LREC 2026

8th Workshop on Indian Language Data: Resources and Evaluation (WILDRE-8)

Workshop Proceedings

Editors
Girish Nath Jha, Kalika Bali, Sobha L, Devendr Kumar

12 May, 2026

Proceedings of WILDRE-8

ISBN 978-2-493814-37-1
EAN 9782493814890

Preface

WILDRE – the 8th Workshop on Indian Language Data: Resources and Evaluation is being organised on May 12th, 2026, under the LREC 2026 platform. India has a high degree of linguistic diversity and has seen concerted efforts by the Indian government and industry to develop language resources. ELRA Language Resources Association and its associate organisations have been very active and successful in addressing the challenges and opportunities related to language resource creation and evaluation. It is, therefore, a great opportunity for resource creators in Indian languages to showcase their work on this platform and to interact and learn from those involved in similar initiatives worldwide.

The broader objectives of the 8th WILDRE will be

- to map the status of Indian Language Resources
- to investigate challenges related to creating and sharing various levels of language resources
- to promote a dialogue between language resource developers and users
- to provide an opportunity for researchers from India to collaborate with researchers from other parts of the world

The call for papers received a good response from the Indian language technology community.

Workshop Organisers

Organizing Committee

- Girish Nath Jha, Professor, School of Sanskrit and Indic Studies, Jawaharlal Nehru University, India
- Kalika Bali, Microsoft Research India Lab, Bangalore
- Sobha L, AU-KBC, Anna University
- Devendr Kumar, MGAHU, Wardha

Table of Contents

Workshop Program

12th May 2026

09:00–09:10 **Welcome by Workshop Chairs**

09:10–09:20 **Inaugural Address by Kavita Bhatia**

09:20–10:00 **Keynote Talk by Amitabh Nag, Bhashini**

10:00–11:00 **Oral Talks I**

10:00–10:15 *MetricalARGS: Studying Metrical Poetry with LLMs*
Chalamalasetti Kranti and Sowmya Vajjala

10:15–10:30 *Semi-automatic Approach for Tamil Discourse Relation Annotation*
Frances Yung, Enosh Peter Ponraj and Vera Demberg

10:30–10:45 *Konkani Daan: A Community-Driven Culturally Grounded Speech Corpus for Low-Resource ASR*
Milind Shivolkar, Vaibhav Gawas and Jyoti Pawar

10:45–11:00 *Is Literal Annotation Enough? Building an Annotation Framework for Metonymic Named Entities in Marathi*
Pratibha Dongare

11:00–11:20 **Coffee Break**

11:20 –11:50 **Oral Talks II**

11:20–11:35 *Bengali-English and Hindi-English Code Mixed Speech Data with Disfluencies*
Anuran Mitra, Tapabrata Mondal, Anirvan Chakravarty and Sivaji Bandyopadhyay

11:35–11:50 *Konkani Wordnet Resources*
Hanumant H. Redkar, Mahadev Gawas, Anjali Desai and Jyoti Pawar

11:50–12:50 Poster Session

11:50–12:50 *Development of Speech Corpus for Low-Resource Language- A case of Sanskrit*
Devendr Kumar, Girish Nath Jha and Khalid Choukri

11:50–12:50 *The shabd portal – searchable lexical resources for Indian languages by Government of India*
mercy lalrohluo hmar, Girish Nath Jha and DHANANJAY SINGH

11:50–12:50 *POS Tagging in Low-Resource Maithili language: Specific Challenges and Nuances*
Shivani Priya, Shruti Jha, Urmila Jha, Girish Nath Jha, Deepali Tiwari and Dr. Jyoti Raj

11:50–12:50 *Preserving Civilisation Memory: A Digital Humanities Approach to The Ramayan*
Shashank Tiwari and Girish Nath Jha

11:50–12:50 *Naamah: A Large Scale Synthetic Sanskrit NER Corpus via DBpedia Seeding and LLM Generation*
Annarao Kulkarni and Akhil Rajeev P

11:50–12:50 *IndEuph-170: Benchmarking Cultural Pragmatics through Euphemism Detection in Indian English*
Debamita Samajdar

