भारतीय प्रौद्योगिकी संस्थान
काशी हिन्दू विश्वविद्यालय

IIT INDIAN INSTITUTE OF TECHNOLOGY BANARAS HINDU UNIVERSITY

Research and Innovation Day

September 05, 2024

Structural Analysis of Hindi Using Formal Methods

Vivek Tripathi

Research Scholar, IIT BHU, Varanasi, India – 221005 [Email: sopan.tripathi@gmail.com]

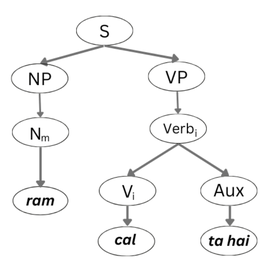
Abstract:
This research applies formal grammatical methods to fragments of ordinary Hindi and aims to provide the foundation for explicit semantics of Hindi. Some syntactic and semantics rules are proposed to describe the nature of Hindi grammar. A basic fragment L_{1H} has been updated several times to incorporate different layers of the complexities of Hindi grammar. The updated layers (L_{2H}, L_{3H} and L_{4H}), also called Fragments, have been explained through basic expressions divided into various categories and a set of formation rules. Then, the tense of the Hindi language was accounted for, and important syntactic generalizations and semantic rules were developed pertaining to the behavior of past and future conjugations of Hindi verbs (in L_{4H}).

Our main theoretical concern is the relation between logical syntax and linguistic syntax. All the fragments are implemented through an in-house software tool developed using Python and linguistic libraries. Hopefully, a new linguistic framework will be developed from this study.

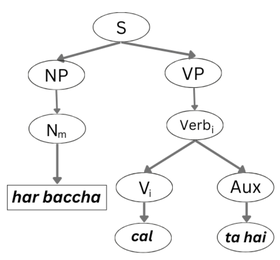
Materials & Methods:
Montague Grammar (MG) and Transformational Grammar (TG) have been two prominent classical fields of study by researchers of 'The School of Philosophy' and 'The School of Linguistics' respectively. This research focuses on the former to analyze the Hindi language and develop its formal counterpart. The framework we developed for Hindi is a combination of MG and TG. Successful syntactic-semantic formalizations have led to promising achievements in natural language processing (NLP) of Hindi.

Syntax of 'The Final Fragment' –

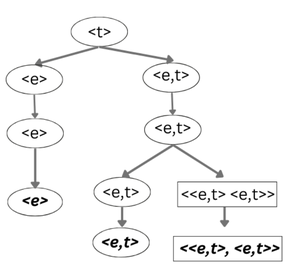
L_{1H}



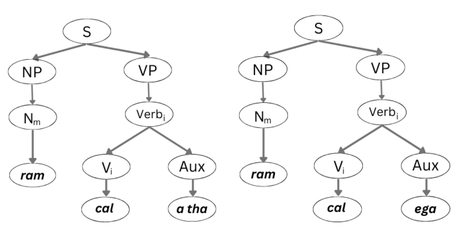
L_{2H}



L_{3H}



L_{4H}

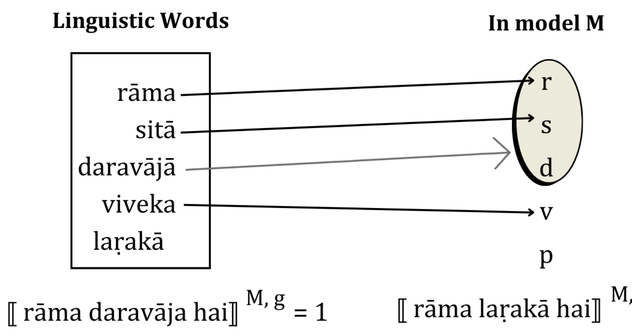


Past and Future auxiliaries are introduced syncategorematically.

The fragments L_{1H}, L_{2H}, L_{3H} and L_{4H} are isomorphic to their corresponding logical languages. The final fragment, L_{5H}, is the total of all the developed fragments and is demonstrated through a software application.
While L_{1H} focuses primarily on developing syntactical rules for referential noun phrases, L_{2H} uses the idea of quantification to frame rules for non-referential noun phrases. L_{3H} classifies linguistic expressions based on their semantic types, and L_{4H} uses the idea of temporal logic to deal with past and future tense auxiliaries of Hindi.

