

Discourse Particle *na* in Hindi: A Corpus Analysis

Meghna Hooda

LING 414: Statistical Methods in Linguistics

Abstract

This study investigates the discourse particle *na* in naturalistic Hindi speech using the IIT Delhi Dialogue Corpus (Pareek et al., 2025). Discourse particles convey pragmatic and attitudinal information beyond the propositional content of a sentence, influencing speaker stance and discourse structuring (Zimmerman, 2011). The research examines whether sentence length and position of discourse particle in the clause influence the occurrence of *na* and whether these factors interact. Using a corpus of 39,224 telephonic conversational sentences, we identify 1,555 sentences containing *na* as a discourse particle. Logistic regression models show that longer sentences increase the likelihood of *na* occurrence and that clause-final positioning significantly moderates this effect. These results provide both statistical and linguistic insights into the functional distribution of *na* in Hindi speech.

Keywords: discourse particle, Hindi, corpus analysis, logistic regression

Introduction

Discourse particles are linguistic elements that contribute meaning beyond the truth-conditional content of a proposition. Rather than altering propositional meaning, they encode information about the speaker's assumptions, attitudes, and reasoning, and about how an utterance is intended to relate to the evolving common ground between interlocutors (Zimmermann 2011). Through these functions, discourse particles play a crucial role in managing interaction, structuring discourse, and signalling how speakers expect their contributions to be interpreted within an ongoing conversation.

In Hindi, a morphologically rich language, discourse particles such as *to*, *hi*, *ji*, and *na* play key roles in spoken interaction (Deo, 2022; Bhalla & Alok, 2025). This study focuses on the distribution of a discourse particle in spoken Hindi, specifically *na*, using naturally occurring

conversational data. The discourse particle shares phonology with negation marker in Hindi *na*. We do not have enough evidence to conclude any connection between the two lexical items. Hence, in our study we exclude negation *na* from the analysis. The discourse particle on the other hand is used to convey pragmatic meaning; the preposition of the sentence remains same without the discourse particle. Despite its prevalence, *na* is underexplored in discourse particle research. Previous studies on Hindi discourse markers (O'Reilly-Brown 2022; Jabbar 2023) highlight their syntactic and pragmatic behaviour, but there is no unified account or a systematic corpus-based analyses. Corpus analysis of the discourse particle gives evidence of its usage in naturalistic speech and help us get an idea about its pattern with other linguistics features.

Research in pragmatics and discourse marker theory consistently shows that discourse particles serve multiple overlapping functions. They may index speaker stance, such as certainty, expectation, or alignment with the addressee; signal pragmatic relations such as inference or appeal to shared knowledge; and contribute to discourse structuring by marking boundaries between units of information (Zimmermann 2011; Groz 2020; Fraser 1999; Schiffrin 1987). Importantly, these functions are not uniformly distributed across utterances but are sensitive to properties of sentence structure and discourse context. In the current paper, we are interested in sentence structuring that may regulate the occurrence of discourse particle.

Literature observes that discourse particles increase increased discourse or syntactic complexity. Longer sentences often encode more discourse content, involve embedding or coordination, and introduce multiple discourse-relevant elements such as arguments, modifiers, or pragmatic inferences. As a result, longer sentences may provide greater opportunity of speakers to use discourse particles to guide interpretation and manage listener expectations. From this perspective, sentence length serves as a useful quantitative proxy for discourse complexity in corpus-based analyses.

In addition to sentence length, structural position plays a crucial role in the distribution of discourse particles. Cross-linguistic research shows that particles frequently occur at clause boundaries, particularly in clause-final positions, where they can comment on or frame the preceding proposition (Groz 2020). This pattern has been observed in languages such as German(Groz 2020) and Japanese, as well as in Hindi, where clause-final particles are often associated with pragmatic enrichment and interactional meaning (Deo 2022). Longer

sentences, by virtue of containing more clauses and boundaries, naturally create more structural sites where such particles may occur.

Henceforth, these insights motivate the central hypotheses of the present study: that the occurrence of *na* in Hindi speech is sensitive to sentence length as a proxy for discourse complexity, and that its distribution is further conditioned by clause-level structural position. By quantitatively modelling these factors, this study aims to clarify how discourse-structural and syntactic properties interact to shape the use of *na* in naturalistic Hindi conversation.

