Vākya to *Vṛkṣa*: A Digital Tool for Understanding Hindi Sentence Structure

Vivek Tripathi[1]
Department of Humanistic Studies, Indian Institute of Technology, BHU, Varanasi, UP-221005, India.

Abstract

Hindi grammar has traditionally been understood and taught through descriptive rules, categories, and examples presented in a linear textual form. While this tradition has ensured continuity of grammatical knowledge, it often leaves the internal organization of sentences implicit rather than visible. As a result, learners frequently engage with grammar as a set of instructions rather than as a structured system of relations. In contemporary digital learning contexts, this raises an important question: how can the structural nature of Hindi sentences be made perceptible without departing from established grammatical intuition?

This paper proposes a shift from *Vākya* (sentence as surface expression) to *Vṛkṣa* (sentence as structured entity) through phrase-structure–based visualization of Hindi sentences. It incorporates Hindi Tree, a web-based tool that transforms Hindi sentences into tree representations, making constituent relations such as subject–predicate organization, object marking, postpositions, auxiliaries, and adjuncts explicitly visible. Rather than replacing traditional grammar, the tool complements it by providing a visual grammar space in which learners can observe how sentence components are structurally connected. From pedagogical and cultural perspectives, Hindi Tree is positioned as an accessible digital aid for grammar learning, teacher instruction, and the development of open, structure-aware resources for Indian languages.

Keywords: Hindi Grammar; Sentence Structure; Tree Representation; Digital Linguistics; Language Pedagogy; *Bhāratiya* Languages.

1. Introduction

Understanding how words combine to form meaningful sentences is a foundational concern in language learning and linguistic analysis. In Hindi, this challenge is compounded by flexible word order, postpositions, auxiliary constructions, and aspectual morphology, all of

[1] ✉ **Corresponding Author: sopan.tripathi@gmail.com**

which often remain implicit in conventional teaching practices. While learners may successfully identify grammatical categories at the word level, they frequently struggle to perceive how these elements group together structurally to produce interpretation. This gap between surface recognition and structural understanding becomes particularly evident for non-native speakers, second-language learners, and even early-stage native learners of Hindi.

Traditional grammar instruction in Hindi has largely emphasized sentence-level correctness (*vākya*), focusing on identifying parts of speech and producing well-formed utterances. However, the hierarchical organization underlying sentences—the structural relationships among noun phrases, verb phrases, postpositional phrases, negation, and auxiliaries—often remains unarticulated. As a result, learners may memorize rules without developing a clear internal model of sentence structure. This limitation is not unique to Hindi, but it is especially salient given the language's rich morphosyntactic patterns and relatively free constituent ordering.

Phrase-structure–based representations offer a way to externalize this hidden organization by making constituent groupings explicitly visible. By representing sentences as hierarchical trees, such representations allow learners to see how words form larger units and how these units relate to one another within a sentence. Importantly, this approach does not introduce new grammatical concepts; rather, it provides a visual articulation of structural intuitions that are already implicit in traditional grammatical understanding. In this sense, tree representations serve as a bridge between abstract grammatical knowledge and concrete perceptual learning.

The present paper discusses constituents through Hindi Tree, a sentence-tree visualization application designed to support grammatical comprehension through explicit structural representation. Hindi Tree focuses on core phrase-structure relationships in Hindi, allowing users to visualize how sentences are organized into constituents such as noun phrases, verb phrases, and postpositional phrases. By interacting with tree structures, learners can explore common sources of confusion—such as postposition attachment, auxiliary placement, negation scope, and modifier grouping—and observe how these elements function within a coherent structural framework.

Within the broader context of the *Bhāratīya Bhāṣā Parivāra*, this work highlights the pedagogical value of making grammatical structure visible without departing from established linguistic intuitions. Rather than proposing a new grammatical theory, the

approach presented here emphasizes clarity, accessibility, and continuity with existing grammatical traditions. By combining phrase-structure visualization with digital interactivity, Hindi Tree demonstrates how technology can support deeper grammatical understanding while remaining grounded in the linguistic character of Indian languages.

2. Related Work: Syntactic Studies and Parsing of Hindi

Hindi has been extensively studied within descriptive and theoretical linguistics, particularly with respect to clause structure, argument realization, and verbal morphology. Early and influential work by Mohanan (1994) provides a detailed account of argument structure in Hindi, addressing case marking, transitivity, and the interaction between syntax and semantics. Studies such as Bhatt (2003) and Butt and Lahiri (2013) further examine the structure of Hindi clauses, focusing on auxiliary constructions, aspectual morphology, and light verb combinations. These works establish a rich theoretical understanding of Hindi syntax, but they are primarily analytic in nature and do not aim to make sentence structure visually accessible for learners or non-specialists.

