

Interpretable Text-Based Prosodic Boundary Prediction for Hindi Children’s Stories

Sakshi Jadhav^{1,*}, Mansi Dhamne^{1,*}, Preeti Rao²

¹Sardar Patel Institute of Technology, ²Indian Institute of Technology Bombay

{sakshi.jadhav23, mansi.dhamne22}@spit.ac.in, prao@iitb.ac.in

Abstract

Prosodic boundary prediction from text remains underexplored for morphologically rich languages such as Hindi. We present a comprehensive study of text-based prosodic boundary prediction for Hindi children’s texts using interpretable linguistic representations. We introduce a new corpus of 54 textbook passages annotated by six fluent Hindi speakers and formulate two complementary tasks: Prosodic Boundary Detection (PBD) and Forbidden Pause Detection (FPD). Analysis of the corpus shows that 42.6% of majority-supported boundaries lack explicit punctuation, motivating the use of richer linguistic features beyond surface cues. Using a 28-feature linguistic representation, our models achieve F_1 scores of 0.89 (PBD) and 0.93 (FPD), while generalizing to an independently annotated Hindi corpus without retraining. We study the optimal feature set in terms of its differences from a previously published similar work in English, and also investigate the cross-lingual portability of the feature sets to draw insights that are supported by the structural differences between Hindi and English.

1 Introduction

Prosodic phrasing determines how words are grouped into meaningful chunks during reading, influencing comprehension of the text. Appropriate pause placement helps readers and listeners resolve syntactic structure and discourse organization, whereas misplaced pauses can fragment meaningful constituents, alter interpretation, and reduce comprehension. Accurate prediction of prosodic boundaries is therefore important for applications such as text-to-speech synthesis, expressive reading tutors, language learning systems, and accessibility technologies (Hirschberg, 2004; Taylor, 2009).

Reading fluency is associated with the production of meaningful prosodic chunks rather than word-by-word articulation. Importantly, the appropriate size of these chunks depends on the reader’s age: younger children typically benefit from shorter prosodic units, and adults naturally adopt similar chunking patterns when reading aloud to younger audiences (Dhamne et al., 2025). Consequently, assessing oral reading fluency (ORF) requires reference information pertaining not only to where pauses should be inserted, but also where they should be avoided in order to prevent adversely affecting comprehension (Sabu and Rao, 2024). This motivates the complementary tasks of Prosodic Boundary Detection (PBD) and Forbidden Pause Detection (FPD) considered in this work.

While punctuation in text provides useful cues to prosodic structure, it is an imperfect proxy for human pause placement. Fluent readers frequently insert pauses where no punctuation exists and ignore punctuation where no perceptible prosodic break is intended (Clifton et al., 2002). The mismatch is particularly evident in Hindi, which uses considerably fewer intra-sentence punctuation marks compared to English despite expressing comparable syntactic and discourse structure. This makes Hindi an ideal setting for investigating the prediction of prosodic phrasing directly from linguistic structure rather than relying punctuation alone.

Most existing work on prosodic boundary prediction has focused on acoustic modeling (Wightman and Ostendorf, 1994) or on text-based prediction for English (Rosenberg et al., 2012; Zheng et al., 2018; Dhamne et al., 2025). Relatively little attention has been paid to text-only prediction for Hindi, and even less to understanding which linguistic cues contribute to prosodic phrasing or whether such cues generalize across languages. These questions are particularly relevant for multi-

*Equal contribution

lingual speech technologies, where large annotated corpora are often unavailable and transferable linguistic representations are desirable.

In this work, we investigate prosodic boundary prediction for Hindi children’s reading passages using text-derived linguistic features. We introduce a newly annotated corpus of 54 stories spanning Grades 3–8 and develop interpretable models for the complementary tasks of Prosodic Boundary Detection (PBD) and Forbidden Pause Detection (FPD). Beyond prediction, we use the models to examine how punctuation, syntax, and grammatical structure contribute to prosodic phrasing, evaluate generalization across independently annotated data, and investigate cross-lingual transfer between Hindi and English. Our contributions are summarized as follows:

- We present a new corpus of Hindi children’s reading passages annotated by six fluent speakers for perceived prosodic boundaries and formulate two complementary prediction tasks: Prosodic Boundary Detection (PBD) and Forbidden Pause Detection (FPD)¹.
- We present the first comprehensive study of interpretable text-derived linguistic features for Hindi prosodic phrasing. Our analysis shows that punctuation is a strong individual cue but is insufficient on its own, while syntactic and grammatical features provide substantial complementary information. We demonstrate that the proposed linguistic representation generalizes across annotation settings and languages by evaluating on an independently annotated Hindi corpus and through zero-shot transfer between Hindi and English, uncovering systematic cross-linguistic differences in prosodic phrasing.
- We provide linguistic analyses of punctuation usage, feature importance, and model errors, offering new insights into the relationship between grammatical structure and prosodic phrasing in Hindi and English children’s texts.

2 Related Work

Prosodic phrasing in spontaneous or in read-aloud speech is closely related to syntactic structure and can therefore be predicted from text alone (Cole

et al., 2010; Selkirk, 2011). Text-based boundary prediction has used POS, syntactic, and contextual features in English (Read and Cox, 2007; Watts et al., 2011; Bell et al., 2006; Braunschweiler and Maia, 2016), Mandarin (Zhang et al., 2016), and Bengali (Ghosh and Rao, 2012).

