

COGNITIVE MODELING OF HINDI SYNTACTIC CHOICE PHENOMENA

SIDHARTH RANJAN



AMAR NATH & SHASHI KHOSLA SCHOOL OF
INFORMATION TECHNOLOGY
INDIAN INSTITUTE OF TECHNOLOGY DELHI
MAY 2023

© Indian Institute of Technology Delhi (IITD), New Delhi, 2023

COGNITIVE MODELING OF HINDI SYNTACTIC CHOICE PHENOMENA

by

SIDHARTH RANJAN

Amar Nath & Shashi Khosla School of Information Technology

Submitted

in fulfillment of the requirements for the degree of
Doctor of Philosophy

to the



Indian Institute of Technology Delhi
May 2023

CERTIFICATE

This is to certify that the thesis entitled **COGNITIVE MODELING OF HINDI SYNTACTIC CHOICE PHENOMENA** submitted by **Mr. SIDHARTH RANJAN** to the **Indian Institute of Technology Delhi** for the award of **Doctor of Philosophy** is a record of the original bona fide research work carried out by him under our supervision and guidance. The thesis has reached the standards fulfilling the requirements of the regulations relating to the degree.

The work presented in this thesis has not been submitted elsewhere, either in part or full to any other University or Institute for the award of any other degree or diploma.

New Delhi

2023

Dr. Sumeet Agarwal

Advisor

Associate Professor

Department of Electrical Engineering

Indian Institute of Technology Delhi

New Delhi

Dr. Rajakrishnan Rajkumar

Co-Advisor

Assistant Professor

Department of Humanities and Social Sciences

IISER Bhopal

Bhopal

ACKNOWLEDGMENTS

In this acknowledgment section of my dissertation work which makes a tiny attempt to investigate *how humans process, comprehend, and produce a language*, I, Sidharth Ranjan, would like to take this opportunity to express my sincere gratitude for the invaluable teaching, consistent guidance, and support provided by my wonderful PhD advisors, Professor Sumeet Agarwal and Professor Rajakrishnan Rajkumar, who were also my master's thesis advisors at IIT Delhi. As I write these words, I feel incredibly nostalgic, realizing that *time sure flies when you are having fun!* It feels like it was just a few days ago I approached both of them, asking if they would be willing to consider me their PhD student extending my master's thesis work at IIT Delhi. And the rest is history!

I would like to express my appreciation to my former teachers from the Linguistics unit at IIT Delhi for their invaluable contribution. Their captivating classroom lectures served as an inspiration for me to pursue a master's thesis in linguistics alongside my undergraduate training in Electrical Engineering. Additionally, I am immensely grateful to my former teachers from my Electrical Engineering study program, especially Professors Subrat Kar, Niladri Chatterjee, and Raghunath K. Shevgaonkar, for their unwavering encouragement in pushing my limits. It would be unfair if I did not acknowledge the two completely different and amazing academic environments that I had the rare privilege of being a part of during my PhD research, namely engineering and technology at IIT Delhi, and pure sciences at IISER Bhopal. The entire credit for this arrangement goes to my visionary PhD advisors, who facilitated my stay at both institutes, especially after Prof. Rajakrishnan moved to IISER Bhopal midway through my PhD research.

I extend my heartfelt thanks to the administration of my home department, the School

of Information and Technology (SIT) at IIT Delhi, as well as the administration of IISER Bhopal, for granting approval for my academic stay at both institutions. Additionally, I would like to express my sincere gratitude to the Department of Humanities and Social Sciences (HSS) at IISER Bhopal for graciously hosting me during all my visits to the institute. My stay at IISERB was fruitful in pursuing my parallel interest in Digital Humanities, where I had the chance to collaborate with Professors Kushal Shah, Arpit Sharma, and Rajakrishnan and applied my insights from computer science and literature for genre studies and literary analysis, even though the two published papers on this theme are not a part of my PhD thesis. The time I spent at IISER Bhopal was also instrumental in shaping my newfound interest in Evolution and Biology, which I aspire to integrate into my future linguistics research. The Evolutionary Intelligence course co-taught by Professors Kushal Shah, Rajakrishnan, and Nagarjun Vijay, was truly eye-opening in this regard.

Special thanks go to Professor Marten van Schijndel for hosting me in his C.Psyd research group at Cornell University. I enjoyed collaborating with him on two papers that formed the basis of the priming research presented in Chapter 5. I am also grateful to the members of C.Psyd group for their invaluable comments and feedback on the reading aloud part of the work in Chapter 7. Additionally, I express my gratitude to Professor Titus von der Malsburg for providing invaluable feedback and suggestions on my AMLaP-2022 poster, which were instrumental in the silent reading part of the work in Chapter 7. I would also like to thank my PhD dissertation committee members, Professors Mausam and Samar Husain, for their invaluable time, interest, and numerous profound queries and suggestions during several thesis progress presentations that helped improve this work. Many thanks to Professor Rahul Garg as well for his suggestions at times. I am grateful to Professors KP Mohanan, Shravan Vasishth, Whitney Tabor, and Silvia Gennari for their astute comments and invaluable feedback on my Cognition-2022 journal article, which served as the foundation of the work presented in Chapter 4.

I am appreciative of each and every person in the two research groups I have had the good fortune to be a part of: LINCOS group at IIT Delhi and the Psycholinguistics group at IISER Bhopal throughout my PhD tenure. Their company has been a striking source

of knowledge and inspiration for me during my stay at each place. At IIT Delhi, many thanks to Tanya Bhatnagar for exciting discussions on arboreal operations in Hindi syntax and her help in developing the tree conversion pipeline deployed in the first part of the work in Chapter 5. Thanks to Ayush Jain and Vishal Singh for collaborating on the third part of the work in Chapter 6. Thanks to Samvit Dammalapati, with whom I had a great time working and collaborating on speech disfluency research, even though the published paper on this theme is not a part of my PhD thesis. At IISER Bhopal, many thanks to Rameez Qureshi, Antony James, and Arman Kazmi, with whom I had a wonderful time working on genre research and literary studies. Thanks to Rupesh Pandey for the logistical help in human data collection conducted in Chapters 4 and 5. Thanks to Ruchira Sakalle, Roshin AR, and Adarsh Tayade for their companionship and discussions in the lab.