The research questions guiding this study are:

1. Does sentence length influence the occurrence of *na* in Hindi speech?
2. Does the clause position of *na* affect its probability of occurrence?
3. Do sentence length and clause position interact to influence *na* use?

Literature Review

Discourse particles serve multiple functions across languages, including marking speaker stance, organizing discourse, and signalling inferential relations (Fraser, 1999; Schiffrin, 1987). Rather than contributing to truth-conditional meaning, these elements operate at the level of discourse management, guiding interlocutors in how an utterance should be interpreted within an unfolding interaction. Their multifunctionality and context sensitivity make them particularly relevant for the study of spoken language, where speakers continuously negotiate common ground and pragmatic expectations. In Hindi, recent work has emphasized the tight connection between the syntactic position of discourse particles and their pragmatic interpretation. For example, Deo (2022) demonstrates that the particle *to* interacts with clause structure to raise and resolve questions, while Bhalla and Alok (2025) show that *ji* encodes honorificity and stance in ways that are strongly conditioned by its placement within the clause. These studies suggest that discourse particles in Hindi are not freely distributed, but instead occupy structurally privileged positions that align with their discourse functions. *na* has received comparatively limited attention, despite its high frequency in spoken Hindi. O'Reilly-Brown (2022) divide *na* as two lexical items on the basis of its position in the sentence: clause final and clause medial. She argues that *na* functions as both a tag-question-like element in

clause final and a discourse marker in clause medial, contributing to speaker inference and inviting alignment or confirmation from the addressee. Importantly, *na* is almost exclusively attested in spoken interaction, underscoring its role in managing real-time discourse rather than encoding grammatical meaning. Jabbar (2023) only focuses on clause final *na* and further develops this analysis by showing that *na* participates in reasonable inference, allowing speakers to signal expectations about shared knowledge or conclusions that follow from the discourse context. Together, these studies characterize *na* as a pragmatically rich particle whose interpretation depends heavily on discourse structure and speaker assumptions. Moreover, these researches show the importance of position of *na* in the sentence on its functionality.

Cross-linguistic research reinforces the idea that clause-final positions are preferred sites for discourse particles. Studies on German demonstrate that particles frequently appear at the right periphery of clauses, where they can comment on, evaluate, or modulate the preceding proposition (Groz, 2020). These positions are especially well-suited for particles that operate on discourse-level meaning, as they allow speakers to frame an utterance after its propositional content has been established. Hindi appears to follow a similar pattern, with clause-final *na* functioning as a marker.

Beyond position alone, discourse particles have also been shown to interact systematically with sentence-level structural properties, such as word order, clause complexity, and dependency length (Vasishth 2004). Corpus-based research on English discourse markers demonstrates that their distribution is sensitive to syntactic complexity and information structure, with higher rates of occurrence in longer and more structurally complex sentences (Haselow, 2011). From this perspective, sentence length can be understood as a proxy for discourse complexity, capturing the cumulative effect of embedding, coordination, and pragmatic layering. These findings motivate the hypothesis that discourse particles are more likely to occur in sentences that place greater interpretive demands on interlocutors.

Despite these theoretical and empirical insights, large-scale corpus-based quantitative studies of Hindi discourse particles remain scarce, particularly studies that model multiple sentence-level predictors simultaneously. Existing work on *na* has primarily relied on qualitative analysis or speaker judgment, leaving open questions about how structural factors such as sentence length and clause position jointly condition its distribution. The present study addresses this gap by using a large, naturalistic spoken corpus and statistical modelling to systematically examine the predictors of *na* occurrence in Hindi speech.

Methods

Data

The analysis utilizes the IIT Delhi Dialogue Corpus, a large-scale corpus of naturalistic Hindi speech (Pareek et al., 2025). The corpus includes:

1. Telephonic conversational data from the CALLFRIEND Hindi project (Canavan & Zipperlen, 1996)
2. Face-to-face task-oriented interactions

For this study, only manually transcribed telephonic data were used, consisting of 39,224 sentences. Of these, 1,555 sentences contain *na* as a discourse particle.