For Hindi and other low-resource Indian languages, prior pause-prediction work has used relatively limited linguistic cues: Vadapalli et al. (2013) used word-terminal syllable identity, while Ghone et al. (2018) directly trained a pause prediction model with data derived from ASR alignment of the text with speech. Hindi linguistic resources nevertheless provide useful structural cues, including clause boundaries (Sharma et al., 2013) and function-word patterns involving case markers, pronouns, and conjunctions (Rao and Yegnanarayana, 1991). Hindi prosody has otherwise been studied predominantly acoustically (Roy, 2017; Jyothi et al., 2014; Pereira et al., 2026), with applications to children’s oral reading fluency assessment (Sabu and Rao, 2024). Text-only Hindi prosodic boundary prediction remains comparatively underexplored.

The closest predecessor is Dhamne et al. (2025), who use linguistically motivated text-only features for prosodic boundary and forbidden-pause prediction in English children’s texts. We extend this approach to Hindi and examine whether structural cues transfer across Hindi and English, whose differing word order and use of postpositions may affect the mapping between syntax and prosody (Dryer, 2013).

3 Data and Methods

3.1 Corpus and Annotation

We use three complementary corpora in this study including two from previously published work. Our primary dataset is a newly annotated Hindi corpus of children’s reading passages collected from Indian school textbooks. The two other corpora are used for external validation and cross-linguistic analyses: (i) the Hindi–English parallel corpus of Pereira et al. (2026), comprising Hindi children’s stories and their English translations, and (ii) the English children’s reading corpus of Dhamne et al. (2025). Table 1 summarizes the corpus characteristics, annotation protocols, and punctuation statistics of all datasets.

Our primary corpus comprises 54 short Hindi passages spanning Grades 3–8 (nine stories per

¹[Link to dataset repository](#)

Dataset	Corpus			Annotation	Punctuation & Prosody		
	Lang.	Stories	Tokens	Scheme	Full Stops	Commas	IntP
This Work	Hindi	54	8,467	Votes (0–6), ≥ 4	751	217	938
Dhamne et al. (2025)	English	54	8,634	Votes (0–7), ≥ 5	701	430	875
Pereira et al. (2026)	Hindi	9	1,191	NB/intP/IP, $\geq 3/5$	113	31	134
Pereira et al. (2026)	English	9	1,064	NB/intP/IP, $\geq 3/5$	111	46	164

Table 1: Summary of the corpora used in this study. IntP counts refer to intermediate phrase boundaries occurring within sentences and exclude sentence-final intonational phrase (IP) boundaries, although they may coincide with non-terminal punctuation (e.g., commas). The table also reports corpus statistics, annotation protocols, and punctuation frequencies for the Hindi and English datasets.

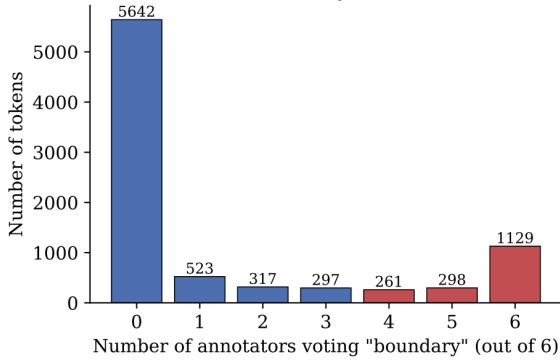


Figure 1: Distribution of boundary vote counts assigned by the six annotators across the Hindi corpus. The distribution is strongly bimodal, with most words receiving either no boundary votes or unanimous agreement.

grade). The corpus contains 8,467 word tokens across 750 sentences and is designed to reflect the increasing linguistic complexity of school reading material. Average story length increases from 132 words in Grade 3 to 177 words in Grade 8, while average sentence length increases from 10.4 to 16.0 words.

Six university students (aged 20–25) independently annotated all passages using a custom web-based annotation tool. All annotators were fluent native Hindi speakers who use Hindi as their primary language of communication. They were instructed to mark every word after which they would naturally pause when reading the passage aloud to children. Annotators were blinded to the grade level of each passage and completed the task independently. Each word therefore received a boundary score between 0 and 6, representing the number of annotators who marked a pause after it. The resulting annotations achieved substantial inter-annotator agreement (Fleiss’ $\kappa = 0.74$). Figure 1 shows the resulting vote distribution. The annotations are strongly bimodal: 66.6% of words receive zero votes, whereas 13.3% receive unani-

mous agreement from all six annotators. Throughout this work, majority agreement (≥ 4 votes) is used to define reference prosodic boundaries unless stated otherwise.

The remaining two corpora provide complementary evidence for evaluating the generality of our findings across annotation protocols and languages. The corpus of Pereira et al. (2026) adopts three-way annotation scheme (NB/intP/IP) based on the human perception of read aloud utterances by native speakers. The bidirectional translation permits the direct comparison of intermediate (intP) and intonational phrase distributions (IP) between Hindi and English, both as distinct from no boundary (NB) words. The English corpus of Dhamne et al. (2025) closely parallels the Hindi corpus of the present work which was patterned after it.

Table 1 also highlights important cross-linguistic differences in punctuation and prosodic phrasing. While full-stop counts, which broadly correspond to sentence-final intonational phrase (IP) boundaries, remain comparable across Hindi and English corpora, English consistently employs substantially more commas. However, the observation of fewer commas in Hindi does not carry over to the number of intermediate boundaries (intP), which is seen to be close to that of English. In our Hindi corpus, among the 1,688 majority-based boundaries (≥ 4 votes), 42.6% carry no punctuation cue whatsoever, while only 12.8% coincide with non-terminal punctuation (e.g., commas). Similarly, in the parallel corpus, only 28 of the 134 Hindi intP boundaries (20.9%) coincide with commas, leaving nearly four-fifths (79.1%) without any explicit punctuation cue. Together, these observations indicate that punctuation alone is an insufficient predictor of Hindi prosodic phrasing, motivating the richer text-based feature set introduced in Section 3.3.