My PhD journey would be incomplete without acknowledging all the excellent instructors of the UG and PG courses who I had the opportunity to work with as a teaching assistant (TA) at both the institutes. I strongly believe that TA duties play a vital role in academic and teaching training, and working with various instructors has helped me realize the hard work and advance preparation required for quality teaching and student training. Along with my PhD advisors, I would like to thank Professors Antara Chatterjee, Adity Singh, and Kushal Shah at IISER Bhopal. I thank Professors Rohan Paul, Jayadeva, Saroj Kaushik, and Sumitava Mukherjee at IIT Delhi.

Now, with this part of acknowledgment, I would like to celebrate life through friendship, love, and family, which ensured that I was sound and ticking during this entire process of attaining knowledge through solitude. Thanks to Rijul Soans, Chetan Ralekar, Preeti Kumari, Wasim Odud, Nikhilesh Thirukovela, Vinayak Gupta, and Dishant Goyal for being fantastic colleagues and friends at IIT Delhi. A sincere thanks to Sakshi for her insightful comments and discussion on Hindi grammar whenever I felt stuck in deciding my own linguistic intuitions. My stay at IISER Bhopal would have been utterly monotonous if I had not met amazing friends there. Special thanks to Gowri Devi, Saumya Gupta, Aswin Soman, Shad Ali Khan, Purbita Das, Dolon Sarkar, and Natasha Negi. At this point, I would also like to thank my parents and siblings for their consistent love and support that

inspired me to pursue this dream. I am glad to follow in the footsteps of my mother, who is the first doctor in our family and has been a constant source of inspiration to me. Soon, I will join her as the second doctor in the family, continuing the legacy she has established.

Towards the end, I would also like to acknowledge the extramural funding received from various funding agencies that supported this work: a) Cognitive Science Research Initiative, Department of Science and Technology, Government of India (DO: DST/CSRI/2018/263) b) Funding for 1 year SRF fellowship at IISER Bhopal via Initiation Grant (Project No. INST/HSS/2018096) c) Travel Grants: Microsoft and Google Research India Travel Award, IIT Delhi Research Scholar Travel Award, School of IT Research Scholar Travel Award, and French National Centre for Scientific Research Travel Award. Sincere thanks to the office staffs Rajesh Kumar and Suresh Mourya at SIT IIT Delhi and Santosh Thakur, Harish, and Pravesh Malviya at HSS IISER Bhopal for taking care of all the approvals, logistics, and finance related matters during the course of my PhD research.

Sidharth Ranjan

Sidharth Ranjan



This thesis is dedicated to my loving and caring parents.

ABSTRACT

Sentence processing research has primarily focused on identifying the cognitive factors that influence comprehension and production difficulty in a sentence while acknowledging that languages offer diverse means of conveying identical concepts (Ferreira, 1996). However, certain syntactic structures consistently emerge as more a favourable choice than the others (Mohanani and Mohanani, 1994). A number of factors are known to influence such syntactic preferences, including information structure, syntactic complexity, and inherent human cognitive capacities *viz.*, limited working memory and information decay. However, the relative impact of these factors on syntactic ordering choices in Hindi is not well understood. Hindi is a verb final language (SOV) exhibiting flexible word ordering patterns and belongs to the Indo-European language family. This thesis focuses on studying linguistic structures and developing computational cognitive models to explore the cognitive biases and processing mechanisms involved in preverbal constituent ordering in Hindi, with the specific goal of understanding language comprehension and production.

To begin with, this work constructs a framework to artificially generate meaning-equivalent grammatical variants of Hindi sentences by linearizing preverbal constituents of the projective dependency trees in the Hindi-Urdu Treebank corpus (Bhatt et al., 2009). Thereafter, the relative impact of various influential theories of language processing in Psycholinguistics, *viz.*, Dependency Locality Theory (Gibson, 1998, 2000), Surprisal Theory (Hale, 2001; Levy, 2008), Expectation Adaptation (Fine et al., 2013), Production-Distribution-Comprehension Theory (MacDonald, 1999, 2013), Interference Theory (Lewis, 1996; Van Dyke, 2002), and finally, Uniform Information Density Hypothesis (Jaeger, 2010; Levy and Jaeger, 2007) are investigated in a serial order to predict the ordering preferences

(reference vs. variants) using a logistic regression model.

The results indicate that while dependency length is a significant factor, its predictive power for Hindi syntactic choice becomes weak in the presence of information status and expectation-based predictors. Notably, trigram surprisal significantly outperforms both dependency length and parser surprisal by a considerable margin, highlighting the primary influence of maximizing lexical predictability in determining preverbal constituent ordering choices in Hindi. After incorporating discourse information into the calculation of surprisal, the aforementioned effect became more pronounced, thus further establishing the influence of discourse predictability in shaping Hindi word-order preferences. We situate our findings within the context of earlier studies on adaptation/priming in comprehension (Fine and Jaeger, 2016; Fine et al., 2013) and production (Bock, 1986a; Gries, 2005). With respect to PDC, our results suggest that Hindi optimizes for processing efficiency in terms of accessibility and minimizes similarity-based interference by avoiding identical case marker repetition. Furthermore, we also found that similarity-based interference significantly predicts dependency length, supporting the view in the literature that the mechanisms underlying locality may be driven by memory interference (Vasishth, 2011). In relation to the efficacy of UID on Hindi word order choices, our results indicate that UID measures are not a significant factor in predicting corpus sentences in the presence of competing control predictors such as lexical surprisal, and do not support a theory of word order based solely on UID. Finally, we discuss the implications of our findings for theories of language production.

