A Formal Grammar of Hindi

A thesis submitted to the
Indian Institute of Technology (BHU), Varanasi
for the award of the degree

of

Doctor of Philosophy

by

Vivek Tripathi

Department of Humanistic Studies
Indian Institute of Technology (BHU)
Varanasi - 221005, India
May 2025

Roll Number: 20191511 Year of Submission: 2025

Abstract

This research presents a formal grammatical framework for the syntactic and semantic analysis of Hindi, drawing on concepts from different logical and mathematical frameworks. It explores the denotations of diverse linguistic expressions, ranging from individual words like ***rāma*** to complex phrases such as ***sītā ko dekhatā hai***. The framework encompasses one-place predicates (e.g., ***cala***), two-place predicates (e.g., ***dekha***), and three-place predicates (e.g., ***de***), while also accounting for gender-specific nouns such as ***rāma*** (masculine) and ***sītā*** (feminine). This work aligns with the tradition of formal semantics, which seeks to analyze natural language using tools and concepts from mathematics and logic.

Here are the key aspects of this research:
I have tried to create a system for identifying grammatical and non-grammatical Hindi sentences. This system mathematically defines valid grammatical strings, ensuring a precise distinction between well-formed and ill-formed Hindi sentences.

Noun phrases (NPs) are analyzed as complex structures that include generalized quantifiers such as ***hara*** and ***koī***. This analysis helps to explain how quantified NPs interact with both intransitive and transitive verbs, addressing a critical aspect of Hindi syntax. For example, the sentence ***hara laṛakā calatā hai*** uses the quantifier ***hara*** similar to the universal quantifier ($\forall$) in logic.

The mathematical theory of types has been used here to categorize predicates, and organize denotations into a hierarchy of distinct types such as D_x and D_y. This systematic approach allows for a more precise analysis of linguistic expressions. For example, sentences like ***rāma calatā hai*** are of type <t>, while individuals like ***rāma*** are of type <t>.

Tenses and aspects have been formalized by using temporal logic (TL). This involves a detailed examination of common Hindi conjugations, such as ***-tā hai***, ***-ā thā***, and ***-egā***, within a structured temporal framework. The approach treats the present tense as a core representation, with past and future tenses being relative to it. This work extends the analysis of Hindi beyond static sentence structures and into the realm of dynamic time-based meanings.

The development of a Python-based software tool implementing the complete grammar. This tool provides a practical demonstration of the developed theoretical framework,

generating parsing trees and semantic computations through a UI. The tool is modular, with components for each of the formalisms developed in the thesis.

Overall, this thesis addresses several core areas of formal linguistics by adopting a modular and incremental approach. The work is divided into fragments that progressively introduce formalisms of propositional logic (L_{1H}), predicate logic (L_{2H}), type theory (L_{3H}) and temporal logic (L_{4H}). Each fragment systematically builds upon the previous one, creating a comprehensive model of Hindi syntax and semantics. This thesis aims to provide a rigorous foundation for formalizing Hindi grammar, contributing to both the theoretical understanding of language and the practical applications in computational linguistics and natural language processing (NLP). The final 'Hindi Fragment' integrates 16 syntactic rules and 14 semantic rules.

Keywords: Natural Language Processing, Syntax of Hindi, Semantics of Hindi, Formal Grammar, Montague Grammar, Parsing, Logic

Contents