3.2 Prediction Tasks

We formulate two supervised prediction tasks based on the annotation scheme described in Section 3.1. Both tasks aim to predict a fluent speaker’s prosody directly from text.

Prosodic Boundary Detection (PBD). This task identifies words that should be followed by a prosodic boundary. A word is labeled as a boundary if it receives majority agreement from the annotators (at least 4 out of 6 votes), corresponding to positions with strong consensus for pausing.

Forbidden Pause Detection (FPD). This task identifies words not expected to be followed by a boundary. Words receiving zero boundary votes are labeled as forbidden pauses, representing positions which all annotators unanimously agreed were the least favoured locations for pausing, possibly disrupting the natural flow and comprehension of the text.

3.3 Features

We represent each word using text-derived linguistic features only, without relying on acoustic, prosodic, or speech-derived information. We initially extracted 95 features spanning lexical, morphological, part-of-speech, syntactic, positional, punctuation, semantic, contextual, and structural information. After the initial feature engineering and exclusions, the resulting 91-feature representation was used as the baseline for sequential feature-group ablation. As shown in Appendix A, the representation was reduced to 28 features with little change in predictive performance. The final feature set retains part-of-speech, syntactic, positional, punctuation, and semantic information.

POS tags and dependency parses are obtained using Stanza (Qi et al., 2020). POS features encode the grammatical category of a token and its immediate context. Syntactic features capture dependency relations, subtree structure, clause transitions, and coordination patterns. Positional features indicate sentence position, while punctuation features describe the local punctuation context around each word. Semantic features are obtained from contextual MuRIL embeddings (Khanuja et al., 2021), which are mean-pooled at the word level and clustered using k-means ($K = 12$); each word is represented by its cluster identity together with distance- and similarity-based statistics. The complete 28-feature inventory is provided in Appendix A.3.

3.4 Training and Evaluation

We employ the XGBoost classifier (Chen and Guestrin, 2016), which is well suited to heterogeneous tabular features and naturally models non-linear feature interactions with minimal preprocessing. An additional advantage is its gain-based feature importance scores, which provide an interpretable view of the linguistic cues driving the predictions. Separate classifiers are trained for Prosodic Boundary Detection (PBD) and Forbidden Pause Detection (FPD) using the same feature representation and training procedure. Both tasks are formulated as binary classification problems using the logistic objective.

The 54 stories are partitioned at the story level into a training set (48 stories) and a held-out test set (6 stories, one from each grade), ensuring that no story appears in both partitions. This prevents lexical and contextual overlap between training and evaluation data while providing an unbiased assessment of generalization to unseen passages.

Model performance is assessed using five-fold GroupKFold cross-validation, with stories serving as the grouping variable to prevent information leakage across folds. Class weights are recomputed independently within each training fold to compensate for class imbalance. The final classifier is retrained on the complete training set and evaluated on the held-out test set. We report precision, recall, F_1 , and accuracy. Cross-validation results are reported as the mean and standard deviation across folds, while held-out test results are reported separately.

4 Results

4.1 Main Results

Table 2 summarizes the performance of the proposed models on both prediction tasks. The low standard deviation across the five cross-validation folds indicates that the models learn stable decision boundaries despite variation in story content. Held-out performance is consistent with the cross-validation estimates, suggesting that the selected model generalizes well to unseen stories. PBD achieves an F_1 score of 0.89 with high recall (0.93), demonstrating that most human-perceived prosodic boundaries can be recovered from text alone. FPD achieves an F_1 score of 0.93, indicating that positions where pausing should be avoided are also highly predictable from linguistic cues. Together, these results show that both positive and

Metric	PBD		FPD	
	CV	Test	CV	Test
Precision	0.84 ± 0.03	0.85	0.93 ± 0.01	0.95
Recall	0.88 ± 0.02	0.93	0.91 ± 0.01	0.90
F_1	0.86 ± 0.01	0.89	0.92 ± 0.01	0.93
Accuracy	0.94 ± 0.01	0.96	0.89 ± 0.01	0.90

Table 2: Performance of the XGBoost Classifier on Prosodic Boundary Detection (PBD) and Forbidden Pause Detection (FPD) on the Hindi 54 stories dataset. Cross-validation (CV) results reported as the mean $\pm$ standard deviation.

Configuration	Precision	Recall	F_1
P	1.00	0.63	0.77
P + C	0.92	0.68	0.78
P + C + Co	0.90	0.75	0.82
C + Co only	0.68	0.18	0.28

Table 3: Performance of the rule-based Prosodic Boundary Detection (PBD) baseline using progressively richer surface heuristics.

negative prosodic constraints can be modeled effectively without relying on acoustic information.

Rule-based Baseline. To assess whether prosodic boundaries can be recovered from simple surface cues alone, we construct an interpretable rule-based baseline for PBD using three progressively richer heuristics: punctuation (P), subordinating connectives (C), and coordinating conjunctions (Co). Unlike the proposed model, the baseline operates directly on the raw text without access to POS tags, dependency structure, or semantic features.