The corpus is distributed in CoNLL-U (.conllu) format, which provides token-level linguistic annotations for each sentence including: part of speech, lemma, morphological features, and syntactic dependency relations. Sentence boundaries are explicitly marked, allowing sentence-level measures such as sentence length to be calculated reliably as the number of tokens per sentence. Note that while counting the sentence length, extra linguistics substance like laughter, pauses, etc. are not included in the count. Each token in the corpus is associated with a unique identifier this allows to determine the relative position of *na* within a clause. Clause-final *na* is operationalized as instances in which *na* occurs before a punctuation.

Importantly, the corpus does not explicitly annotate discourse particles as a separate category. As a result, instances of *na* were identified through string-based searches. Because *na* can also function as a negation marker when it appears before the verb, all such sentences negation uses(67) were excluded from the analysis.

Statistical Analysis

To examine the factors conditioning the occurrence of the discourse particle *na* in Hindi speech, a series of logistic regression models were fitted. Logistic regression was selected as the appropriate analytical framework because the dependent variable is binary, indicating whether

na is present or absent in a given sentence. The dependent variable was coded as 1 for sentences containing discourse-particle *na* and 0 for sentences without *na*.

Two primary predictors were included in the analysis. The first independent variable was sentence length, treated as a continuous variable. Sentence length was used as a quantitative proxy for syntactic and discourse complexity, under the assumption that longer sentences tend to encode more pragmatic content and discourse structure. The second independent variable was the clause-final position of *na*, encoded as a binary categorical variable, with clause-final occurrences coded as 1 and non-clause-final occurrences coded as 0. This variable captures the well-documented sensitivity of discourse particles to clause-peripheral positions.

The analysis proceeded in two stages. Model 1 included sentence length as the sole predictor, allowing for an assessment of the baseline relationship between sentence length and the probability of *na* occurrence. Model 2 extended this analysis by including clause-final position as an additional predictor, as well as an interaction term between sentence length and clause position. The inclusion of the interaction term made it possible to test whether the effect of sentence length on *na* occurrence differs depending on the structural position of the particle within the clause.

Model fit and statistical significance were evaluated using standard diagnostics for generalized linear models. Predictor effects were assessed using estimated coefficients, standard errors, z-values, and associated p-values. Overall model fit was evaluated by comparing residual deviance values across models, with lower residual deviance indicating improved fit to the observed data. All statistical analyses were conducted in the R programming environment, using generalized linear modelling functions suitable for binomial outcome variables.

Results

Descriptive Statistics

The dataset consists of a total of 41,979 sentences, of which 1,457 sentences contain the discourse particle *na*, while 40,415 sentences do not. Among the sentences with *na*, 67 utterances are classified as negative *na* which are excluded from the analysis. We also observed that Clause-final *na* appears in 68% of sentences containing *na*.

Sno,	Sentence Type	Sentences
1	With <i>na</i>	1,457
2	Without <i>na</i>	40,415
3	Total (1+2)	41,917
4	With neg <i>na</i>	67

