1–27. 32, 33

Joshi, A. K., Levy, L. S., & Takahashi, M. (1975). Tree adjunct grammars. *Journal of computer and system sciences*, *10*(1), 136–163. 162

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: from eye fixations to comprehension. *Psychological Review*, *87*(4), 329–354. 72

Just, M. A., & Carpenter, P. A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review*, *98*, 122–149. 32, 73, 167, 169

Just, M. A., & Varma, S. (2002). A hybrid architecture for working memory: Reply to MacDonald and Christiansen. *Psychological Review*, *109*, 55–65. 32, 73, 169

Just, M. A., & Varma, S. (2007). The organization of thinking: What functional brain imaging reveals about the neuroarchitecture of complex cognition. *Cognitive, Affective, and Behavioral Neuroscience*, *7*(3), 153–191. 167

Kahane, S., Nasr, A., & Rambow, O. (1998). Pseudo-projectivity: A polynomially parsable non-projective dependency grammar. In *Proceedings of ACL-COLING.* 48

Kaplan, R. M. (1972). Augmented transition networks as psychological models of sentence

comprehension. *Artificial Intelligence*, *3*, 77–100. 45

Karttunen, L. (1977). Syntax and semantics of questions. *Linguistics and Philosophy*, *1*, 3–44. 20, 21, 22, 98

Keller, F. (1996). How do humans deal with ungrammatical input? Experimental evidence and computational modeling. In D. Gibbon (Ed.), *Natural language processing and speech technology: Results of the 3rd KONVENS conference* (pp. 27–34). Berlin: Mouton de Gruyter. 17, 88, 104, 127, 128, 150

Keller, F. (2010). Cognitively plausible models of human language processing. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics: Short Papers* (pp. 60–67). Uppsala. 7, 61, 79, 114

Kimball, J. (1973). Seven principles of surface structure parsing in natural language. *Cognition*, *2*, 15–47. 30, 45, 165

Kiss, K. (1993). Wh-movement and specificity. *Natural language and linguistic theory*, *11*(1), 85–120. 25

Kitahara, H. (1993). Deducing “superiority” effects from the Shortest Chain Requirement. In *Harvard working papers in linguistics* (pp. 109–119). Cambridge, MA: Harvard University Department of Linguistics. 104

Kluender, R. (1992). Deriving island constraints from principles of predication. In H. Goodluck & M. Rochemont (Eds.), *Island constraints: Theory, acquisition, and processing* (pp. 223–258). Dordrecht: Kluwer. 17, 29, 88

Kluender, R. (1998). On the distinction between strong and weak islands: A processing perspective. *Syntax and Semantics*, *29*, 241–279. 17, 29, 159, 163, 164, 171

Kluender, R., & Kutas, M. (1993). Subjacency as a processing phenomenon. *Language and cognitive processes*, *8*, 573–633. 29

Kroch, A. (1989). Amount quantification, referentiality, and long *wh*-movement. *U. Penn Working Papers in Linguistics*, *5*(2), 21–36. 27

Kuhlmann, M. (2007). *Dependency structures and lexicalized grammars*. Unpublished doctoral

dissertation, Universität des Saarlandes. 44

Kush, D., & Lindahl, F. (2011). *On the escapability of islands in Scandinavian.* (Talk given at 2011 Linguistics Society of America Annual Meeting, Pittsburgh.) 91, 92, 104, 131, 132

Legendre, G., Wilson, C., Smolensky, P., Homer, K., & Raymond, W. (2006). Wh-questions. In P. Smolensky & G. Legendre (Eds.), *The harmonic mind: from neural computation to optimality-theoretic grammar* (pp. 183–230). Cambridge, MA: MIT Press. 15

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177. 43, 44

Levy, R., Fedorenko, E., Breen, M., & Gibson, T. (Submitted). *The processing of extraposed structures in english.* 71

Lewis, R. L. (1993). *An architecturally-based theory of human sentence comprehension*. Unpublished doctoral dissertation, Carnegie Mellon University. 30, 167, 170