Table 3 shows that punctuation alone provides a strong baseline ($F_1 = 0.77$), reflecting its close association with major prosodic boundaries. Incorporating connective and coordination rules yields only a modest improvement ($F_1 = 0.82$). In contrast, removing punctuation causes recall to collapse from 0.75 to 0.18 ($F_1 = 0.28$), demonstrating that lexical heuristics alone recover only a small fraction of true boundaries. The proposed text-based model (Table 2) substantially outperforms the strongest rule-based baseline ($F_1 = 0.89$), primarily through higher recall (0.93 vs. 0.75), indicating that it learns richer linguistic cues beyond explicit punctuation and surface lexical markers.

4.2 Generalization Across Annotation Protocols

To evaluate robustness, we apply the trained PBD model, without retraining, to the independently annotated Hindi corpus of Pereira et al. (2026). This corpus differs from ours in both annotator pool (five teachers) and annotation protocol (majority threshold $\geq 3/5$). Despite these differences, the model achieves 94.4% accuracy, with a boundary-class precision of 86.2%, recall of 87.2%, and F_1 score of 0.87. Recall increases monotonically with annotator agreement, rising from 47% for boundaries receiving the minimum majority (3/5 votes) to 98% for unanimously annotated (5/5) boundaries. Thus, the model reliably recovers boundaries on which human readers strongly agree, while most remaining errors occur at inherently ambiguous locations.

4.3 Zero-Shot Cross-Lingual Transfer

The experiments above establish that the proposed representation generalizes across datasets and annotation protocols within Hindi. To evaluate whether the learned linguistic representation transfers across languages, we constructed a shared 22-feature representation comprising features that have consistent definitions and can be extracted for both Hindi and English. We implemented a common feature-extraction pipeline using Stanza and applied the same feature definitions to both languages. We excluded the semantic representation features used in the Hindi model because these features are dependent on a specific pretrained language model rather than on the shared linguistic feature definitions, making them less suitable for a controlled cross-lingual feature comparison.

For the cross-lingual experiments, the 22-feature model was trained on the 48-story training split of the source language and evaluated on the six-story held-out test split of the target language, without retraining. Thus, for Hindi→English transfer, the model was trained on 48 Hindi stories and tested on six English stories; for English→Hindi transfer, it was trained on 48 English stories and tested on six Hindi stories. For the in-language experiments, the same 48/6 story-level train/test partition was used within each language.

Table 4 reports both in-language and zero-shot cross-lingual performance using this shared representation. Restricting the Hindi representation from 28 to 22 features results in only a modest

Train	Test	Precision	Recall	F_1
Hindi	Hindi	0.77	0.95	0.85
English	English	0.80	0.93	0.86
English	Hindi	0.72	0.72	0.72
Hindi	English	0.51	0.79	0.62

Table 4: In-language and zero-shot cross-lingual Prosodic Boundary Detection (PBD) performance using the shared 22-feature representation. The upper block reports in-language performance, while the lower block evaluates zero-shot transfer without retraining. Metrics are reported for the boundary class.

reduction in in-language performance (Hindi: F_1 0.89→0.85), indicating that the shared feature set preserves most of the predictive information of the full representation.

Both transfer directions perform substantially above the trivial always-non-boundary baseline, demonstrating that many linguistic cues governing prosodic phrasing are shared across Hindi and English. However, transfer is asymmetric. Training on English and evaluating on Hindi yields an F_1 score of 0.72, whereas the reverse direction achieves only 0.62. The Hindi-trained model maintains high recall (0.79) but suffers a marked drop in precision (0.51), suggesting that it predicts boundaries in English contexts that are not typically realized as prosodic breaks. Section 5.2 examines the linguistic differences underlying this transfer asymmetry.

4.4 Feature Analysis

Figure 2 compares the contribution of each feature category to the two prediction tasks using normalized XGBoost gain. The two models rely on noticeably different linguistic cues. For PBD, punctuation is the dominant feature group, accounting for 61% of the total gain, followed by syntactic features (31%). In contrast, FPD depends primarily on syntactic structure (60%), with punctuation contributing only 28%.

This difference is consistent with the nature of the two tasks. Prosodic Boundary Detection seeks locations where readers naturally introduce pauses, making explicit punctuation an informative cue. Forbidden Pause Detection instead requires recognizing words that belong to tightly integrated syntactic constituents where pausing would disrupt comprehension. Consequently, dependency structure becomes substantially more informative than punctuation. POS, positional, and semantic fea-

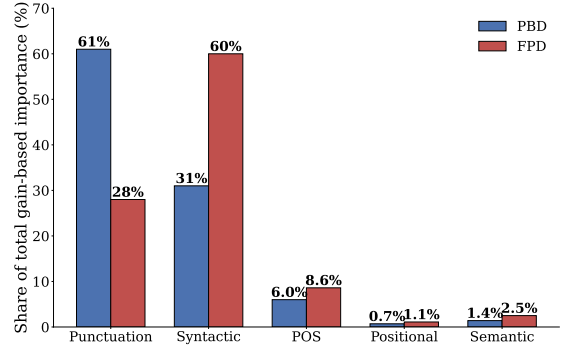


Figure 2: Relative contribution of linguistic feature categories to Prosodic Boundary Detection (PBD) and Forbidden Pause Detection (FPD), measured using normalized XGBoost gain. While punctuation dominates PBD, FPD relies primarily on syntactic features, indicating that recognizing locations where pauses should not occur depends more strongly on grammatical structure.