Table 1: Number of sentences obtained from the corpus

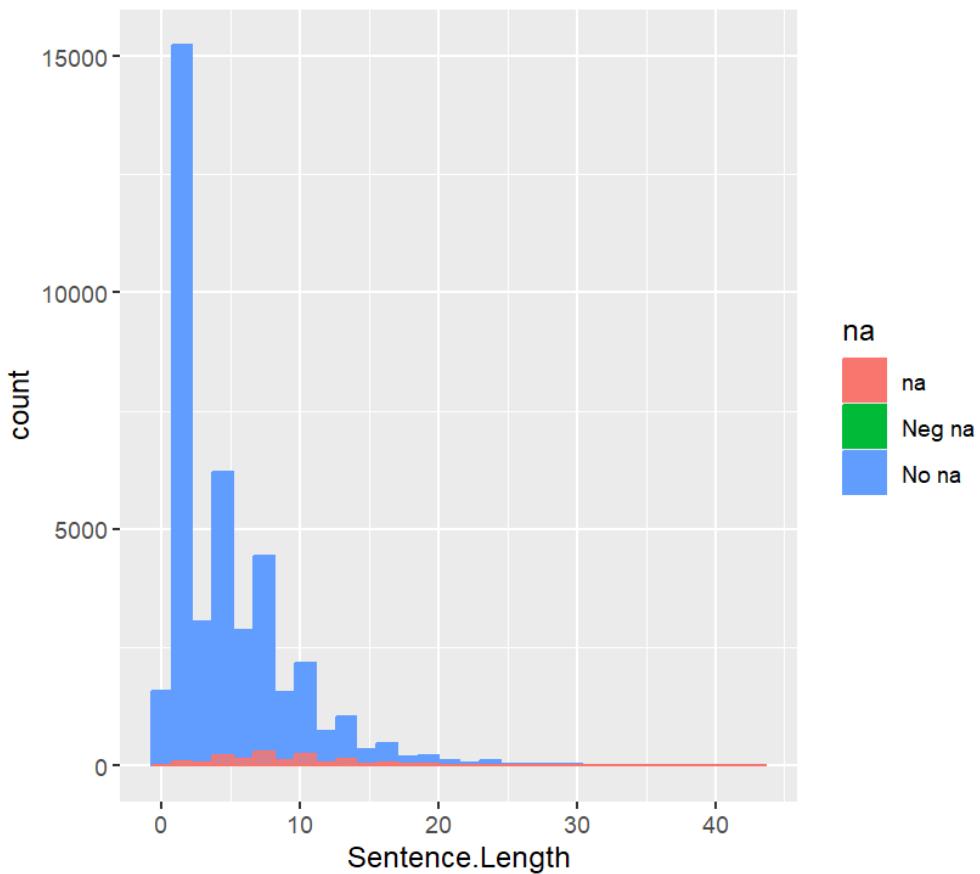


Figure 1: Distribution of sentence length across the corpus in three Sentence Type. Sentences containing negation *na* is very small hence not visible on the histogram.