Lewis, R. L. (1996). Interference in short-term memory: the magical number two (or three) in sentence processing. *Journal of Psycholinguistic Research*, *25*(1), 93–113. 30, 31, 32, 167, 170

Lewis, R. L. (1999). Specifying architectures for language processing: Process, control, and memory in parsing and interpretation. In M. Crocker, M. Pickering, & C. Clifton Jr (Eds.), *Architectures and mechanisms for language processing* (pp. 56–89). Cambridge, England: Cambridge University Press. 30, 31, 32, 37, 73, 167, 170

Lewis, R. L., & Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science*, *29*, 1–45. 3, 5, 30, 32, 34, 35, 36, 37, 38, 39, 61, 64, 66, 67, 68, 72, 73, 74, 75, 76, 116, 159, 167, 170

Lewis, R. L., Vasishth, S., & Van Dyke, J. (2006). Computational principles of working memory in sentence comprehension. *Trends in Cognitive Sciences*, *10*(10), 447-454. 32, 35, 36, 37, 39

Lin, C.-J., Weng, R. C., & Keerthi, S. S. (2008). Trust region newton method for large-scale regularized logistic regression. *Journal of Machine Learning Research*, *9*, 627–650. 59

Maling, J. (1978). An asymmetry with respect to *wh*-islands. *Linguistic Inquiry*, *9*(1), 75–89. 104

Maling, J., & Zaenen, A. (1982). A phrase structure account of Scandinavian extraction phenomena. In P. Jacobson & G. K. Pullum (Eds.), *The nature of syntactic representation* (pp. 229–282). Dordrecht: D. Reidel Publishing Company. 96, 97

Manning, C., & Schütze, H. (1999). *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press. 79

Manzini, M. R. (1992). *Locality: a theory and some of its consequences* (Vol. 19). Cambridge, MA: MIT Press. 158

Manzini, M. R. (1994). Syntactic dependencies and their properties: strong islands. In *UCL working papers in linguistics* (Vol. 6). London: University College London. 158

Marcus, M. P. (1980). *A theory of syntactic recognition for natural language*. Cambridge, MA: MIT Press. 45, 165, 167, 170

McDonald, R., Pereira, F., Ribarov, K., & Hajiĉ, J. (2005). Non-projective dependency parsing using spanning tree algorithms. In *Proceedings of HLT and EMNLP* (pp. 523–530). 45

McElree, B., Foraker, S., & Dyer, L. (2003). Memory structures that subserve sentence comprehension. *Journal of Memory and Language*, *48*, 67–91. 31, 32, 35, 36, 73

Mel'čuk, I. (1988). *Dependency syntax: Theory and practice*. Albany, NY: State University of New York Press. 13

Miller, G., & Chomsky, N. (1963). Finitary models of language users. In R. Luce, R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (pp. 419–491). New York, NY: John Wiley. 109, 110

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *The Psychological Review*, *63*, 81–97. 30

Milward, D. (1994). Dynamic dependency grammar. *Linguistics and philosophy*, *17*, 561–605. 45

Nivre, J. (2004). Incrementality in deterministic dependency parsing. In *Proceedings of the*

workshop on incremental parsing (ACL) (p. 50-57). 46, 47, 49

Nivre, J. (2006). *Inductive dependency parsing*. Dordrecht: Springer. 46, 49

Nivre, J. (2008). Sorting out dependency parsing. In A. Ranta (Ed.), *GoTAL 2008* (Vol. 5221, pp. 16–27). Berlin: Springer-Verlag. 46, 47, 49

Nivre, J. (2009). Non-projective dependency parsing in expected linear time. In *Proceedings of the 47th annual meeting of the ACL and the 4th IJCNLP of the AFNLP* (pp. 351–359). Suntec, Singapore. 8, 46, 49, 50, 53, 59

Nivre, J., Boguslavsky, I. M., & Iomdin, L. L. (2008). Parsing the SynTagRus Treebank of Russian. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING 2008)* (pp. 641–648). Manchester, UK. 54