11:50–12:50 *Integrating Syntactic and Discourse Signals through Multi-Encoder Fusion in NMT for Low-Resource Indian Language Pairs*
Sobha Lalitha Devi, Vijay Sundar Ram and Pattabhi RK Rao

11:50–12:50 *NE-LID: A Fast and Accurate Language Identification System for Northeast Indian Languages*
Badal Nyalang

11:50–12:50 *Integrating Cultural Wisdom and Digital Technologies for Children's Moral and Emotional Development*
Ms. Garima and Girish Nath Jha

12:50–13:00 Closing by Sobha Lalitha Dev

MetricalARGS: Studying Metrical Poetry with LLMs

Chalamalasetti Kranti[1], Sowmya Vajjala[2]
[1]Department of Linguistics, University of Potsdam, Germany,
[2]National Research Council, Ottawa, Canada
kranti.chalamalasetti@uni-potsdam.de, sowmya.vajjala@nrc-cnrc.gc.ca

Abstract

Many classical languages have well-studied traditions of poetic meter which enforce constraints on a poem in terms of syllable and phoneme patterns. Such advanced literary forms offer opportunities for probing deeper reasoning and language understanding in Large Language Models (LLMs) and their ability to follow strict pre-requisites and rules in generating text. In this paper, we introduce MetricalARGS, the first taxonomy of poetry-related NLP tasks designed to evaluate LLMs on metrical poetry across four dimensions: **A**nalysis, **R**etrieval, **G**eneration, and **S**upport. We discuss how these tasks relate to existing NLP tasks, addressing questions around datasets and evaluation metrics. Taking the metrical poetry of Telugu language as our example, we illustrate how the taxonomy can be used with LLMs in practice through a quantitative and qualitative evaluation. MetricalARGS highlights the broader possibilities for understanding the capabilities and limitations of today's LLMs through the lens of metrical poetry. We believe MetricalARGS can also serve as a reference taxonomy for studying and comparing metrical poetry across Indian languages as a starting point, and can be extended to other languages with established metrical poetry traditions.

Keywords: Poetry Analysis, Metrical poetry, Telugu

1. Introduction

There has been a consistent interest in the generation and analysis of the poetic form in NLP and computational creativity research. Although the language of focus is mainly English, there has been some research in a few other relatively high-resource languages such as Chinese. While automated poetry generation/translation has been the most commonly studied problem, there are also several other related topics when we consider more constrained literary-linguistic systems such as meter in poetry, which are governed by their own set of rules and requirements. The language specific nature of these rules, and the customization needed in terms of datasets, evaluation methods etc was a bottleneck in extending such studies in computational creativity into other languages in NLP research.

The advent of LLMs that are capable of in-context learning with very few (or no) examples opens up a possibility of extending this research to other world languages with rich and long standing poetic traditions such as Sanskrit and other Indian languages including Telugu. Many of them involve intricate rules and formal specifications that go beyond free-form generation of poems, which we refer to as metrical poetry throughout this paper (*chandas* in Sanskrit, and similar words in other Indian languages). Existing research on poetry related topics in NLP has largely focused on individual tasks in isolation, with no unified taxonomy to connect them. In this paper, we take the first steps in addressing this gap by first creating a taxonomy of metrical poetry tasks covering four dimensions - Analysis, Retrieval, Generation, and Support, which we call

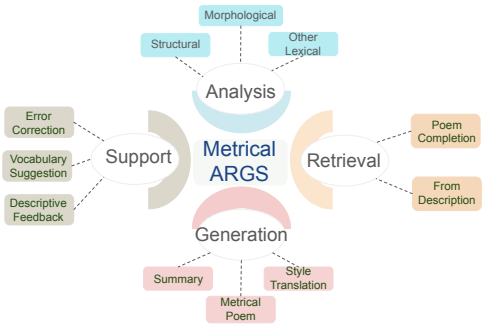

Figure 1: The MetricalARGS taxonomy of tasks for metrical poetry, spanning four dimensions: Analysis, Retrieval, Generation, and Support.

MetricalARGS (Figure 1). We then demonstrate how to use the taxonomy to study the capabilities of LLMs by considering a small test set constructed for Telugu, a Dravidian language with centuries of established metrical poetic tradition (Rao and Shulman, 2020).