Towards the end, this work proposes a theoretically motivated novel metric, FORWARD SURPRISAL (*i.e.*, probability of target word given the two words in the upcoming context) to account for planning processes in both the comprehension and production systems beyond the most commonly used surprisal metric estimated using preceding context. Consequently, this work vouches for an integrated model of the comprehension and production processes which are traditionally considered distinct cognitive systems in the psycholinguistics literature (Pickering and Garrod, 2013).

सार

वाक्य प्रसंस्करण अनुसंधान ने मुख्य रूप से उन संज्ञानात्मक कारकों की पहचान करने पर ध्यान केंद्रित किया है जो एक वाक्य में समझ और उत्पादन कठिनाई को प्रभावित करते हैं, यह स्वीकार करते हुए कि भाषाएं समान अवधारणाओं को व्यक्त करने के विविध साधन प्रदान करती हैं (फरेरा, 1996)। हालांकि, कुछ सिंटेक्टिक संरचनाएं लगातार दूसरों की तुलना में अधिक अनुकूल विकल्प के रूप में उभरती हैं (मोहनन और मोहनन, 1994)। सूचना संरचना, वाक्यात्मक जटिलता, और निहित मानव संज्ञानात्मक क्षमता, सीमित कार्यशील स्मृति और सूचना क्षय सहित ऐसी वाक्यात्मक प्राथमिकताओं को प्रभावित करने के लिए कैड कारक जाने जाते हैं। हालांकि, हिंदी में वाक्य-विन्यास के विकल्पों पर इन कारकों के सापेक्ष प्रभाव को अच्छी तरह से नहीं समझा गया है। हिंदी एक क्रिया अंतिम भाषा है जो लचीले शब्द क्रम पैटर्न को प्रदर्शित करती है और इंडो-यूरोपीय भाषा परिवार से संबंधित है। यह थिसिस भाषा की समझ और उत्पादन को समझने के विशिष्ट लक्ष्य के साथ हिंदी में पूर्ववर्ती घटक क्रम में शामिल संज्ञानात्मक पूर्वाग्रहों और प्रसंस्करण तंत्रों का पता लगाने के लिए भाषाई संरचनाओं का अध्ययन करने और कम्प्यूटेशनल संज्ञानात्मक मॉडल विकसित करने पर केंद्रित है।

आरंभ करने के लिए, यह काम हिंदी-उर्दू ट्रीबैंक कॉर्पस (भट्ट एट अल, 2009) में प्रोजेक्टिव डिपेंडेंसी ट्री के प्रीवर्बल घटकों को रैखिक करके हिंदी वाक्यों के समकक्ष व्याकरणिक रूपों को कृत्रिम रूप से उत्पन्न करने के लिए एक रूपरेखा तैयार करता है। इसके बाद, भाषा की समझ के विभिन्न प्रभावशाली सिद्धांतों का सापेक्ष प्रभाव, जैसे, निर्भरता स्थानीयता सिद्धांत (गिब्सन, 1998, 2000), आश्चर्य सिद्धांत (हेल, 2001; लेवी, 2008), अपेक्षा अनुकूलन (फाइन एट अल, 2013): प्रोडक्शन-डिस्ट्रीब्यूशन-कॉम्प्रिहेंशन थ्योरी (मैकडॉनल्ड, 1999, 2013), इंटरफेरेंस थ्योरी (लुईस, 1996; वैन डाइक, 2002), और अंत में, यूनिफॉर्म इंफॉर्मेशन डेंसिटी हाइपोथिसिस (जैगर, 2010; लेवी और जैगर, 2007) की जांच लॉजिस्टिक रिग्रेसन मॉडल का उपयोग करके अनुक्रम वरीयताओं (संदर्भ बनाम वेरिएंट) की भविष्यवाणी की जाती है।

परिणामों से संकेत मिलता है कि जहां निर्भरता की लंबाई एक महत्वपूर्ण कारक है, वहीं सूचना की स्थिति और अपेक्षा-आधारित भविष्यवक्ताओं की उपस्थिति में हिंदी वाक्य-विन्यास पसंद के लिए इसकी भविष्यवाणी शक्ति कमजोर हो जाती है। विशेष रूप से, ट्रिग्राम सरप्रिसल डिपेंडेंसी लेंथ और पार्सर सरप्रिसल दोनों को काफी अंतर से बेहतर प्रदर्शन करता है, जो हिंदी में प्रीवर्बल घटक अनुक्रम विकल्पों को निर्धारित करने में लेक्सिकल भविष्यवाणी को अधिकतम करने के प्राथमिक प्रभाव को उजागर करता है। आश्चर्य की गणना में प्रवचन की जानकारी को शामिल करने के बाद, उपरोक्त प्रभाव अधिक स्पष्ट हो गया, इस प्रकार हिंदी शब्द-क्रम वरीयताओं को आकार देने में प्रवचन की भविष्यवाणी के प्रभाव को और स्थापित किया गया। हम अनुकूलन प्राइमिंग पर पहले के अध्ययनों के संदर्भ में अपने निष्कर्षों को समझने में (फाइन एंड जैगर, 2016; फाइन एट अल, 2013) और प्रोडक्शन में (बॉक, 1986ए; ग्रिस, 2005) व्यवस्थित करते हैं।

पीडीसी के संबंध में, हमारे परिणाम बताते हैं कि हिंदी प्रसंस्करण दक्षता के लिए अनुकूलन करती है और समान केस मार्कर पुनरावृत्ति से बचकर समानता-आधारित हस्तक्षेप को कम करती है। इसके अलावा, हमने यह भी पाया कि समानता-आधारित हस्तक्षेप काफी हद तक निर्भरता की लंबाई की भविष्यवाणी करता है, साहित्य में इस दृष्टिकोण का समर्थन करता है कि तंत्र अंतर्निहित स्थानीयता स्मृति हस्तक्षेप (वशिष्ठ, 2011) द्वारा संचालित हो सकती है। हिंदी शब्द क्रम विकल्पों पर यूआईडी की प्रभावकारिता के संबंध में, हमारे परिणाम संकेत देते हैं कि यूआईडी शब्द क्रम का समर्थन नहीं करते हैं।