From a descriptive and typological perspective, comprehensive overviews of Hindi and related Indo-Aryan languages are provided in works such as Kachru (2006), which situate Hindi within the broader Indo-Aryan family and document its syntactic properties, including verb-final order, postpositional phrases, and relatively flexible constituent ordering. While such descriptions are foundational, they typically rely on prose and example sentences rather than explicit structural visualization.

In computational linguistics, substantial efforts have been made to develop syntactic resources and parsers for Hindi. The Hindi Treebank, developed by Begum et al. (2008), represents one of the most significant contributions in this area, providing large-scale annotated data for syntactic analysis. Much of the computational work on Hindi parsing has adopted dependency-based representations, often inspired by the *Paṇinian* grammatical framework (e.g., Bharati et al., 1995; Bharati et al., 2009). Dependency parsers for Hindi and other Indian languages have been developed and evaluated extensively (e.g., Begum et al., 2010; Rao et al., 2010; Husain et al., 2014), primarily with the goal of supporting downstream NLP tasks such as machine translation and information extraction.

While these computational resources are invaluable for automated processing, they are not designed for pedagogical exploration or for helping learners' reason about sentence structure. Dependency representations emphasize head–dependent relations and semantic roles, often abstracting away from constituent grouping. As a result, they offer limited support for learners who wish to understand how words form larger structural units such as noun phrases and verb phrases.

Constituency-based representations for Hindi are comparatively less common, and where they exist, they are largely intended for parser training or theoretical analysis rather than interactive use. Moreover, existing treebanks and parsers are typically not accessible through open, learner-oriented interfaces. Similar observations hold for syntactic resources developed for structurally related languages such as Marathi, Bengali, Punjabi, and Tamil, where dependency parsing has received far more attention than constituency-based visualization.

The present work positions itself at the intersection of these traditions. Building on established descriptive and theoretical insights into Hindi syntax, and informed by prior computational work on parsing, it addresses a distinct but underexplored need: the availability of an open, constituency-based visualization tool for understanding sentence structure. Rather than proposing a new grammatical theory or competing with existing parsers, the approach adopts a phrase-structure perspective to make syntactic organization perceptible and interpretable for learners, teachers, and other stakeholders.

3. Structural Challenges in Understanding Hindi Sentences

Despite familiarity with vocabulary and basic grammatical categories, learners of Hindi often experience difficulty in understanding how sentences are internally organized. These difficulties typically do not arise from incorrect word forms, but from uncertainty about how words combine into larger structural units. In Hindi, where case marking is selective, auxiliaries are multi-part, and adjuncts may freely intervene, linear order alone is insufficient to reveal syntactic relationships. As a result, learners may recognize individual elements correctly while misinterpreting their grammatical roles.

A recurring source of confusion concerns the use of case marking, particularly the postposition *ko*. Consider the contrast in (1):

(1a) *viveka mobāila dekha rahā hai*
Vivek mobile see PROG AUX
'Vivek is watching a mobile.'

(1b) *viveka rāma ko dekha rahā hai*
Vivek Ram ACC see PROG AUX
'Vivek is looking at Ram.'

Although both sentences contain the same verb *dekha*, only the second requires the postposition *ko*. Learners often treat *ko* as a general marker of objects, yet its distribution is selective. The contrast illustrates that *ko* is not merely a marker of verb–object relation, but interacts with semantic and pragmatic properties such as animacy and individuation. Without a clear representation of how noun phrases relate to the predicate, learners may incorrectly assume that all direct objects require identical marking. Structural visualization helps clarify that *mobāila* and *rāma* occupy similar syntactic positions, while differing in how case marking is realized.

Auxiliary constructions introduce a different kind of structural ambiguity. Consider the contrast in (2):

(2a) *maĩ gira gayā*
I fall went
'I fell.'

(2b) *Maĩ gira gayā hũ̃*
I fall went AUX
'I have fallen.'

In (2a), *gayā* is often misinterpreted as a tense auxiliary, leading learners to assume that the sentence already contains a complete auxiliary structure. However, the structurally complete predicate emerges in (2b), where *hũ̃* functions as the auxiliary anchoring the clause. The example illustrates how surface similarity can obscure underlying structure: without visual representation, learners may conflate participial elements with auxiliary verbs, resulting in an incomplete or incorrect analysis of predicate formation.

