Grammar as Infrastructure for Indian AI

Vivek Tripathi[1], Department of Humanistic Studies, IIT (BHU), Varanasi
[1]https://iamalinguist.github.io

Abstract

AI for Indian languages is dominated by corpus-based models that overlook deep grammar, leading to errors in interpretation, agreement, and meaning retention—especially in complex languages like Hindi. This work presents grammar as infrastructure for AI, treating formal linguistic modeling as foundational rather than supplementary. Using Hindi as a prototype, it addresses challenges like light verbs, auxiliaries, postpositional subjects, and syntax–semantics mismatches through type theory, update semantics, and formal parsing. The approach enhances machine reasoning and semantic accuracy while supporting language pedagogy and computational modeling. Building linguistically aware AI for Hindi offers a scalable framework for other Indian languages.

Objective

- Establish grammar as a foundational layer for AI in Indian languages.
- Demonstrate how formal models improve interpretation beyond statistical methods.
- Use Hindi as a scalable prototype for grammar-based AI across Indian languages.
- Present grammar-driven solutions to light verbs, agreement, auxiliaries, and semantic composition.
- Advocate for future AI pipelines grounded in linguistic theory rather than post-hoc corrections.

Introduction

AI systems for Indian languages are largely built on statistical learning and translation-based pipelines. While effective for surface-level tasks, they fail when confronted with deep linguistic phenomena such as compositional meaning, agreement mismatches, and structural ambiguity. Hindi, like most Indian languages, exhibits features—light verbs, postpositional subjects, free word order, auxiliary stacking, and meaning shift—that challenge syntax–semantics alignment [4,5].

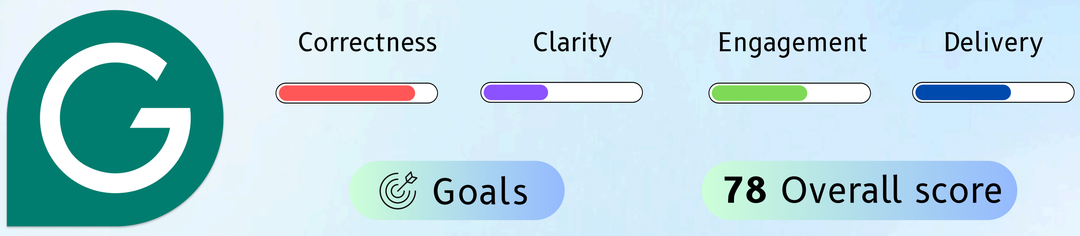

Fig 1. Grammarly works because it has grammar-aware infrastructure underneath. AI systems without grammar can only approximate language correctness, not understand it.

This work positions grammar [cf. Fig 1] not as annotation, but as infrastructure. Without formal grammatical modeling, AI systems continue to misinterpret meaning and produce unreliable outputs. Using Hindi as the prototype, this research demonstrates how formal syntax and semantics provide a scalable foundation for Indian language AI.

Methodology

- Type-theoretic and update-semantics modeling to analyze auxiliaries, meaning shifts, and dynamic interpretation [1,2,3].
- Syntax–semantics interface design to address light verb constructions and non-isomorphic structures.
- Constituency parsing infrastructure through Hindi Tree, an open-access parser [7].
- Montague-style compositional semantics [1] for meaning computation, inference, and instructional use.
- Agreement and argument structure modeling to account for postpositional subjects and role interpretation.
- Pedagogical semantics frameworks to improve language acquisition and reasoning.

Results and Discussion

The research resolves several major limitations in AI handling of Hindi:

- **Light verbs and meaning decay**: Demonstrated that meaning is distributed across constructions [4,5], requiring type-shifting models.
- **Auxiliary dynamics**: Introduced a formal update-semantic treatment for tense, aspect, and modality.
- **Syntax–semantics misalignment**: Showed why surface parsing fails and how compositional models restore interpretive accuracy.
- **Postpositional subjects and agreement**: Proposed a new theory that captures non-canonical agreement patterns.
- **Compositional meaning for AI and pedagogy**: Developed semantic triads (CPI) to support reasoning and instruction.
- **Computational tools**: Built a constituency parser [7] as grammatical infrastructure rather than annotation.

Conclusions

This work demonstrates that AI for Indian languages cannot rely solely on corpora, annotations, or statistical generalization. Hindi provides a clear case where phenomena such as light verbs, agreement shifts, auxiliaries, and non-isomorphic syntax require grammar-based modeling for accurate interpretation. Formal grammar is not an academic add-on — it is the missing infrastructure for machine reasoning, semantic fidelity, and linguistic pedagogy [1,2]. Developing grammatical infrastructure for Hindi establishes a scalable template that can be extended to other Indian languages [6].

Translational Potential

- AI and NLP systems: Integration into parsing, translation, speech interfaces, and reasoning models.
- Educating Indian Languages: Semantic-pedagogical frameworks for teaching Hindi and other Indian languages.
- Multilingual AI infrastructure: Scalable adaptation to Marathi, Bangla, Tamil, Punjabi, and others.
- Cognitive and decision-making models: Use in inference engines, dialogue systems, and meaning computation.
- Policy and standardization: Supports national initiatives for linguistic AI and digital governance.

Acknowledgements

Supported by the Department of Humanistic Studies, IIT (BHU), with PhD guidance from Dr. Nirmalya Guha and valuable review insights from Dr. Sanjukta Ghosh and Dr. V. Ramanathan.

References

[1] R. Montague, Formal Philosophy, Yale Univ. Press, 1974.
[2] H. Kamp, U. Reyle, From Discourse to Logic, Kluwer Acad. Publ., 1993.
[3] J. Carpenter, Type-Logical Semantics, MIT Press, 1997.
[4] T. Bhatt, "Light Verb Constructions in Hindi," Lingua 113 (2003) 1239–1272.
[5] A. Mohanan, Argument Structure in Hindi, CSLI Publ., 1994.
[6] A. Bojar, D. Zeman, "Improving SMT for Indian Languages Using Syntactic and Morphological Analysis," Proc. ICON, 2006, pp. 15–24.
[7] V. Tripathi, "Hindi Tree: An Open-Access Constituency Parser for Hindi," Language Resources and Evaluation, 2025 (submitted).