अंत में, यह काम एक सैद्धांतिक रूप से प्रेरित उपन्यास मीट्रिक, फॉरवर्ड सरप्रिसल (यानी, आगामी संदर्भ में दो शब्दों को दिए गए लक्ष्य शब्द की संभावना) का प्रस्ताव करता है, जो कि सबसे अधिक इस्तेमाल किए जाने वाले आश्चर्यजनक मीट्रिक से परे दोनों समझ और उत्पादन प्रणालियों में नियोजन प्रक्रियाओं के लिए है। पूर्ववर्ती संदर्भ का उपयोग करके अनुमान लगाया गया। नतीजतन, यह कार्य समझ और उत्पादन प्रक्रियाओं के एक एकीकृत मॉडल के लिए प्रतिज्ञा करता है, जिन्हें आमतौर पर मनोविज्ञान साहित्य (पिकरिंग और गारोड, 2013) में विशिष्ट संज्ञानात्मक प्रणाली माना जाता है।

Table of Contents

	Page
Certificate	i
Acknowledgments	ii
Abstract	vi
List of Figures	xiv
List of Tables	xix
1 Introduction	1
1.1 Contributions and Outlook	2
1.2 Thesis Overview	7
Chapters	
2 Background	8
2.1 Decay and Capacity Approaches	8
2.1.1 Dependency Locality Theory	10
2.2 Interference Theory	13
2.3 Information Theoretic Approaches	17
2.3.1 Surprisal Theory	17
2.3.2 Uniform Information Density	24
2.4 Factors influencing Hindi constituent ordering	27
2.4.1 Semantic Factors	27
2.4.2 Information Status Considerations	28
2.4.3 Prosody	31
2.5 Summary	31
3 Data and Methods	33
3.1 Data	33
3.1.1 Dependency Treebank	34
3.1.2 Constituency Treebank	34
3.1.3 Read-aloud speech corpus	35
3.1.4 Silent reading corpus	35
3.2 Methods	36
3.2.1 Variant Generation	36
3.2.2 Human Evaluation	38
3.2.3 Ranking Model	39

3.3	Cognitive Measures	41
3.4	Summary	43
4	Locality and Expectation Effects	45
4.1	Dependency length minimization in language processing	47
4.1.1	DLT and Comprehension	47
4.1.2	DLT and Syntactic Choice	48
4.1.3	Computational Simulations	49
4.2	Surprisal minimization in language processing	50
4.3	Data and Methods	55
4.4	Results	59
4.4.1	Corpus Study	59
4.4.2	Correlation and Regression Experiments	60
4.4.3	Prediction Accuracy Experiments	63
4.4.4	Construction-wise Experiments	71
4.5	Discussion	85
4.5.1	Low Impact of Dependency Locality	86
4.5.2	Success of Lexical Surprisal	88
4.5.3	Dependency Length and Case	90
4.5.4	Pronoun Placement	93
4.5.5	Relationship with Information Locality	93
4.5.6	Implications for Language Production	94
4.6	Conclusion	97
5	Syntactic and Discourse Context Effects	100
5.1	Constituency Treebank and Berkeley Parser Evaluation	100
5.1.1	Hindi Treebanks	102
5.1.2	Creating a Constituency Treebank	104
5.1.3	Parser Training and Evaluation	110
5.1.4	Evaluation of PCFG Syntactic Surprisal	118
5.1.5	Discussion	125
5.2	Discourse and Lexical Repetition Effects	125
5.2.1	Data and Methods	126
5.2.2	Regression Experiments	127
5.2.3	Prediction Experiments	130
5.2.4	Construction-wise Experiments	132
5.2.5	Predicting GIVENNESS	135
5.2.6	Discussion	137
5.3	Discourse Predictability Effects	138
5.3.1	Data and Method	141
5.3.2	Regression Analysis	143
5.3.3	Prediction Accuracy	145
5.3.4	Success of Adaptive LSTM Surprisal	148
5.3.5	What causes priming?	148
5.3.6	Human Evaluation	150
5.3.7	Discussion	151

5.4	Dual Mechanism Priming Effects	153
5.4.1	Data and Method	155
5.4.2	Verb-Specific Priming	156
5.4.3	Double Object construction	156
5.4.4	Success of Adaptive LSTM Surprisal	159
5.4.5	Conjunct Verb Construction	162
5.4.6	Success of Lexical Repetition Surprisal	163
5.4.7	Human Evaluation	164
5.4.8	Variance Inflation Factor Analysis	166
5.4.9	Discussion	169
6	Assessment of Production Theories	172
6.1	Production-Distribution-Comprehension Theory	172
6.1.1	Background	176
6.1.2	Data and Methods	178
6.1.3	Processing Efficiency Experiments	179
6.1.4	Case Markers and Processing Efficiency	183
6.1.5	Interference Experiments	186
6.1.6	Discussion	190
6.2	Interference Effects in Dependency Locality	191
6.2.1	Background	193
6.2.2	Data and Methods	196
6.2.3	Predicting dependency length	202
6.2.4	Construction-wise analysis	204
6.2.5	Predicting Semantic Similarity	206
6.2.6	Discussion	208
6.3	Uniform Information Density Theory	210
6.3.1	UID Measures	213
6.3.2	Data and Methods	214
6.3.3	Regression and Classification Experiments	215
6.3.4	UID and Syntactic constructions	220
6.3.5	Discussion	225
7	Modeling Aloud and Silent Reading Times	228
7.1	Reading Aloud	229
7.1.1	Background	231
7.1.2	Data and Methods	233
7.1.3	Predicting Reading Aloud Time	240
7.1.4	Forward Surprisal	241
7.1.5	Parafoveal Preview and Word Length Effects	241
7.1.6	Word Class and Duration	243
7.1.7	Word Class Prediction and PCFG Surprisal	244
7.1.8	Discussion	246
7.2	Silent Reading	249
7.2.1	Data and Methods	250
7.2.2	Predicting Eye-tracking Reading Times	255