From a methodological standpoint, studying metrical poetry with LLMs is important for two main reasons. First, it provides a rigorous testbed for understanding the capabilities of LLMs. Metrical verse requires models to coordinate multiple layers of linguistic competence, including phonology and prosody (to identify and count syllables correctly: a quantitative constraint), morphology and rhythmic structure (to follow metrical and phonological patterns: a structural constraint), and syntax and semantics (to maintain coherence, meaning, and thematic flow: a semantic constraint), all while preserving stylistic and aesthetic consistency. Second, metrical systems are defined by explicit, algorithmic

Proceedings of WILDRE-8, pages 1–13

rules that govern syllable patterns, rhyme positions, and line breaks. This makes them inherently computational in nature and opens up opportunities for developing modeling approaches and evaluation methods that can be integrated with mainstream NLP.

Beyond its methodological significance, studying metrical poetry with LLMs also carries important cultural and pedagogical value. Introduction to the basics of poetic meter happens in the high school level in the standard educational system of many Indian languages, including Telugu, which we use as our test language. Therefore, exploring the relevance of LLMs for metrical poetry also holds a strong pedagogical potential in supporting student learners as well as adult learners. It could also revitalize interest in a classical literary form of the language, assist with cross-linguistic studies of the poetic form, and support other digital humanities research.

With this motivation, we make the following contributions in this paper:

1. We create a taxonomy of tasks around the analysis and generation of metrical poetry and connect them to standard NLP tasks, outlining dataset and evaluation considerations for each task (Section 3).
2. Taking Telugu as the example language, we demonstrate how LLMs can be used for each of these tasks. Our case study (Section 5) serves as an illustrative probe to identify the potential and limitations of using LLMs for metrical poetry related tasks.

To our knowledge, this is the first paper to propose a unified taxonomy of tasks for metrical poetry in NLP, exploring how and where LLMs can support them, rather than focusing on one specific task. Further, this is also the first paper that assesses LLMs on Telugu metrical poetry.

2. Related Work

Most work related to the poetic form in NLP research has focused on poem generation (Ghazvininejad et al., 2016; Gonçalo Oliveira, 2017; Lau et al., 2018; Van de Cruys, 2020; Ormazabal et al., 2022) including recent research involving LLMs (Belouadi and Eger, 2023; Yu et al., 2024; Qu et al., 2025; Koziev and Fenogenova, 2025). In terms of the studied languages, while several world languages such as English (Chakrabarty et al., 2022; Walsh et al., 2024), Chinese (Pan et al., 2023; Ma et al., 2023), and Arabic (ElOraby et al., 2022; Alghallabi et al., 2025) are more widely studied, there is a smaller amount of research on languages such as Portuguese (Valença and Calegario, 2025) and Russian (Koziev and Fenogenova, 2025). Among the Indian languages, previous work focused only on Sanskrit meter (e.g., Jagadeeshan et al., 2026).

Tasks such as poetry analysis (Kao and Jurafsky, 2012; Kesarwani et al., 2017; Gopidi and Alam, 2019; Kurzynski et al., 2024; Sandhan et al., 2025; Jadhav et al., 2025) and translation into a given style/language (Genzel et al., 2010; Ghazvininejad et al., 2018; Chakrabarty et al., 2021; Wang et al., 2024) were also explored in the past, and there is a small amount of research on scansion and metrical analysis (Agirrezabal et al., 2017; Valença and Calegario, 2025; Agirrezabal et al., 2016). There is also some interest in exploring the pedagogical relevance of NLP based tools for poetry generation and analysis (Zhipeng et al., 2019; Rosa et al., 2025), and in the development of tools to scansion a poem i.e., identify the metrical patterns in a poem through rules (Terdalkar and Bhattacharya, 2023). To our knowledge, individual tasks are considered in isolation so far, disconnected from each other, due to a lack of a common taxonomy. Among the Indian languages, only Sanskrit has been studied to some extent across these tasks in the past NLP research (Sandhan et al., 2025; Jadhav et al., 2025; Terdalkar and Bhattacharya, 2023). Despite consistent academic interest in this topic within the NLP community, there has been no categorization of its specific sub-tasks. This paper addresses the issue of building a common taxonomy across related tasks for metrical poetry and introduces a new language, Telugu, into this line of research.