Feature Set	#	Precision	Recall	F_1
Punctuation	4	0.63	0.75	0.69
+ Syntactic	12	0.80	0.95	0.86
+ POS	7	0.82	0.96	0.88
+ Positional	1	0.82	0.95	0.88
+ Semantic	4	0.84	0.95	0.89

Table 5: Progressive feature-group ablation for Prosodic Boundary Detection (PBD). Feature groups are added cumulatively in decreasing order of gain-based importance.

tures contribute comparatively little to either task.

Table 5 shows that punctuation alone provides a competitive baseline ($F_1 = 0.69$), but syntactic features yield the largest improvement, followed by a smaller gain from POS features, while positional and semantic features contribute only marginally.

5 Analysis

5.1 Error Analysis

Although the PBD model achieves a recall of 0.93, the few missed boundaries are concentrated at locations with the weakest annotator agreement. Most occur at the minimum majority threshold (4/6 votes), indicating that the model primarily struggles with boundaries that are inherently ambiguous even for human readers.

GT: एक अमीर बाप ने | अपने आलसी बेटे को बुलाकर कहा कि...

Model: एक अमीर बाप ने अपने आलसी बेटे को बुलाकर कहा कि...

GT: गर्मी पड़ने लगी थी | धीरे-धीरे | उसके तेवर तीखे हो चले |

Task	Representative Example
PBD (False Positive)	जब कोई युवा पुरुष अपने घर से बाहर निकलकर बाहरी संसार [में] अपनी स्थिति जमाता है तब पहली कठिनाता उसे मित्र चुनने में आती है।
FPD (False Negative)	चूहा बच्चे की माँ के पास गया और मन [में] निश्चय किया कि वह दूध लाएगा।

Table 6: Representative postpositional constructions responsible for errors in both prediction tasks. Bracketed tokens indicate the location where the model makes an incorrect prosodic decision.

Model: गर्मी पड़ने लगी थी | धीरे-धीरे उसके तेवर | तीखे हो चले |

GT: यह हेल-मेल | बढ़ते-बढ़ते | मित्रता के रूप में परिणत हो जाता है |

Model: यह हेल-मेल बढ़ते-बढ़ते | मित्रता के रूप में परिणत हो जाता है |

Green (|) indicates a correctly predicted boundary, while red (|) indicates a boundary mismatch.

Across these examples, the missed boundaries occur at locations where the intended pause reflects discourse organization rather than a strong local syntactic or punctuation cue. Human readers introduce a boundary because a meaningful discourse unit has been completed, whereas the model instead favors a more locally coherent syntactic grouping. Such cases are inherently difficult to resolve from text alone.

A more systematic pattern emerges when examining the errors made by both prediction tasks. In PBD, 41.4% of the false positives occur immediately after common Hindi postpositions such as में, से, ने, and को, where the model incorrectly predicts a boundary inside a tightly integrated phrase. The same constructions account for 25% of FPD false negatives (25.0%), where the model incorrectly allows a pause at positions unanimously rejected by the annotators. Representative examples are shown in Table 6.

These errors arise because Hindi postpositions do not correspond to a single prosodic function. The same postposition may either mark the end of a discourse unit, where a pause is appropriate, or occur within a tightly bound argument phrase, where a pause is dispreferred. Distinguishing between these cases requires resolving the attachment and scope of the postpositional phrase within the sentence. Although our feature representation incorporates dependency-based syntactic information, postpositional attachment remains a well-known challenge for automatic Hindi parsing

Category	IP	intP
<i>Hindi</i>		
Verb / Copula	97.4%	31.3%
Postposition	0.0%	17.2%
Conjunction / Relativiser	0.0%	0.7%
Content word(noun-dominant)	2.6%	50.7%
<i>English</i>		
Content word (noun-dominant)	87.6%	81.7%
Verb / Auxiliary	8.0%	11.6%
Preposition	3.5%	3.0%
Conjunction	0.9%	3.7%

Table 7: Grammatical category of the word immediately preceding intonational phrase (IP) and intermediate phrase (intP) boundaries in the Hindi-English parallel corpus. Percentages are computed independently for each boundary type and language.

(Ambati et al., 2010). The observed errors therefore appear to reflect limitations of the underlying syntactic analysis rather than an absence of relevant linguistic information in the feature representation. Transliterations and English translations are provided in Appendix B.

5.2 Explaining Cross-Lingual Transfer

Section 4.3 showed that the shared 22-feature representation transfers across languages, but with a clear asymmetry: a classifier trained on English transfers more effectively to Hindi than the converse. To better understand this behavior, we analyze the grammatical category of the word immediately preceding prosodic boundaries in the Hindi-English parallel corpus.

Table 7 reveals a clear cross-linguistic difference in where prosodic boundaries tend to occur. In Hindi, IP boundaries almost always follow verbs or copulas (97.4%), reflecting the language’s predominantly verb-final clause structure. Intermediate phrase boundaries are more evenly distributed, occurring after verbs (31.3%) and noun-dominated content words (50.7%), with a smaller proportion following postpositions (17.2%).

English exhibits a markedly different pattern. Both IP and intP boundaries overwhelmingly follow content words (87.6% and 81.7%, respectively), while comparatively few occur after verbs or auxiliaries. Since English is not consistently verb-final, prosodic phrase boundaries are distributed across a wider range of syntactic environments than in Hindi.

These differences help explain the transfer results reported in Section 4.3. A classifier trained on Hindi learns a strong association between prosodic

boundaries and verb-final constituent completion. When applied to English, this bias produces many false positive boundary predictions, resulting in high recall but substantially lower precision. In contrast, the English-trained model learns a more distributed notion of boundary placement that transfers more naturally to Hindi, leading to the comparatively stronger English→Hindi zero-shot performance.