7.2.3	Parafoveal Preview and Word Length Effects	257
7.2.4	Word Class and Duration	259
7.2.5	Word Class Prediction and PCFG Surprisal	262
7.2.6	Discussion	264
8	Conclusion	267
8.1	Key results	268
8.2	Minimization of Dependency length and Surprisal	269
8.3	Maximization of Discourse Predictability	271
8.4	Non-uniform Distribution of Information	273
8.5	Integrated model of Production and Comprehension	274
8.6	Future work	276
8.6.1	Written <i>vs</i> spoken Hindi	277
8.6.2	Hindi read-aloud speech corpus	277
8.6.3	Multi-modal Hindi corpus	278
8.6.4	Multi-construction Hindi corpus	278
8.6.5	Testable hypotheses	280
8.7	Summary	281
	Bibliography	283
	Appendix A Supplementary Materials	322
A.1	Locality and Expectation Effects	322
A.1.1	Dependency Length Calculation	322
A.1.2	Linguistic constructions in HUTB	324
A.1.3	Locality and Non-Locality Cases	324
A.1.4	Hindi Case Markers	326
A.1.5	Regression Results	326
A.1.6	Locality and Differential Case Marking	328
A.2	Constituency Treebank and Berkeley Parser Evaluation	329
A.3	Evaluation of PCFG surprisal on SPR data	335
A.3.1	Predicting Reading Time	337
A.3.2	Reading Time and Relative Clause Type	338
A.4	Uniform Information Density Effects on Syntactic Choice in English	340
A.4.1	Pairwise Classification for English Word-Order Choices	342
A.4.2	UID effects in English Constructions	346
A.4.3	Conclusion	353
	Vita	357
	Biography	359

List of Figures

Figure	Page
2.1 Subject Relative Clause (top; dependency length = 4) and Object Relative Clause (bottom; dependency length = 6)	11
2.2 Type of Interference Effects (Image adapted from Van Dyke and Johns (2012))	16
2.3 Syntactic Reduction: Optional <i>that</i> -complementizer (Image adapted from Jaeger (2010))	26
(a) <i>that</i> -complementizer insertion for UID (Eg. 12)	26
(b) <i>that</i> -complementizer elision for UID (Eg. 13)	26
2.4 Syntactic (<i>Constituent structure</i>) mapping with discourse elements [Image adapted from Butt and King (1996)]	29
(a) <i>c</i> -structure with discourse mapping	29
(b) <i>c</i> -structure with discourse mapping for Example 15	29
2.5 <i>f</i> -structure and <i>i</i> -structure [Image adapted from Butt and King (1996)]	29
(a) <i>f</i> -structure for Example 15	29
(b) <i>i</i> -structure for Example 15	29
3.1 HUTB dependency tree and relation labels	35
(a) Dependency tree	35
(b) Dependency relations	35
3.2 HUTB constituency tree and syntactic labels	36
(a) Constituency tree	36
(b) Constituency labels	36
4.1 Dependency length distribution for the Hindi-Urdu Treebank corpus	46
4.2 Dependency length and constituent ordering patterns for English head-medial and Japanese head-final structures ¹ (Overall dependency length (DL) of the structure indicated above each subfigure)	49
(a) Postverbal short-long order for a head-medial structure	49
(b) Postverbal long-short order for a head-medial structure	49
(c) Preverbal short-long order for a head-final structure	49
(d) Preverbal long-short order for a head-final structure	49
4.3 Hindi preverbal (20750 pairs) and postverbal (2599 pairs) constituent sequences	60

4.4	Scatterplot matrix showing correlations between predictors (Pearson’s coefficient of correlation also shown)	61
4.5	Prediction accuracy of surprisal (number of cases provided in parentheses; McNemar’s two-tailed significance against previous bar indicated using: $***p < 0.001$; $**p < 0.01$)	66
4.6	Example tree depicting identical case inflection <i>me</i> marking heads of two preverbal constituents	67
4.7	Classification accuracy for bins of dependency length difference (bin-wise number of data points in parentheses)	70
4.8	Example tree depicting a sentence with a conjunct verb	74
4.9	trigram surprisal profile	75
4.10	Dependency parser surprisal profile (case markers are excluded as they are represented as a feature on the preceding head noun during parsing)	76
4.11	Construction-wise classification accuracy (number of cases in parentheses; McNemar’s two-tailed significance against previous bar: $***p < 0.001$)	78
4.12	Bin-wise distribution of active and passive sentences (number of data points in parentheses)	80
4.13	Average case density and dependency length for the 25 most-frequent HUTB verbs (average dependency length and case density values for the entire dataset depicted as dotted lines parallel to X and Y axes respectively)	90
5.1	Example HUTB dependency tree and relation labels	103
	(a) Dependency tree	103
	(b) Dependency relations	103
5.2	Example HUTB constituency tree and syntactic labels	103
	(a) Constituency tree	103
	(b) Constituency labels	103
5.3	Handling Unary Constraints	105
	(a) Original tree	105
	(b) Transformed tree	105
5.4	Handling Empty Categories	106
	(a) Original tree	106
	(b) Transformed tree	106
5.5	Handling Split Roots	107
	(a) Original tree	107
	(b) Transformed tree	107
5.6	Removal of empty categories and Unary nodes	108
	(a) Generated PST	108
	(b) Modified PST	108
5.7	Handling split roots in coordination structures	109
	(a) Generated PST	109
	(b) Modified PST	109
5.8	Parsing Pipeline	112
5.9	Impact of POS tagging on parsing performance	114
5.10	Scatterplot matrix showing correlations between predictors (Pearson’s coefficient of correlation also shown)	120