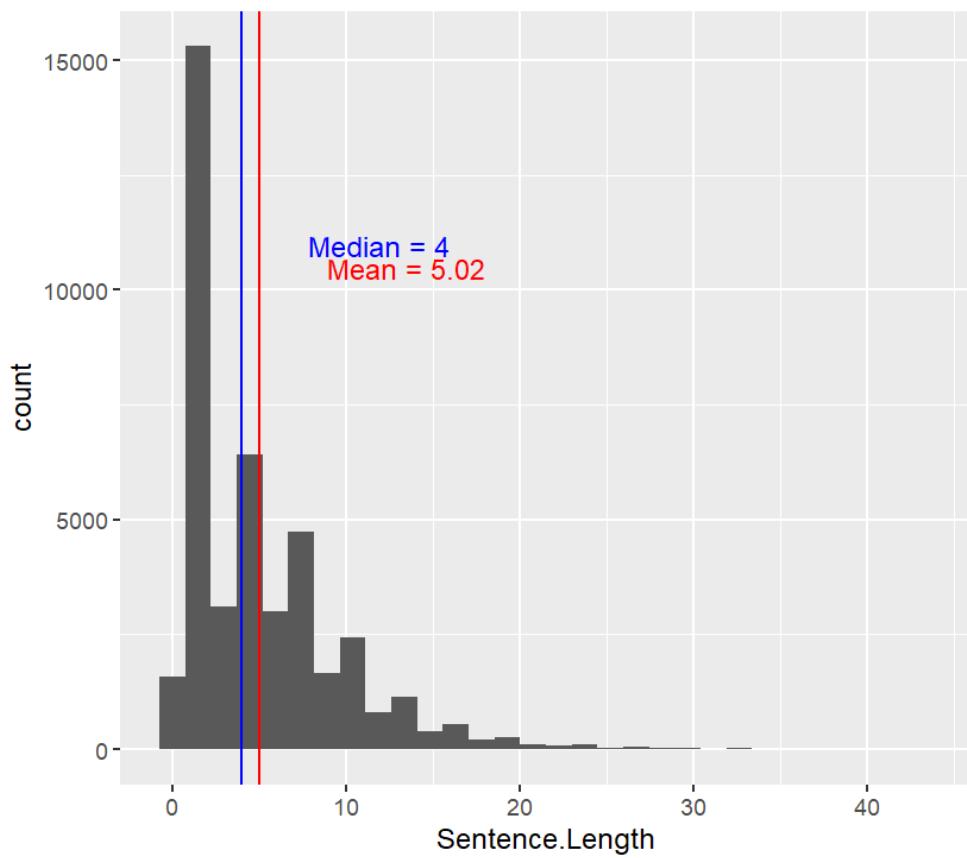


Figure 2: Distribution of sentence length in the corpus

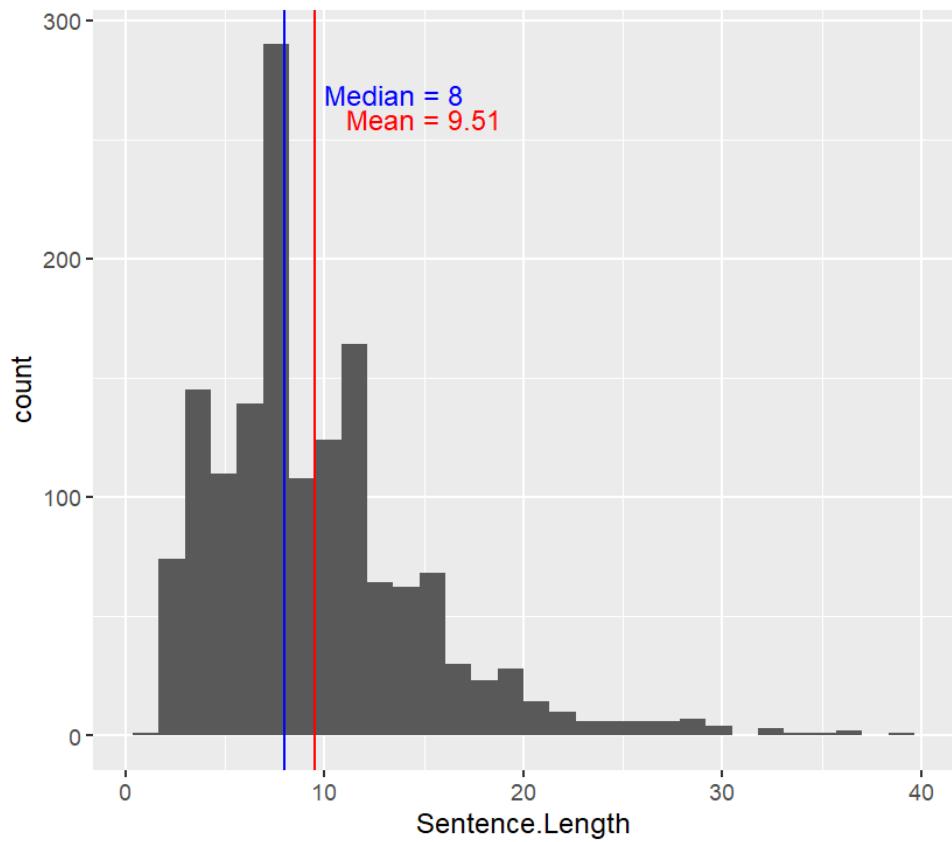


Figure 3: Distribution of sentence length in sentences with *na*.

Figure 1 and 2 shows notable difference between sentences that contain *na* and those that do not. Sentences containing *na* have an average length of 8 words, which is higher than the average length of sentences without *na* (4 words). This indicates that sentences with *na* tend to be longer, suggesting that the presence of *na* may be associated with more complex or elaborated syntactic structures. The fact that sentences containing *na* are, on average, longer than those without *na* provides preliminary support for our hypothesis that the occurrence of *na* is influenced by sentence length.

Logistic Regression

Model 1: Sentence Length → Presence of *na*

Coefficient	Estimate	Std. Error	z-value	p-value
Intercept	-4.234	0.044	-95.58	<2e-16
Sentence Length	0.138	0.004	34.34	<2e-16

Model 1 includes a single predictor: Sentence Length. The model estimates a highly significant positive effect of sentence length on the probability that a sentence contains *na* ($\beta = 0.138$, $z = 34.34$, $p < 2e-16$). This coefficient indicates that for each additional unit increase in sentence length (i.e., each extra word), the log-odds of encountering *na* increase by approximately 0.138. The fitted curve is shown in Figure 4. The narrow confidence band around the fitted curve indicates that the estimate is stable across sentence lengths despite variability in raw observations. The model's residual deviance (11872 on 41908 degrees of freedom) indicates good fit for a binary categorical outcome, further supporting the reliability of the sentence-length predictor. Overall, Model 1 provides strong statistical evidence that longer sentences are more likely to contain the discourse particle *na*.