More broadly, these findings suggest that while the proposed feature representation captures structural cues that are shared across languages, the relative importance of those cues remains language dependent. Zero-shot transfer is therefore constrained less by the feature representation itself than by genuine differences in how prosodic phrasing aligns with syntactic structure across languages.

6 Conclusion

We presented a study of text-based prosodic boundary prediction for Hindi children’s reading passages using linguistically interpretable features. We introduced a new manually annotated text corpus and formulated two complementary prediction tasks: Prosodic Boundary Detection (PBD) and Forbidden Pause Detection (FPD), both important for ORF assessments. Using an interpretable feature representation, our models achieve strong performance on both tasks while generalizing successfully to an independently annotated corpus collected under a different annotation protocol.

Beyond predictive performance, our analyses provide several insights into the relationship between punctuation, syntax, and prosody. We show that although punctuation is the strongest individual cue, it is insufficient to explain human prosodic phrasing in Hindi, where a substantial proportion of intermediate phrase boundaries occur without any explicit punctuation. Instead, syntactic and grammatical features provide complementary information that significantly improves prediction beyond surface cues alone. Furthermore, our cross-lingual experiments demonstrate that many of these linguistic cues transfer across Hindi and English, while also revealing systematic language-specific differences in how prosodic boundaries align with grammatical structure.

Taken together, our findings suggest that, even in a language where punctuation provides only partial guidance, much of human prosodic phras-

ing is already encoded in the grammatical structure of the text itself. Looking ahead, extending this framework to additional Indian languages will help distinguish language-specific from language-independent prosodic cues and support the development of multilingual prosody prediction systems. Another promising direction is to combine interpretable linguistic features with contextual neural representations, enabling improved prediction (at the cost of losing some linguistic interpretability). Finally, integrating textual and acoustic information could facilitate richer modeling of boundary strength, prominence, and expressive reading, benefiting applications such as expressive text-to-speech synthesis, computer-assisted reading instruction, and speech accessibility technologies.

Limitations

Our objective in this work is not to maximize predictive performance using large neural encoders, but to understand the linguistic signals underlying prosodic phrasing. Recent contextual language models such as MuRIL, IndicBERT, or multilingual transformer encoders could potentially improve prediction accuracy by learning richer lexical and contextual representations. However, such representations are difficult to interpret, making it challenging to attribute predictions to specific linguistic phenomena or to analyze which cues transfer across languages.

Accordingly, we deliberately adopt an interpretable feature-based representation comprising syntactic, grammatical, positional, semantic, and punctuation-derived features. This enables feature attribution, controlled ablation studies, and detailed linguistic analyses that would be considerably less transparent with dense neural embeddings. The resulting representation also permits direct comparison across Hindi and English by constructing a shared language-agnostic feature space.

Another limitation concerns the interpretation of the cross-lingual transfer results. While we observe an asymmetry between English-to-Hindi and Hindi-to-English transfer, this finding requires further validation because the English and Hindi corpora were annotated by different annotator pools and employ different majority-voting thresholds. Consequently, the observed transfer differences may reflect not only underlying linguistic differences between the two languages but also varia-

tions in dataset characteristics and annotation protocols.

Beyond cross-lingual evaluation, our experiments are also restricted to children’s reading material. Although this domain provides relatively controlled syntax and vocabulary, prosodic strategies may differ in conversational speech, news reading, literary text, or spontaneous dialogue. Finally, our analysis focuses on textual prediction alone and does not incorporate acoustic information, which would likely improve prediction of discourse-driven boundaries that are only weakly signaled by syntax.