5.11	Classification accuracy for bins of dependency length difference (bin-wise number of data points in parentheses)	122
5.12	Pearson correlation coefficients between predictors	128
5.13	Example HUTB dependency tree and relation labels	140
	(a) Dependency tree	140
	(b) Dependency relations	140
5.14	Pearson's coefficient of correlation between different pairs of predictors	143
5.15	Information profile for the reference-variant pair 32a and 32b	149
5.16	Information profiles for the reference-variant pair 34a and 34b	161
6.1	Example HUTB dependency tree	178
6.2	Mean trigram surprisal per sentence of reference and variant sentences (95% confidence intervals indicated)	180
6.3	Mean syntactic surprisal per sentence of reference and variant sentences (95% confidence intervals indicated)	181
6.4	Lexical surprisal profiles of the normal version of Hindi	184
6.5	Lexical surprisal profiles of the caseless artificial version of Hindi	185
6.6	Parafoveal preview in reading; adapted from Schotter et al. (2012)	197
6.7	HUTB Dependency Tree	198
6.8	Pearson's coefficient of correlation between different pairs of predictors	202
6.9	Correlation plot of predictors in HUTB corpus (158891 data points)	217
	(a) UID based on trigram surprisal	217
	(b) UID based on dependency parser surprisal	217
6.10	Mean of difference between mean predictor values of variants and reference sentences (95% classification intervals indicated)	219
	(a) Entire dataset (158891)	219
	(b) Active construction (143721)	219
	(c) Passive constructions (15170)	219
6.11	Information variation (lexical surprisal) in bits/word across reference-variant sentences shown in Example 44	223
6.12	Information variation (syntactic surprisal) in bits/word across a pair of reference-variant sentences shown in Example 44	224
7.1	DRC model of visual word recognition and reading aloud; adapted from Coltheart et al. (2001)	233
7.2	Integration and storage cost calculations for the sentence 'Saumya narrated a story to Ramesh', with head-dependent distance indicated above each dependency link; example sentence adapted from Husain et al. (2015)	235
7.3	Pearson's correlation coefficients amongst the different predictors and word duration	236
7.4	Word duration and information profiles of sentences containing a question marker (<i>kis</i> ; top figure) and particle (<i>toh</i> ; bottom figure)	245
7.5	Hypothetical eye-movement record with the shaded area representing the region of interest. Adapted from Roberts and Siyanova-Chanturia (2013)	252
7.6	Pearson's correlation coefficients amongst the different predictors and word duration	254

A.1	Example dependency tree (word_position) corresponding to Example 46 . . .	322
A.2	Mean of difference between mean predictor values of variants and reference sentences (95% classification intervals indicated)	327
	(a) Dependency length	327
	(b) trigram surprisal	327
	(c) Parser surprisal	327
A.3	Subjects with and without case marking	328
A.4	Objects with and without case marking	328
A.5	NP internal structure and Modifier attachment error	329
	(a) Gold tree	329
	(b) Parser output	329
A.6	Coordination and NP internal structure Error	330
	(a) Gold tree	330
	(b) Parser output	330
A.7	Split-verb compound Error	331
	(a) Gold tree	331
	(b) Parser output	331
A.8	Split-verb compound Error	332
	(a) Gold tree	332
	(b) Parser output	332
A.9	Bad Sense and VGF Attachment Error	333
	(a) Gold tree	333
	(b) Parser output	333
A.10	Bad Sense within Passives	334
	(a) Gold tree	334
	(b) Parser output	334
A.11	Reading Time vs Conditions	336
A.12	Surprisal vs Conditions	336
A.13	Correlation between predictors	337
A.14	Mean of difference between mean predictor values of variants and reference sentences (95% classification intervals indicated)	342
	(a) Trigram surprisal (8385 data points)	342
	(b) PCFG log likelihood	342
A.15	Correlation plot of predictors in Brown corpus (8385 data points)	345
	(a) UID based on trigram surprisal	345
	(b) UID based on PCFG parser surprisal	345
A.16	Trigram information variation (hartleys/word) across a pair of reference-variant sentences	349
	(a) Example 1	349
	(b) Example 2	349
A.17	Mean of difference between mean predictor values of variants and reference sentences (95% classification intervals indicated)	351
	(a) Entire dataset (8385)	351
	(b) Dative constructions (505)	351
	(c) Postverbal adjuncts (2213)	351

A.18	Trigram information variation (hartleys/word) across a pair of reference-variant sentences	352
	(a) Example 1	352
	(b) Example 2	352
A.19	Density plot over lexical based UID metrics for Brown corpus	354
	(a) Referent sentences (6415 data points)	354
	(b) Variant sentences (8385 data points)	354
A.20	Density plot over syntactic based UID metrics for Brown corpus	355
	(a) Referent sentences (6415 data points)	355
	(b) Variant sentences (8385 data points)	355
A.21	Density plot over lexical based UID metrics for HUTB corpus	355
	(a) Referent sentences (7586 data points)	355
	(b) Variant sentences (158891 data points)	355
A.22	Density plot over syntactic based UID metrics for HUTB corpus	356
	(a) Referent sentences (7586 data points)	356
	(b) Variant sentences (158891 data points)	356

List of Tables

Table	Page
3.1 Constituent-wise statistics (Percentage of constituents in parentheses; ASL is average sentence length in words)	37
3.2 Joachims' transformation	40
4.1 Regression model containing three predictors (158891 data points; all predictors except intercept term significant $p < 0.001$)	61
4.2 Prediction performance (158891 data points; each row refers to a distinct model)	63
4.3 Construction-wise statistics	71
4.4 Regression and prediction results for conjunct verb constructions (116020 data points)	72
4.5 Regression coefficients of construction-wise models (number of data points in parentheses)	77
4.6 Regression and prediction results for overall (72833 points), direct object (DO; 1663 points) and indirect object (IO; 1353 points) fronted cases	81
4.7 Constituent length statistics in HUTB reference sentences	86
4.8 Average distance (number of intervening words) of various preverbal nouns from the root verb	91
4.9 Hindi construction distribution; Written = 1.27 million sentences; Spoken = 17,766 sentences	96
5.1 Subtrees counts post modification.	107
5.2 Our Dataset	107
5.3 Parser performance on unseen test split using Berkeley POS Tagger (Data points = 1230 PSTs). The grammar induced from the treebank contained 34 POS tags, 535 phrasal categories, and 93,040 binary and 542 unary production rules.	114
5.4 Parser performance on unseen test split using NLTK maxent POS Tagger (Data points = 1230 PSTs). The grammar induced from the treebank contained 32 POS tags, 533 phrasal categories, and 1,08,594 binary and 540 unary production rules.	115