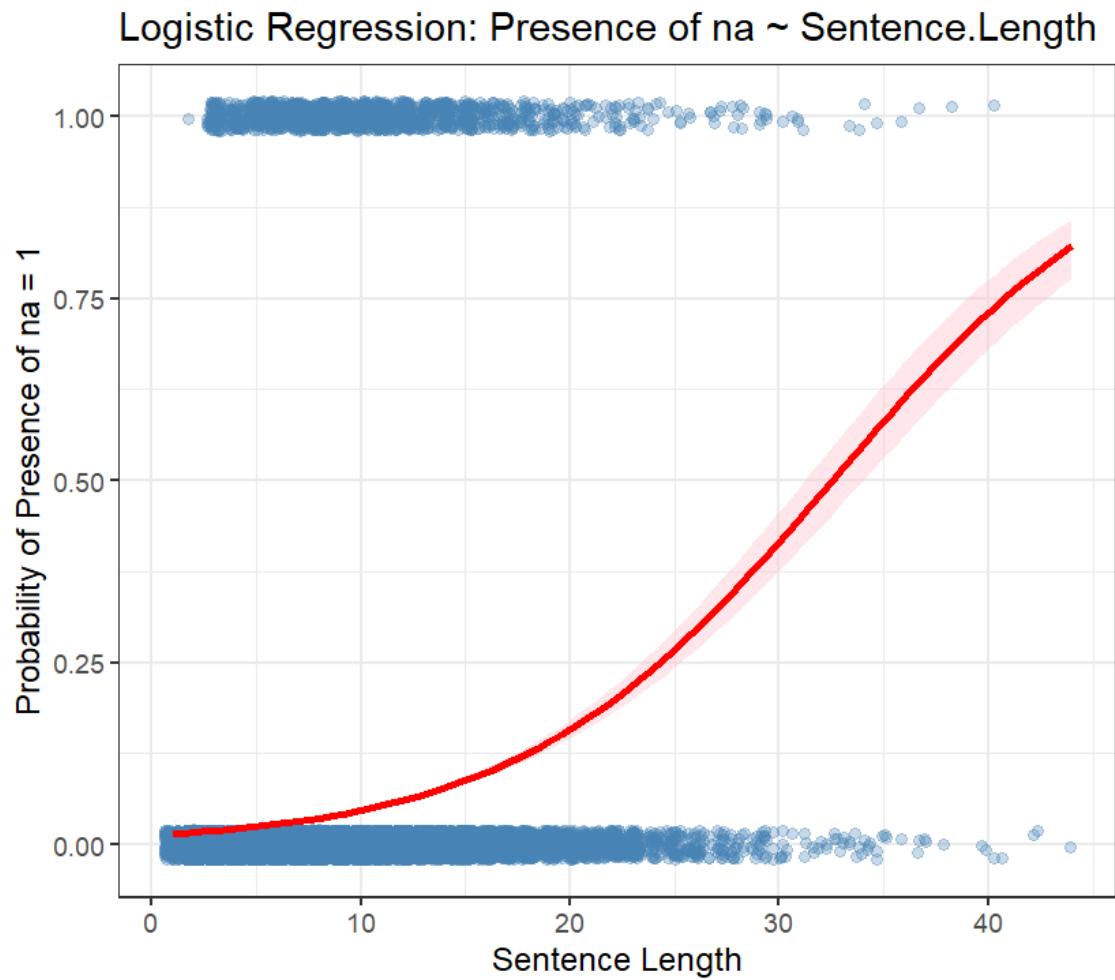


Figure 4: Visualising logistic regression curve showing the probability presence of na as a function of sentence length. The line presents the model prediction, and the points represent sentence length.

Model 2: Sentence Length + Clause-final Position + Interaction

Coefficient	Estimate	z-value	p-value
Intercept	-4.928	-65.33	<2e-16
Sentence Length	0.186	31.36	<2e-16
Clause-final	1.242	13.28	<2e-16
Sentence.Length × Clause-final	-0.091	-10.83	<2e-16

While Model 1 establishes a positive association between sentence length and the presence of *na*, discourse particles are also known to be sensitive to structural position. To account for this, Model 2 incorporates (i) Clause-Final Position (0 = not clause final, 1 = clause final) and (ii) the interaction between Sentence Length and Position. The fitted logistic curves for each position are shown in Figure 6.