Nivre, J., Hall, J., Nilsson, J., Chanev, A., Eryiğit, G., Kübler, S., et al. (2007). MaltParser: a language-independent system for data-driven dependency parsing. *Natural Language Engineering*, *13*(2), 95–135. 46

Nivre, J., & Nilsson, J. (2005). Pseudo-projective dependency parsing. In *Proceedings of ACL-COLING* (pp. 99–106). 48

Nivre, J., Nilsson, J., & Hall, J. (2006). Talbanken05: A Swedish treebank with phrase structure and dependency annotation. In *Proceedings of Language Resources and Evaluation Conference (LREC)* (pp. 24–26). 53

Nivre, J., Rimell, L., McDonald, R., & Gómez-Rodríguez, C. (2010). Evaluation of dependency parsers on unbounded dependencies. In *Proceedings of the 23rd International Conference on Computational Linguistics* (pp. 833–841). Stroudsburg, PA, USA: Association for Computational Linguistics. 6, 80

Pesetsky, D. (1987). Wh-in-situ: movement and unselective binding. In E. Reuland & A. ter Meulen (Eds.), *The representation of (in)definiteness* (pp. 98–129). Cambridge, MA: MIT Press. 24

Pesetsky, D. (2000). *Phrasal movement and its kin*. Cambridge, MA: MIT Press. 22

Phillips, C. (2006). The real-time status of island phenomena. *Language*, *82*, 795–823. 32

Phillips, C. (In Press). Some arguments and non-arguments for reductionist accounts of syntactic phenomena. *Language and Cognitive Processes*. 4, 29, 160, 172

Pickering, M. J., Barton, S., & Shillcock, R. (1994). Unbounded dependencies, island constraints and processing complexity. In J. Charles Clifton, L. Frazier, & K. Rayner (Eds.), *Perspectives on sentence processing* (pp. 199–224). London: Lawrence Erlbaum. 30, 32

Pritchett, B. L. (1992). Parsing with grammar: Islands, heads, and garden paths. In H. Goodluck & M. Rochemont (Eds.), *Island constraints: Theory, acquisition and processing* (pp. 321–349). Dordrecht: Kluwer. 29, 167, 168

Pritchett, B. L. (1993). Subjacency in a principle-based parser. In R. C. Berwick (Ed.), *Principle-based parsing: Computation and psycholinguistics* (pp. 301–345). Dordrecht: Kluwer Academic Publishers. 167, 168, 170

Rambow, O., & Joshi, A. K. (1994). A processing model for free word-order languages. In C. Clifton, Jr., L. Frazier, & K. Rayner (Eds.), *Perspectives on sentence processing* (pp. 267–301). Hillsdale, NJ: Erlbaum. 32, 33

Rayner, K., Kambe, G., & Duffy, S. A. (2000). The effect of clause wrap-up on eye movements during reading. *The quarterly journal of experimental psychology*, *53A*(4), 1061–1080. 72

Reinhart, T. (1976). *The syntactic domain of anaphora*. Unpublished doctoral dissertation, Massachusetts Institute of Technology. 18

Reinhart, T. (1995). *Interface strategies: optimal and costly computations*. Cambridge, MA: MIT Press. 104

Rizzi, L. (1990). *Relativized minimality*. MIT Press. 2, 5, 20, 25, 158, 172

Roark, B. (2001). *Robust probabilistic predictive syntactic processing: motivations, models, and applications*. Unpublished doctoral dissertation, Brown University. 43

Ross, J. R. (1967). *Constraints on variables in syntax*. Unpublished doctoral dissertation, MIT. 2, 14, 15, 18, 20, 158

Sag, I. A., Hofmeister, P., Arnon, I., Snider, N., & Jaeger, T. F. (2008). *Processing accounts of superiority affects.* (Submitted) 22

Satta, G. (1992). Recognition of linear context-free rewriting systems. In *Proceedings of the Association for Computational Linguists (ACL)* (pp. 89–95). 47

Schuler, W., AbdelRahmen, S., Miller, T., & Schwartz, L. (2010). Broad-coverage parsing using human-like memory constraints. *Computational Linguistics*, *36*(1), 1–30. 7, 30