5.5	Parser performance on unseen test split using ISCNLP POS Tagger (Data points = 1230 PSTs). The grammar induced from the treebank contained 31 POS tags, 518 phrasal categories, and 97,236 binary and 526 unary production rules.	115
5.6	Parser performance on training dataset containing both train and development splits (Data points = 11082 PSTs)	115
5.7	Parser performance on training dataset containing both train and development splits (Data points = 11082 PSTs)	115
5.8	Parser performance on training dataset containing both train and development splits (Data points = 11082 PSTs)	115
5.9	Bracketing errors using POS tags emitted by Berkeley parser when trained on gold tags	117
5.10	Bracketing errors using POS tags emitted by Berkeley parser when trained on tags obtained from ISCNLP tagger	117
5.11	Parser performance on different syntactic constructions of test dataset with gold POS tagset (Data points = 1230 PSTs)	118
5.12	Regression model containing three predictors (158891 data points; all predictors except intercept term significant $p < 0.001$)	120
5.13	Individual and collective prediction accuracies (** $p < 0.001$ McNemar's two-tailed significance compared to model on previous row)	121
5.14	Regression and classification results on non-canonical word-order choices	123
5.15	Regression model (72833 data points, all significant predictors denoted by $ t >2$)	129
5.16	Regression model on data set where Reference: Given-New , Variant: New-Given (7333 data points, all significant predictors denoted by $ t >2$)	130
5.17	Regression model on Type B dataset (4659 data points, Reference: New-Given , Variant: Given-New , all significant predictors denoted by $ t >2$)	130
5.18	Prediction performances (Full data set (72833 points), Active (71160 points) and Passive (1673 points) cases; each row refers to a distinct model; ** $p < 0.001$ McNemar's two-tailed significance compared to model on previous row)	130
5.19	Impact of discourse distance over and above every other predictors in the classification model (number of cases provided in parentheses; McNemar's two-tailed significance against previous bar indicated using: ** $p < 0.001$; * $p < 0.01$).	132
5.20	Regression model on non-locality data set ($N = 25381$; all significant predictors denoted by $ t >2$)	132
5.21	Regression model on Active data set ($N = 71160$; all significant predictors denoted by $ t >2$)	132
5.22	Regression model on Passive data set ($N = 1673$; all significant predictors denoted by $ t >2$)	132
5.23	Regression model on conjunct-verb data set ($N = 51617$; all significant predictors denoted by $ t >2$)	134
5.24	Regression model on DO-fronted data set ($N = 71160$; all significant predictors denoted by $ t >2$)	135
5.25	Regression model on IO-fronted data set ($N = 1673$; all significant predictors denoted by $ t >2$)	135

5.26	Prediction performances (Conjunct Verb (51617 points), Direct objects (DO; 1663 points) and indirect object (IO; 1353 points) fronted cases; each row refers to a distinct model; *** McNemar’s two-tailed significance compared to model on previous row)	136
5.27	Regression model on GIVENNESS data ($N = 978$; all significant predictors denoted by $ t >2$)	136
5.28	Prediction performances (GIVENNESS dataset (598 points); GIVEN-NEW class (1): 299; NEW-GIVEN class (0): 299; each row refers to a distinct model; *** McNemar’s two-tailed significance compared to model on previous row)	137
5.29	Regression model on full data set ($N = 72833$; all significant predictors denoted by $ t >2$)	144
5.30	Discourse adaptation regression model on DO/IO fronted cases (all significant predictors denoted by $ t >2$)	146
5.31	Prediction performances (Full data set (72833 points), Direct objects (DO; 1663 points) and indirect object (IO; 1353 points) fronted cases; each row refers to a distinct model; *** McNemar’s two-tailed significance compared to model on previous row)	146
5.32	Predictor scores for reference-variant pairs	148
5.33	Effect of adaptation on discourse sentences (Prev1: Preceding one sentence in discourse, Prev5: Preceding five sentences in discourse)	150
5.34	Prediction performance (Direct objects (DO: 1663 points), Indirect Objects (IO: 1353 points)); Baseline denotes <i>base1+g</i> shown in Table 5.31; bold denotes McNemar’s two-tailed significance compared to baseline model in the same row	151
5.35	Targeted human evaluation — Agreement human/corpus : Percentages of times human judgement matches with corpus reference choice; Model corpus : Percentages of corpus choice correctly predicted by the classifier containing all the predictors (<i>base1 + g</i> as per Table 5.31); Model human : Percentages of human label correctly predicted by the classifier containing all the predictors (<i>base1 + g</i> as per Table 5.31)	152
5.36	Prediction performance of verb-specific and subject-objects alternations (72833 points); Baseline denotes <i>base1</i> shown in Table 5.31; bold denotes McNemar’s two-tailed significance compared to baseline model in the same row)	157
5.37	Regression model on lemma verb GIVE data set (14094 data points; all significant predictors denoted by $ t >2$)	158
5.38	Levin’s verb semantic classes and case density (i.e., number of case markers per constituent in a sentence)	158
5.39	Regression model on double object construction S-DO-IO data set (9278 data points; all significant predictors denoted by $ t >2$)	159
5.40	Argument ordering and case density (i.e., number of case markers per constituent in a sentence)	159
5.41	Levin’s verb classes within S-DO-IO data points from Table 5.36	160
5.42	Predictor scores for reference-variant pairs	161
5.43	Regression model on conjunct verb data set ($N = 51617$; all significant predictors denoted by $ t >2$)	163