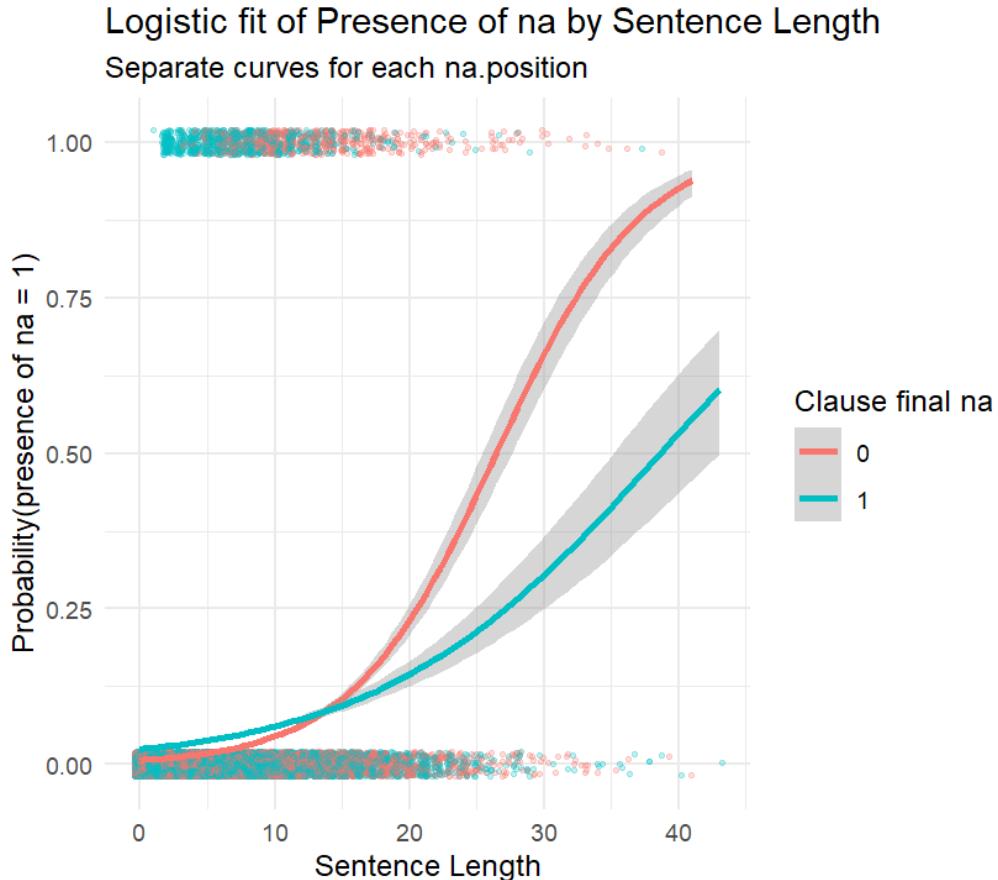


Figure 6: Visualising logistic regression curve showing the probability presence of *na* as a function of sentence length. The line presents the model prediction, and the points represent sentence length.

The model estimates reveal several critical patterns:

1. Main effect of Sentence Length: The positive effect of sentence length remains robust ($\beta = 0.186, p < 2e-16$), indicating that, regardless of position, longer sentences increase the likelihood of *na*.
2. Main effect of Clause-Final Position Clause-final *na* is significantly more probable than non-final *na* ($\beta = 1.242, p < 2e-16$). This corresponds to more than a 3.4-fold increase in odds ($\exp(1.242) \approx 3.46$). This means that, all else equal, a sentence containing *na* at

the clause boundary is much more likely than a sentence containing *na* in a non-final position.

3. Interaction between Sentence Length and Position: The negative interaction ($\beta = -0.091, p < 2e-16$) indicates that the effect of sentence length is weaker for clause-final *na*.

Discussion

The findings indicate that both sentence length and clause-final position significantly influence the occurrence of *na* in Hindi speech. Longer sentences generally provide more opportunities for pragmatic marking, consistent with prior studies of discourse complexity (Fraser, 1999; Groz, 2020). Clause-final *na* enhances this effect, highlighting the importance of syntactic position in discourse marking. The interaction effect suggests that the impact of sentence length is reduced when *na* occurs in clause-final position, implying that *na*'s discourse function may be particularly tied to its syntactic placement rather than sentence length alone.