Adjunct attachment provides another challenge, particularly with intransitive verbs. Consider (3):

(3) *sītā ghara mẽ ā cukī hai*
Sita home in come PERF AUX
'Sita has come home.'

Here, *ā* is an intransitive verb, and *ghar mẽ* functions as a locative adjunct rather than an object. Learners, however, may incorrectly infer transitivity due to the presence of a postpositional phrase. Structural visualization makes it clear that *ghar mẽ* does not form an object constituent with the verb, but attaches as an adjunct within the predicate. Without such visualization, learners may incorrectly generalize that any postpositional phrase following a verb signals an object relation.

A final contrast highlights the difference between bare and case-marked noun phrases in object position:

(4a) *maĩ kāma kiyā hũ̃*
I work did AUX
'I have done work.'

(4b) *maĩ kāma ko kiyā hũ̃*
I work ACC did AUX
'I have done the work.'

Although both sentences are structurally similar, the presence of *ko* in (4b) signals a different interpretation of the noun phrase. In (4a), *kāma* functions as a bare object integrated into the verbal predicate, whereas in (4b), *kāma ko* is construed as a more individuated object. Learners often struggle to articulate this distinction, relying instead on memorized patterns. Tree-based representation helps make visible how both constructions are structurally related while differing in internal noun phrase marking.

Taken together, these examples demonstrate that many challenges in Hindi grammar comprehension are structural rather than lexical. Learners may know the meanings of words and even recognize correct forms, yet remain uncertain about how those forms are organized within the sentence. The absence of visible constituent structure makes it difficult to reason about case marking, auxiliary composition, adjunct attachment, and object interpretation. Addressing these challenges requires an approach that foregrounds structural organization, enabling learners to move beyond surface patterns toward principled grammatical understanding.

4. Hindi Tree: An Open Phrase-Structure Visualization System for Hindi

The structural challenges discussed in the previous section point to a broader gap in Hindi language pedagogy and resources: while learners are frequently exposed to grammatical terminology and sentence-level rules, there exists little support for making sentence structure itself explicitly visible. In contrast to several widely studied languages, Hindi lacks open, learner-oriented tools that allow users to explore constituency structure interactively. Existing resources either assume advanced theoretical knowledge or focus narrowly on correctness rather than comprehension. Hindi Tree (HT) is designed to address this gap by providing an open-access, phrase-structure–based visualization system that foregrounds structural understanding for a wide range of stakeholders, including language learners, teachers, and researchers.

HT adopts a phrase-structure perspective in which sentences are represented as hierarchically organized groupings of constituents such as noun phrases, verb phrases, and postpositional phrases. This choice is both pedagogical and methodological. Phrase-structure grammar (PSG) offers a relatively flat and transparent representation of syntactic organization, allowing users to see how words group together without requiring engagement with advanced theoretical constructs such as movement, projection levels, or derivational operations. By privileging constituency over derivation, HT maintains structural integrity while remaining accessible to non-specialists.

Consider the sentence:

(5a) *sītā rāma ko dekha rahī hai*
Sita Ram ACC see PROG AUX
'Sita is seeing Ram.'

In its linear form, the sentence presents a sequence of words whose grammatical relations must be inferred implicitly. When this sentence is represented as a hierarchical structure in HT (see Figure 1), *sītā* forms the subject noun phrase, *rāma ko* forms an object noun phrase within the predicate, and *dekha rahī hai* functions as a unified verbal complex. By explicitly displaying these groupings, the tree representation clarifies the role of case marking, distinguishes arguments from modifiers, and makes auxiliary structure visible in a way that linear representation does not.

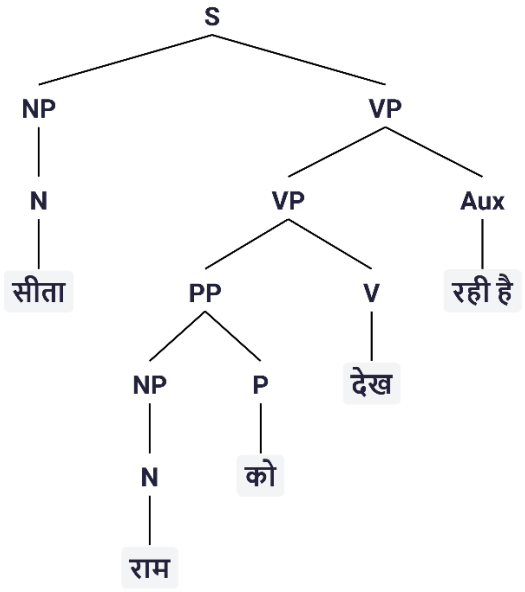

Figure 1. Linear sentence (*vākya*) versus hierarchical phrase-structure representation (*vṛkṣa*) in Hindi.