5.44	Prediction performances (Conjunct Verb (51617 points); each row refers to a distinct model; *** McNemar’s two-tailed significance compared to model on previous row)	163
5.45	Targeted human evaluation — Agreement human/corpus : Percentages of times human judgement matches with corpus reference choice; Model corpus : Percentages of corpus choice correctly predicted by the classifier containing all the predictors; Model human : Percentages of human label correctly predicted by the classifier containing all the predictors	165
5.46	VIF scores for each regression model; each column denotes regression model on corresponding dataset indicated in column header	167
5.47	VIF scores for each regression model after removing <i>trigram surprisal</i> and <i>base lstm surprisal</i> measures from them; each column denotes regression model on corresponding dataset indicated in column header	168
5.48	Regression model on full dataset without correlated features (72833 points; all significant predictors denoted by $ t >2$)	168
5.49	Classification performance without correlated features; + denotes features are added incrementally; *** McNemar’s two-tailed significance compared to model on previous row	168
6.1	Hindi case markers (Butt and King, 1996)	176
6.2	Classification accuracies of surprisal for natural and caseless Hindi (175801 data points)	179
6.3	Values of case features extracted from tree in Figure 6.1.	187
6.4	Learned weights of some case-sequence predictors.	187
6.5	Pairwise classification and ranking accuracy (*** denotes McNemar’s two-tailed significance $p < 0.001$ over the baseline model).	188
6.6	Linear regression model predicting average dependency length on full data set (1996); significant predictors denoted in bold	203
6.7	Four different regression models predicting average dependency length in various binned data set; Column values represent regression coefficient of different predictors in the regression model; dl = Average dependency length ; Bin-wise number of data points in parentheses; Forward and backward surprisal, and same case bigram features not shown as they are not significant (NS) in the models; Avg dl (Min, 1 Quartile, Mean, 3 Quartile, Max) = 0.37, 1.36, 1.83, 2.20, 6.20	204
6.8	Linear regression model predicting average dependency length on DO-fronted dataset (133); significant predictors denoted in bold	205
6.9	Linear regression model predicting average dependency length on IO-fronted dataset (101); significant predictors denoted in bold	205
6.10	Linear regression model predicting average dependency length on conjunct-verb dataset (1158); significant predictors denoted in bold	206
6.11	Linear regression model predicting semantic similarity on full data (1996); significant predictors denoted in bold	206
6.12	Performance of logistic regression models	216
6.13	Performance of logistic regression models (* denotes McNemar’s two-tailed significance $p < 0.05$ over the baseline model)	218

6.14	UID and Hindi syntactic construction (‘+’ stands for ‘Lexical+Syntactic surprisal +’)	220
6.15	Regression Model	225
7.1	Grammatical category-wise descriptive statistics in TDIL and HUTB corpora	234
7.2	Fixed effects of an LMM predicting reading aloud time (15607 data points; all predictors are significant for the $ t =2$ threshold)	240
7.3	Fixed effects of LMM (with word length as interaction term) predicting reading aloud time (15607 data points; all significant predictors denoted by $ t >2$)	242
7.4	Fixed effects of LMM (with word class as interaction term) predicting reading aloud time (15607 data points; all significant predictors denoted by $ t >2$)	242
7.5	Prediction accuracy for content and function word classification (on the entire dataset of 15607 data points) via Generalized LMs where features are added incrementally (all differences between successive pairs of models significant at $p < 0.001$ via McNemar’s test)	246
7.6	Grammatical category-wise descriptive statistics in Potsdam-Allahabad (PAC) and HUTB corpora	250
7.7	Fixed effects of an LMM predicting first pass reading time (FPRT; 40323 data points; significant predictors are in bold with $ t =2$ threshold)	256
7.8	Fixed effects of an LMM predicting regression path duration (RPD; 40323 data points; Significant predictors are in bold with $ t =2$ threshold)	256
7.9	Fixed effects of an LMM predicting total fixation time (TFT; 40323 data points; all predictors are significant for the $ t =2$ threshold)	257
7.10	Regression model predicting the first-pass reading using various cognitive measures (40323 data points; all significant predictors denoted by $ t >2$)	258
7.11	Regression model predicting the regression path duration using various cognitive measures (40323 data points; all significant predictors denoted by $ t >2$)	259
7.12	Regression model predicting the total fixation time using various cognitive measures (40323 data points; all significant predictors denoted by $ t >2$)	260
7.13	Fixed effects of an LMM predicting first pass reading time (FPRT; 40323 data points; all predictors are significant for the $ t =2$ threshold)	261
7.14	Fixed effects of an LMM predicting regression path duration (RPD; 40323 data points; all predictors are significant for the $ t =2$ threshold)	262
7.15	Fixed effects of an LMM predicting total fixation time (TFT; 40323 data points; all predictors are significant for the $ t =2$ threshold)	263
7.16	Prediction accuracy for content and function word classification (on the entire dataset of 56652 data points) via GLMs where features are added incrementally (all differences between successive pairs of models significant at $p < 0.001$ via McNemar’s test)	264
A.1	Dependency length calculation for tree in Figure A.1 (corresponding to Example 46)	323
A.2	Selected constructions in our dataset (#reference sentences/#variants)	324
A.3	Example reference sentences (italicized) and their variants (correct and excluded)	325
A.4	Hindi case markers (Butt and King, 1996)	326

A.5	Probability of predicting the positive class using a model trained using normalized predictor vectors (95% CI given in parentheses in each cell; # data points given in the header row)	327
A.6	The main effect of syntactic surprisal, distance, and their interaction on reading times at the critical region	339
A.7	The main effect of syntactic surprisal, distance, of RC Type, and their interaction on reading times at the critical region	340
A.8	English: Performance of logistic regression models consisting of 8385 data points	343
A.9	English: Performance of logistic regression models consisting of 8385 data points (* denotes McNemar’s two-tailed significance $p < 0.05$ over the baseline model)	344
A.10	English Classification Results: UID effects in different constructions (‘+’ stands for ‘Lexical+Syntactic surprisal +’)	348
A.11	Coefficients and their significance levels for a model that includes all predictors as main effects in English	353