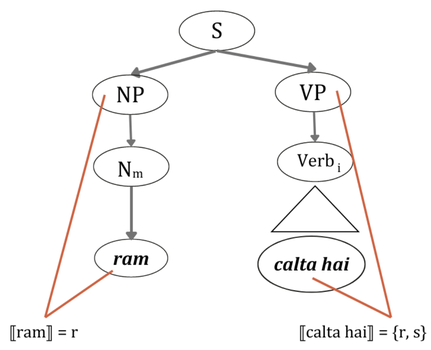
Semantics of 'The Final Fragment' –

(1) The semantic value of α i.e. ⟦α⟧, where α is a linguistic unit, is calculated w.r.t a model M and an interpretation function I. Following is an example of a model M.



(2) In this model M, the interpretation function interprets semantic-value of rāma, sītā, daravājā, laṛakā etc. as {r}, {s}, {r, s, d}, {ϕ} respectively.

(3) Each well-formed formation (wff) of the fragment contains linguistic units. These units are treated either as an individual or as a predicate. The predicates act as a function and individuals as arguments. Such distinction is identified based on syntax. For example, in wff 'rāma calatā hai.', ⟦rāma⟧ outputs {r} and ⟦cal⟧ outputs {r, s} and ⟦tā hai⟧ outputs null value. Thus, all auxiliaries are semantically vacuous.



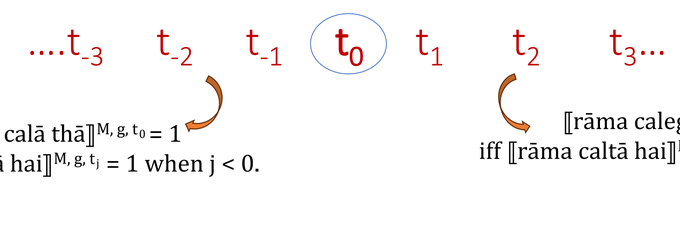
There may be multiple models M where linguistic units' semantic value (SV) can be different. Thus, we use the value assignment function (g) that assigns SV wr.t to a model M. Thus, the denotations of each wff are w.r.t M & g.

⟦rāma caltā hai⟧^{M,g} = 1 iff ⟦rāma⟧ ϵ ⟦caltā hai⟧
⟦rāma caltā hai⟧^{M,g} = 1 iff {r} ϵ {r, s}, otherwise 0.

Thus, ⟦rāma caltā hai⟧^{M,g} = 1

Software output does not incorporate diacritical marks currently.

(4) The SV of past and future tense-based wffs are computed using the linear time sequence. If t_0 corresponds to the present moment, represented by the 'tā hai' auxiliary, then time-moments before and after represent past and future tenses, respectively.



Novelty and Key Achievements:
This research study provided a scientific understanding of the functioning of syntax and semantics of Hindi. The fragments with different attributes provide a complete framework for understanding the linguistic behavior of quantified and non-quantified Hindi expressions. Type logic provides an alternative view of understanding behaviors of vectors, i.e. light verbs when adjoining root or non-root part of a verb.

The Hindi Tree (THT) Parser automatically builds linguistic syntax trees for Hindi sentences. This tool can parse an annotated sentence and display every tree that satisfies the syntax rules.

Publications:
Tripathi, V., & Rathod, D. (2024). Semantic model for fragment of Hindi (Part 1). Rupkatha Journal, 16(1). https://doi.org/10.21659/rupkatha.v16n1.03g
Tripathi, V., & Rathod, D. (2024). Semantic model for fragment of Hindi (Part 2). Rupkatha Journal, 16(2). https://doi.org/10.21659/rupkatha.v16n2.02

Contact: Vivek Tripathi
Email- vivektripathi.rs.hss20@iitbhu.ac.in

Contact: Chairman, RID-2024
Email- RID2024@iitbhu.ac.in