Postpositional phrases in Hindi may function either as arguments or as adjuncts, a distinction that is often difficult for learners to articulate using linear representations alone. This becomes particularly evident in sentences with intransitive predicates, where the presence of a postpositional phrase does not necessarily imply an object relation. When such constructions are visualized in HT (see Figure 2), adjunct phrases such as *ghara mẽ* in the sentence such as *sītā ghara mẽ ā cukī hai*, are shown attaching to the predicate as modifiers rather than forming an object constituent with the verb. By making this attachment distinction explicit, the tree representation helps learners distinguish argument structure from optional modification, a contrast that is frequently obscured in surface-level analysis.

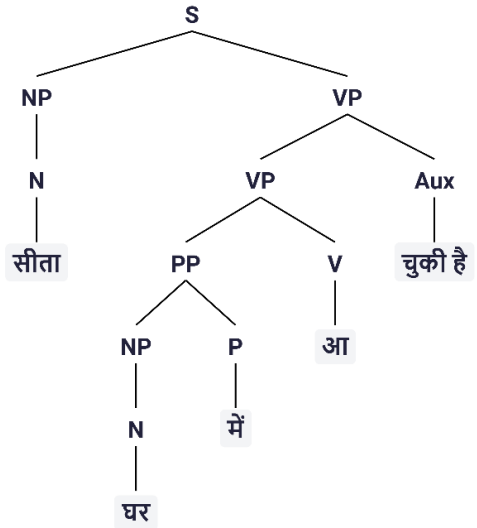

Figure 2. Adjunct versus argument attachment in Hindi predicates.

Hindi predicates frequently consist of a lexical verb combined with participial material and a finite auxiliary, forming multi-word verbal expressions whose internal organization is not always transparent to learners. In linear representations, these elements may appear as loosely concatenated forms, encouraging the misconception that tense or aspect is marked by a single word. When such constructions are visualized in HT (see Figure 3), the predicate is represented as a unified structural unit in which participial elements and auxiliaries are grouped together. This visualization clarifies the role of auxiliary anchoring and helps distinguish participial morphology from finite tense marking, thereby supporting a more accurate understanding of predicate formation in Hindi.

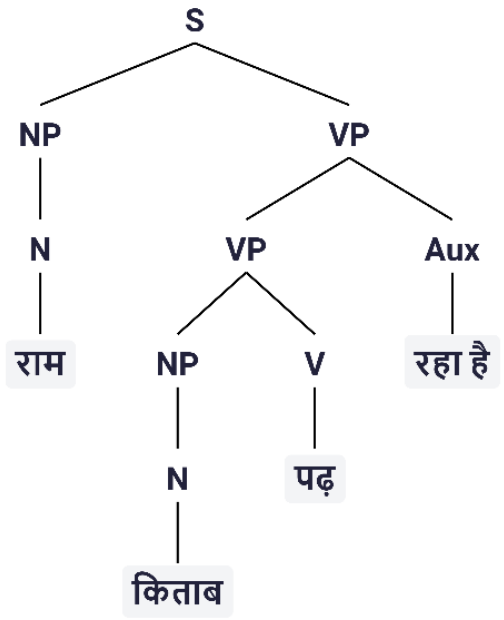

Figure 3. Internal structure of Hindi predicates with participial forms and auxiliaries.

While treebanks and parsers exist for research purposes, they are typically not designed for learner interaction, nor are they easily accessible to non-specialists. HT positions itself as an open, exploratory interface to constituency structure, enabling users to engage with grammatical organization without requiring prior training in formal syntax.

5. Pedagogical, Cultural, and Cross-Linguistic Significance

The pedagogical value of HT lies in its ability to make grammatical structure perceptible without altering the substance of traditional grammatical knowledge. In many instructional settings, Hindi grammar is taught through linear sentences, word-level categorization, and rule memorization. While this approach enables learners to produce well-formed sentences, it

often leaves the internal organization of those sentences implicit. HT addresses this limitation by externalizing structure, allowing learners to observe how grammatical units are organized and related within a sentence.