References

- Bharat Ram Ambati, Samar Husain, Joakim Nivre, and Rajeev Sangal. 2010. [On the role of morphosyntactic features in Hindi dependency parsing](#). In *Proceedings of the NAACL HLT 2010 First Workshop on Statistical Parsing of Morphologically-Rich Languages*, pages 94–102, Los Angeles, CA, USA. Association for Computational Linguistics.
- Peter Bell, Tina Burrows, and Paul Taylor. 2006. [Adaptation of prosodic phrasing models](#). In *Proceedings of Speech Prosody*.
- Norbert Braunschweiler and Rannieri Maia. 2016. [Pause prediction from text for speech synthesis with user-definable pause insertion likelihood threshold](#). In *Proceedings of Interspeech*, pages 3191–3195.
- Tianqi Chen and Carlos Guestrin. 2016. [XGBoost: A scalable tree boosting system](#). In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794. ACM.
- Charles Clifton, Katy Carlson, and Lyn Frazier. 2002. [Informative prosodic boundaries](#). *Language and Speech*, 45(2):87–114.
- Jennifer Cole, Yoonsook Mo, and Soondo Baek. 2010. [The role of syntactic structure in guiding prosody perception with ordinary listeners and everyday speech](#). *Language and Cognitive Processes*, 25(7–9):1141–1177.
- Mansi Dhamne, Sneha Raman, and Preeti Rao. 2025. [Predicting prosodic boundaries for children’s texts](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 31875–31885, Suzhou, China. Association for Computational Linguistics.
- Matthew S. Dryer. 2013. [Order of subject, object and verb](#). In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.
- Atish Shankar Ghone, Rachana Nerpagar, Pranaw Kumar, BiraChandra Singh, Prakash B. Pimpale, and Sasikumar M. 2018. Automatic pause boundary and pause duration detection for text-to-speech synthesis systems in Indian languages. In *Proceedings of the 15th International Conference on Natural Language Processing*, pages 28–34, International Institute of Information Technology, Hyderabad, India. NLP Association of India.
- Krishnendu Ghosh and K. Sreenivasa Rao. 2012. [Data-driven phrase break prediction for Bengali text-to-speech system](#). In *Contemporary Computing*, volume 306 of *Communications in Computer and Information Science*, pages 118–129, Berlin, Heidelberg. Springer.
- Julia Hirschberg. 2004. [Pragmatics and intonation](#). In Laurence R. Horn and Gregory Ward, editors, *The Handbook of Pragmatics*, pages 515–537. Blackwell Publishing, Malden, MA.
- Preethi Jyothi, Jennifer Cole, Mark Hasegawa-Johnson, and Vandana Puri. 2014. [An investigation of prosody in hindi narrative speech](#). In *Proceedings of the 7th International Conference on Speech Prosody*, pages 623–627. International Speech Communication Association.
- Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, Shruti Gupta, Subhash Chandra Bose Gali, Vish Subramanian, and Partha Talukdar. 2021. [MuRIL: Multilingual representations for indian languages](#). *Preprint*, arXiv:2103.10730.
- Mildred Pereira, Preeti Rao, and Hansjörg Mixdorff. 2026. [Comparing acoustic cues to phrase boundary in Hindi and Indian English](#). In *Proceedings of Speech Prosody 2026*, pages 476–480.
- Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. [Stanza: A Python natural language processing toolkit for many human languages](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108. Association for Computational Linguistics.
- G. V. Ramana Rao and B. Yegnanarayana. 1991. [Word boundary hypothesization in hindi speech](#). *Computer Speech & Language*, 5(4):379–392.
- Ian Read and Stephen Cox. 2007. [Stochastic and syntactic techniques for predicting phrase breaks](#). *Computer Speech & Language*, 21(3):519–542.
- Andrew Rosenberg, Raul Fernandez, and Bhuvana Ramabhadran. 2012. [Phrase boundary assignment from text in multiple domains](#). In *Proceedings of the 13th Annual Conference of the International Speech Communication Association (INTER-SPEECH)*, pages 2558–2561, Portland, OR, USA.

- Somnath Roy. 2017. A systematic review of hindi prosody. *arXiv preprint arXiv:1705.03247*.
- Kamini Sabu and Preeti Rao. 2024. [Predicting children’s perceived reading proficiency with prosody modeling](#). *Computer Speech & Language*, 84:101557.
- Elisabeth Selkirk. 2011. [The syntax-phonology interface](#). In John Goldsmith, Jason Riggle, and Alan C. L. Yu, editors, *The Handbook of Phonological Theory*, 2nd edition, pages 435–484. Wiley-Blackwell.
- Rahul Sharma, Soma Paul, Riyaz Ahmad Bhat, and Sambhav Jain. 2013. Automatic clause boundary annotation in the Hindi treebank. In *Proceedings of the 27th Pacific Asia Conference on Language, Information, and Computation (PACLIC 27)*, pages 499–504, Taipei, Taiwan.
- Paul Taylor. 2009. *Text-to-Speech Synthesis*. Cambridge University Press, Cambridge, UK.
- Anandaswarup Vadapalli, Peri Bhaskararao, and Kishore Prahallad. 2013. Significance of word-terminal syllables for prediction of phrase breaks in text-to-speech systems for indian languages. In *Proceedings of the 8th ISCA Speech Synthesis Workshop*. International Speech Communication Association.
- Oliver Watts, Junichi Yamagishi, and Simon King. 2011. Unsupervised continuous-valued word features for phrase-break prediction without a part-of-speech tagger. In *Proceedings of Interspeech*, pages 2157–2160.
- Colin W. Wightman and Mari Ostendorf. 1994. [Automatic labeling of prosodic patterns](#). *IEEE Transactions on Speech and Audio Processing*, 2(4):469–481.
- Zhengchen Zhang, Fuxiang Wu, Chenyu Yang, Minghui Dong, and Fugen Zhou. 2016. [Mandarin prosodic phrase prediction based on syntactic trees](#). In *Proceedings of the 9th ISCA Speech Synthesis Workshop (SSW9)*, pages 160–165.
- Yibin Zheng, Jianhua Tao, Zhengqi Wen, and Ya Li. 2018. Blstm-crf based end-to-end prosodic boundary prediction with context sensitive embeddings in a text-to-speech front-end. In *Interspeech*, pages 47–51.

Feature group	Features
A-priori excluded	word, lemma, prev_word, next_word
Orthographic	has_visarga, has_chandrabindu, has_hyphen, ends_with_matra, has_anusvara
Syllable-based	n_syl, syllables_since_last_major_boundary_child, syllables_remaining_till_breath_child, breath_pause_flag_child, is_monosyllabic, is_polysyllabic
Corpus-derived	word_freq, word_log_freq, word_idf, is_rare, unigram_surprisal, bigram_surprisal, pmi_prev, pmi_next
Quote	is_quote_boundary, precedes_quote_open, precedes_quote_close, follows_quote_open, follows_quote_close, inside_quotes, follows_quote
Word length	word_length, prev_word_length, next_word_length
Function / lexical class	is_function_word, is_content_word, is_stopword, is_postposition_lex
Contextual POS	prev_is_verb, next_is_verb, prev_is_noun, next_is_noun, prev_is_adjective, next_is_adjective, prev_is_postposition, next_is_postposition, prev_is_conjunction, next_is_conjunction
Dependency magnitude / head context	dependency_depth, distance_to_head, subtree_size, head_upos, head_dep_relation
Structural / punctuation	follows_exclamation, follows_semicolon, is_sentence_end, follows_dash, precedes_dash, follows_question, follows_comma, follows_danda, is_between_conjunctions
Positional	tokens_to_story_end, word_index_in_sentence, sentence_length_in_tokens, normalized_position_in_sentence, tokens_to_sentence_end, word_position_normalized
Total removed	67 features

Table 8: Feature groups removed during feature reduction. Four lexical identity features were excluded *a priori*; the remaining groups were discarded through sequential ablation.