Limitations of the study include the exclusive use of telephonic data, which may not capture task-oriented discourse patterns but provide a very naturalistic data. Future work could explore additional factors, such as verb transitivity, dependency length, and prosodic features, to refine models of *na* distribution. In our exploratory model we see that including verb type (transitive/intransitive) in model 1 did not significantly affect the probability of *na* ($p > 0.7$). However, with more data we possibly can get significant effect on its negative slope that we observe in our current results ($\beta = -0.023326$) with baseline of Intransitive.

Further, clause-final positioning of *na* is highly frequent: 68% of sentences containing *na* exhibit the particle at the end of the clause. This positional preference aligns with prior observations in discourse particle studies and underscores the functional role of *na* in marking clause boundaries or pragmatic nuances.

Conclusion

This study demonstrates that logistic regression provides a robust framework for modelling the occurrence of *na* in naturalistic Hindi speech. The analysis yields several key findings. First,

longer sentences are slightly more likely to contain *na*, suggesting that sentence length creates environments conducive to the use of this discourse particle. Second, clause-final positioning is a strong predictor of *na* occurrence, highlighting the importance of syntactic placement in its distribution. Third, the interaction between sentence length and clause-final position indicates that the effect of sentence length on *na* usage is moderated by its position within the clause.

These findings show the complex interplay between structural and positional factors in shaping the use of *na*. For future research, incorporating additional syntactic variable such as clause type, part of speech of words around its dependencies, etc. and working with more data could provide a more comprehensive understanding of the functional and distributional patterns of discourse particles in Hindi. Such extensions would further clarify how *na* contributes to cohesion, emphasis, and pragmatic meaning in natural speech.

Code link: <https://github.com/MeghnaHooda/Corpus-analysis-of-discourse-markers-na-in-Hindi/tree/main>

References

- Ahmad Jabbar. 2023. *The Hindi-Urdu NA and reasonable inference*. Proceedings of the 59th Meeting of the Chicago Linguistic Society (CLS 59).
- Bhalla, O., & Alok, D. 2025. *The Hindi Discourse Particle ji: Honorificity and Form-Meaning Interaction*. Proceedings of the 27th Seoul International Conference on Generative Grammar (SICOGG-27), 21–30.
- Canavan, A., & Zipperlen, G. 1996. *Callfriend Hindi LDC96S52*. Web Download. Philadelphia: Linguistic Data Consortium.
- Deo, A. 2022. *Could be stronger: Raising and resolving questions with Hindi = to*. Language, 98(4), 716–748.
- Fraser, B. 1999. *What are discourse markers?* Journal of Pragmatics, 31(7), 931–952.
- Groz, H. 2020. *Discourse Particle*. In L. Matthewson, C. Meier, H. Rullmann & T. E. Zimmermann (Eds.), *The Companion to Semantics*. Wiley.

- Haselow, A. (2011). Discourse marker and modal particle: The functions of utterance-final then in spoken English. *Journal of pragmatics*, 43(14), 3603-3623.
- Jabbar, A. 2023. *The Hindi-Urdu NA and reasonable inference*. CLS 59.
- O'Reilly-Brown, M. 2022. *Naan as a Tag Question and a Discourse Marker in Hindi-Urdu. Formal Approaches to South Asian Languages*.
- Pareek, B., Zafar, M., Hooda, M., Yadav, K., Vaidya, A., & Husain, S. 2025. *IIT Delhi Dialogue Corpus: A Quantitative Analysis of a Spoken Corpus of Hindi*. Language Resources and Evaluation. doi:10.1007/s10579-025-09867-8
- Schiffrin, D. 1987. *Discourse markers*. Cambridge University Press.
- Vasisht, S. (2004). Discourse context and word order preferences in Hindi. *Yearbook of South Asian Languages*, 113-127.
- Zimmerman, M. 2011. *Discourse Particles*. In P. Portner, C. Maienborn, & K. von Heusinger (Eds.), *Semantics* (HSK 33.2), 2011–2038. Berlin: Mouton de Gruyter.