From a learning perspective, visualizing phrase structure supports a shift from recognition-based learning to structural reasoning. Instead of memorizing isolated rules for postpositions, auxiliaries, or negation, learners can see how these elements systematically attach within a sentence. HT is also pedagogically inclusive. Its phrase-structure representations are accessible to a wide range of users, including school students, second-language learners, teacher trainees, and early-stage researchers.

Within the broader cultural context of the *Bhāratīya Bhāṣā Parivār* (BBP), HT aligns naturally with long-standing grammatical intuitions. Indian grammatical traditions have historically emphasized relational understanding among linguistic units, even when such relations were not expressed through explicit diagrams. The transition from *vākya* to *vṛkṣa* offered by HT can thus be seen as a visual articulation of those insights. At the same time, the relevance of HT extends beyond Hindi alone. Because the system is grounded in phrase-structure representation rather than language-specific rules or scripts, it is naturally adaptable to other languages that share similar structural properties. Many Indian languages exhibit comparable syntactic characteristics, such as verb-final (SOV) word order, postpositions following noun phrases, auxiliary elements occurring at the clause periphery, and relatively flexible constituent ordering. For such languages, the pedagogical challenge of making sentence structure visible closely parallels that of Hindi, making tree-based visualization equally effective.

To illustrate this, HT is applied to sentences from several structurally related languages, including Punjabi, Gujarati, Marathi, Bengali, Tamil, and Sanskrit. We can observe using Figure 4 that for languages such as Punjabi (ਮੈਂ ਪੜ੍ਹਦਾ ਹਾਂ; *maĩ paṛhdā hã̃*) and Gujarati (હું વાંચું છું; *hũ vãcũ chũ*), HT can be adopted in a relatively straightforward manner due to the presence of overt auxiliary constructions similar to Hindi. In contrast, for languages such as Marathi (मी वाचतो; *mī vācatō*), Bengali (আমি পড়ি; *āmi poṛi*), Tamil (நான் படிக்கிறேன்; *nāṉ paṭikkiṟēṉ*), and Sanskrit (अहं पठामि; *ahaṃ paṭhāmi*), script differences do not pose a constraint; however, adaptation requires the induction of an additional

morphological layer to decompose verbal forms and make their internal structure explicit for phrase-structure visualization.

As shown in Figure 4, sentences conforming to an SOV profile across different scripts can be represented using the same constituent-based framework. These visualizations highlight shared structural patterns—such as verb-final predicates—while allowing language-specific features to remain visible. The figure demonstrates that HT functions as a flexible visualization framework for a family of structurally similar languages rather than as a language-specific parser.

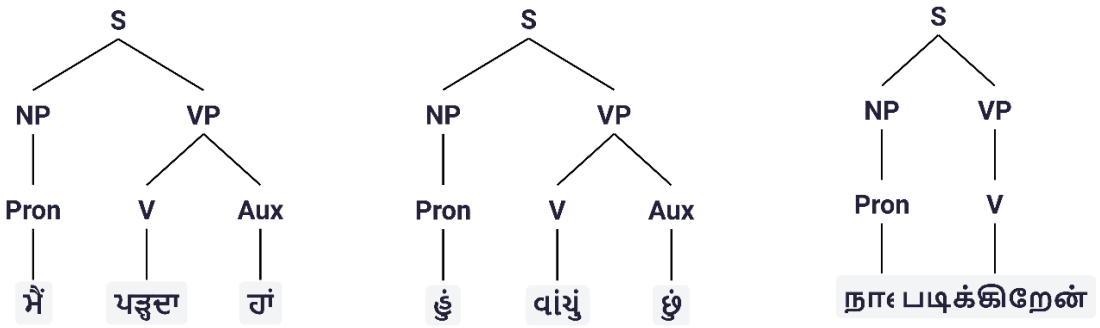

Figure 4. Phrase-structure tree visualizations for structurally similar and different languages (Punjabi, Gujarati and Tamil), illustrating the adaptability of HT to SOV languages across scripts.

Importantly, this cross-linguistic applicability does not depend on script uniformity. Since HT operates on constituent grouping rather than orthographic form, it can accommodate languages written in Devanagari, Perso-Arabic, Brahmic, or other scripts, provided that basic lexical segmentation and grammatical categories are available. This script-agnostic design allows the same visualization framework to be extended across languages while preserving structural integrity.