Sprouse, J., Wagers, M., & Phillips, C. (To Appear). *A test of the relation between working memory capacity and syntactic island effects.* 90, 96, 113, 129, 136, 152, 154

Stabler, E. P. (1994). The finite connectivity of linguistic structures. In Charles Clifton Jr., L. Frazier, & K. Rayner (Eds.), *Perspectives on sentence processing* (pp. 303–336). Hillsdale, NJ: Erlbaum. 30

Stabler, E. P. (1997). Derivational minimalism. In *Proceedings of Logical Aspects of Computational Linguistics (LACL 1996)* (pp. 68–95). 4, 162

Stevens, S. S. (1975). *Psychophysics: Introduction to its perceptual, neural, and social prospects*. New York, NY: John Wiley. 85

Stevenson, S. (1994). Competition and recency in a hybrid network model of syntactic disambiguation. *Journal of Psycholinguistic Research*, *23*(4), 295–322. 33

Stowe, L. (1986). Evidence for on-line gap location. *Language and cognitive processes*, *1*, 227–245. 5, 30, 32, 104

Szabolcsi, A., & den Dikken, M. (2002). Islands. In Lisa Lai Shen Cheng & R. P. E. Sybesma (Eds.), *The second GLOT international state of the article book* (pp. 123–241). Berlin: Mouton de Gruyter. 15, 17

Szabolcsi, A., & Zwarts, F. (1993). Weak islands and an algebraic semantics for scope taking. *Natural language semantics*, *1*, 235–284. 26, 27, 140, 158

Tabor, W., Juliano, C., & Tanenhaus, M. (1997). Parsing in a dynamical system: An attractor-based account of the interaction of lexical and structural constraints in sentence processing. *Language and Cognitive Processes*, *12*(2/3), 211–271. 33, 34, 43, 169

Tabor, W., & Tanenhaus, M. K. (1999). Dynamical models of sentence processing. *Cognitive Science*, *23*(4), 491–515. 43, 169

Tesnière, L. (1959). *Éléments de syntaxe structurale*. Editions Klincksiek. 11, 12

Van Dyke, J., & Lewis, R. L. (2003). Distinguishing effects of structure and decay on attachment and repair: A cue-based parsing account of recovery from misanalyzed ambiguities. *Journal of Memory and Language*, *49*, 285–316. 32, 37

Van Dyke, J. A., & McElree, B. (2006). Retrieval interference in sentence comprehension. *Journal of Memory and Language*, *55*, 157–166. 32, 37

Vasishth, S., Brüssow, S., Lewis, R. L., & Drenhaus, H. (2008). Processing polarity: How the ungrammatical intrudes on the grammatical. *Cognitive Science*, *32*(4). 39

Vasishth, S., Brüssow, S., Lewis, R. L., & Drenhaus, H. (2008). Processing polarity: How the ungrammatical intrudes on the grammatical. *Cognitive Science*, *32*(4), 685–712. 73

Vosse, T., & Kempen, G. (2000). Syntactic structure assembly in human parsing: A computational model based on competitive inhibition and a lexicalist grammar. *Cognition*, *75*, 105–143. 32, 167, 169

Wanner, E., & Maratsos, M. (1978). An ATN approach in comprehension. In M. Halle, J. Bresnan, & G. Miller (Eds.), *Linguistic theory and psychological reality* (pp. 119–161). Cambridge, MA: MIT Press. 3, 32, 33, 165

Warren, T., & Gibson, E. (2002). The influence of referential processing on sentence complexity. *Cognition*, *85*, 79–112. 32, 37, 73

Yamada, H., & Matsumoto, Y. (2003). Statistical dependency analysis with support vector machines. In *Proceedings of the 8th International Workshop on Parsing Technologies (IWPT)* (pp. 195–206). Nancy, France. 45, 59

Yngve, V. H. (1960). A model and an hypothesis for language structure. *Proceedings of the American Philosophical Society*, *104*(5), 444–466. 30

Yoshida, M. (2006). *Constraints and mechanisms in long-distance dependency formation*. Unpublished doctoral dissertation, University of Maryland, College Park. 17