A Feature Inventory

This appendix provides the complete feature inventory used by the proposed models. We first summarize the feature groups removed during the feature engineering process and then list the final 28-feature representation used for the Hindi experiments. The reduced 22-feature language-agnostic representation used for cross-lingual transfer is described in Section 4.3.

A.1 Feature Groups Removed

The initial feature engineering pipeline contained 95 text-derived features spanning lexical, orthographic, syllabic, corpus-derived, contextual, positional, syntactic, and punctuation-based information. Four lexical identity features were excluded *a priori*, while the remaining feature groups were removed through sequential ablation, yielding the final 28-feature representation used throughout the paper. Table 8 summarizes the discarded feature groups.

A.2 Sequential Feature Reduction

Feature groups were removed sequentially rather than individually to identify a compact representation while preserving predictive performance. Most reductions produced only minor changes in Boundary F_1 , indicating substantial redundancy in the initial feature set. The largest degradation occurred when sentence-position features were removed prematurely; these features were therefore

restored and re-evaluated after the remaining reductions. Table 9 summarizes the complete reduction sequence.

A.3 Final Feature Inventory

The final Hindi representation consists of 28 linguistically motivated features spanning POS, dependency structure, sentence position, punctuation context, and semantic information. Table 10 lists the complete feature inventory used by both prediction tasks.

Step	Additional group removed	Features	No-Bdy F ₁	Bdy F ₁	Δ previous
Baseline	None	91	0.9740	0.8883	–
1	Orthographic	86	0.9747	0.8908	+0.0025
2	Syllable-based	80	0.9753	0.8939	+0.0031
3	Corpus-derived	72	0.9709	0.8860	–0.0079
4	Quote features	65	0.9708	0.8872	+0.0012
5	Word-length	62	0.9695	0.8814	–0.0058
6	Function / lexical-class	58	0.9716	0.8877	+0.0063
7*	<i>Exploratory</i> : sentence/story position	52	0.9627	0.8586	–0.0291
8	Contextual POS flags	48	0.9709	0.8848	–0.0029
9	Dependency magnitude / head context	43	0.9756	0.9019	+0.0171
10	Zero-importance structural/punctuation flags	34	0.9750	0.8995	–0.0024
11	Remaining positional measures	28	0.9729	0.8924	–0.0071

Table 9: Sequential feature reduction from the initial 91-feature representation to the final 28-feature model. Step 7 denotes an exploratory positional ablation that was reverted after a substantial drop in performance.

Feature Type	Retained Features
POS	upos, is_verb, is_noun, is_postposition, is_conjunction, prev_upos, next_upos
Dependency	dep_relation, prev_dep, next_dep, is_root, is_leaf, num_children, crosses_clause_boundary_to_next, depth_change_to_next, is_first_word_of_subtree, is_last_word_of_subtree, precedes_coordinator, is_conjunct_head
Positional	is_sentence_start
Punctuation	is_punctuation_following, followed_by_boundary_punct, tokens_until_next_punct, tokens_since_last_boundary_punct
Semantic	semantic_cluster_id, semantic_cluster_dist, semantic_cluster_sim, semantic_margin
Total	28 features

Table 10: Complete inventory of the 28 linguistic features retained in the final Hindi representation after feature reduction.

B Supporting Error Analysis Examples

The following examples in Table 12 correspond to the representative postpositional constructions discussed in Table 6. They illustrate contexts where the model makes incorrect prosodic decisions around the Hindi postposition में (*mem*, “in”), which may either occur within a tightly integrated phrase or coincide with a larger discourse constituent.

Eg.	Sentence
1	Transliteration: <i>Ek amīr bāp ne apne ālasī beṭe ko bulākar kahā ki...</i> Translation: A wealthy father called his lazy son and said...
2	Transliteration: <i>Garmī parne lagī thī, dhīre-dhīre uske tevar tīkhe ho chale.</i> Translation: The weather had begun to turn hot, and gradually his temperament became harsher.
3	Transliteration: <i>Yah hel-mel baṛhte-baṛhte mitratā ke rūp meṃ pariṇat ho jātā hai.</i> Translation: This growing familiarity gradually develops into friendship.

Table 11: Hindi examples discussed in Section 5.1, with transliterations and their respective English translations.

Eg.	Sentence
PBD	Transliteration: <i>Jab koī yuvā puruṣ apne ghar se bāhar nikalkar bāharī sansār meṃ apnī sthiti jamātā hai tab pahalī kaṭhīnatā use mitra chunne meṃ ātī hai.</i> Translation: When a young man leaves his home, establishes himself in the outside world, the first difficulty he faces is choosing friends.
FPD	Transliteration: <i>Cūhā bacce kī mā ke pās gayā aur man meṃ niścaya kiyā ki vah dūdh lāegā.</i> Translation: The mouse went to the child’s mother and resolved in his mind that he would bring milk.

Table 12: Transliterations and English translations of the representative postpositional constructions discussed in Section 5.1.