Finally, the open and exploratory nature of HT contributes to a more inclusive ecosystem of language-learning resources. Many existing grammatical tools are either proprietary or targeted exclusively at specialists, limiting their usefulness for learners and teachers. HT addresses this gap by offering an open interface to sentence structure. In doing so, it supports not only Hindi grammar education but also the broader goal of developing shared pedagogical resources for languages within the *Bhāratīya Bhāṣā Parivār*.

6. Conclusion

This paper has argued that many persistent difficulties in understanding Hindi grammar arise not from a lack of grammatical knowledge, but from the absence of visible structural representation. Learners are often able to identify words, inflections, and surface patterns, yet struggle to perceive how these elements combine into larger grammatical units. By shifting attention from linear sentence strings to hierarchical constituent organization, phrase-structure visualization offers a principled way to bridge this gap between recognition and understanding.

Through the presentation of HT, the paper has demonstrated how a phrase-structure–based approach can make sentence organization explicit without introducing theoretical complexity. By focusing on observable constituency relations—such as noun phrase formation, postpositional attachment, auxiliary construction, and adjunct placement—HT enables learners to reason about structure rather than rely on memorized rules. The choice to ground the system in phrase-structure grammar ensures conceptual clarity and pedagogical accessibility while preserving the structural integrity of Hindi sentences.

Importantly, the contribution of HT extends beyond Hindi alone. The inclusion of tree visualizations from structurally similar languages such as Punjabi, Gujarati, Marathi, Bengali, Tamil, and Sanskrit illustrates that the underlying approach is adaptable across languages that share an SOV profile and postpositional structure. This adaptability highlights the potential for developing shared, structure-aware pedagogical resources within the *Bhāratīya Bhāṣā Parivāra*, without erasing language-specific distinctions or imposing uniform analytical assumptions.

Equally significant is the open and exploratory nature of the system. In a context where accessible constituency-based resources for language learners remain limited, HT offers an open access for engaging with grammatical structure. Rather than functioning as a prescriptive grammar checker or a theory-driven parser, it serves as a reflective tool that encourages comparison, inquiry, and structural insight.

In sum, HT illustrates how visualizing phrase structure can enhance grammatical comprehension, support pedagogy, and foster cross-linguistic awareness while remaining grounded in established linguistic intuition. By making structure visible, the approach contributes to a deeper and more principled understanding of Hindi and related languages,

demonstrating the value of constituency-based visualization as a bridge between language tradition, learner cognition, and digital innovation.

7. Reference

Begum, R., Husain, S., Bai, L., Sharma, D. M., Sangal, R., & Bharati, A. (2008). Developing a dependency treebank for Hindi. *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008)*, 1–8.

Begum, R., Husain, S., Sharma, D. M., Sangal, R., & Bharati, A. (2010). Dependency annotation scheme for Indian languages. *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2010)*, 1–8.

Bharati, A., Chaitanya, V., Sangal, R., & Ramakrishnamacharyulu, K. V. (1995). *Natural language processing: A Paninian perspective*. Prentice-Hall of India.

Bharati, A., Sangal, R., Sharma, D. M., & Bai, L. (2009). *AnnCorra: Treebanks for Indian languages—Guidelines for annotating Hindi treebank (Version 2.0)*. Language Technologies Research Centre, International Institute of Information Technology Hyderabad.

Bhatt, R. (2003). Locality in correlatives. *Natural Language & Linguistic Theory, 21*(3), 485–541. https://doi.org/10.1023/A:1024128401581

Butt, M., & Lahiri, A. (2013). Diachronic pertinacity of light verbs. *Lingua, 135*, 7–29. https://doi.org/10.1016/j.lingua.2013.06.006

Husain, S., Sharma, D. M., Sangal, R., & Bharati, A. (2014). A transition-based dependency parser for Hindi. *Proceedings of the COLING 2014 Workshop on Indian Language Data: Resources and Evaluation*, 1–10.

Kachru, Y. (2006). Hindi. In G. Cardona & D. Jain (Eds.), *The Indo-Aryan languages* (pp. 566–601). Routledge.

Mohanan, T. (1994). *Argument structure in Hindi*. CSLI Publications.

Rao, D., Husain, S., Gadde, P., & Sangal, R. (2010). Parsing Indian languages: Dependency-based approaches. In *Proceedings of the Eighth International Conference on Natural Language Processing (ICON-2010)*. Association for Computational Linguistics.