COGNITIVE MODELING OF HINDI SYNTACTIC CHOICE PHENOMENA

SIDHARTH RANJAN

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

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 (Mohanan and Mohanan, 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 सिंटैक्टिक 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 कैइ कारक जाने जाते हैं। हालाँकि, 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 थीसिस 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, यूनिफ़ॉर्म इंफॉर्मेशन डेंसिटी हाइपोथीसिस (जैगर, 2010; लेवी और जैगर, 2007) की जांच लॉजिस्टिक रिग्रेशन 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 अनुकूलनँप्राइमिंग पर पहले के अध्ययनों के संदर्भ में अपने निष्कर्षों को समझने में (फाइन एंड जेगर, 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

List of Figures

Figure **Page**

List of Tables

Table **Page**