3. Metrical Poetry and NLP Tasks

Meter in poetry can be described as a controlled linguistic system that provides a rhythmic structure to the poems. Meters are typically characterized by rules governing the syllabic and/or sound patterns of the words in a poem, which control the eventual makeup of the poem. We will use syllable based meters as our use case for the rest of this paper. Although many languages of the world have established metrical traditions specific to that language, the typical process of creating a metrically compliant poem across languages consists of similar steps such as: choosing a metrical form, composing lines that fit the meter, abiding by its rhyme and pattern restrictions, and achieving some form of balance between form and meaning. Thus, it may be possible to construct a generalized taxonomy of tasks that can support comparisons across languages considering the common process as the starting point.

With that motivation, we identify a task taxonomy related to understanding and producing metrical poems (Figure 1) and relate these tasks to the standard tasks studied in NLP research (see Figure 2),

as that would enable us to define standard metrics for training and evaluation. We identify four broad categories of tasks: *Analysis, Retrieval, Generation and Support*, and describe them further, looking into potential means of dataset construction and evaluation. Since all of them require deeper understanding and reasoning about the linguistic structure, we do not include reasoning as a distinct task. We take the Telugu meter as a basis for this categorization, but we believe the taxonomy would be adaptable and extendable to other languages.

3.1. Analysis

Analysis tasks concentrate on taking an existing poem and studying its characteristics. We classify analysis into three sub-groups: structural analysis, morphological analysis and other analysis, which are explained below.

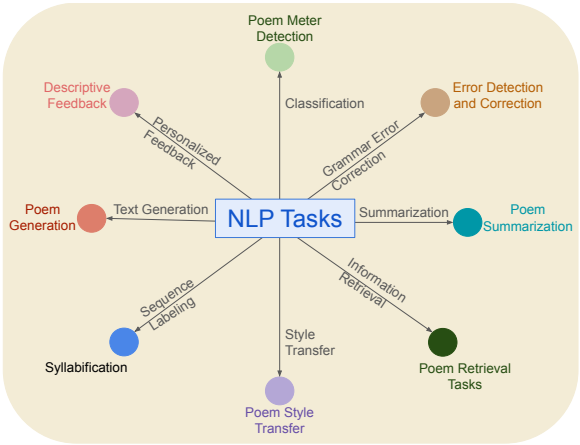

Figure 2: Mapping between METRICALARGS subtasks and established NLP tasks. Each poetry-specific task (outer nodes) aligns with a well-known NLP task (center), illustrating how METRICALARGS builds on existing NLP paradigms while extending them to metrical poetry.

Structural Analysis: This refers to the poem's adherence to a specific metrical form, and can be split into two broad tasks:

1. **Syllabification and Syllable Classification**: Given a line from a poem, or the full poem, the task is to identify the correct syllable pattern (or other phonetic length patterns in non-syllabic meters) in the word sequence, and group them together into the appropriate syllable sequences. For example, in Indian languages such as Sanskrit and Telugu, syllable groups are typically 1–4 syllables long, characterized by combinations of heavy (*guru*) and light (*laghu*) syllables, and this task would focus on arranging the given input into such identifiable groups.
2. **Mapping a syllable pattern to a meter**: Scansion is the process of identifying the metrical structure of a verse, which involves multiple sub-tasks such as counting the syllables, determining rhyme and other pattern based rules, and verifying them with the rules of the available meters to assign the given pattern to a meter.

The first task can be compared to a standard sequence labeling task such as part-of-speech tagging or named entity recognition in NLP, and the evaluation measures inherited from sequence labeling research can be adapted to this task. Metrical verse identification is a classification problem, as the number of known metrical patterns in a language is typically a fixed number[1]. Thus, both the tasks can be measured in terms of accuracy of some form i.e., the number of correctly identified patterns/meters. In terms of data collection, while there are no available datasets of this nature, many classical poems in Telugu and other Indian languages such as Sanskrit are published with an indicator that identifies their meter. Third party rule-based metrical analysis software already exist for some Indian languages such as Telugu[2], Sanskrit[3], which can be used to build a large scale dataset to support a larger evaluation or fine-tuning of the structural analysis abilities of LLMs, or for building verification tools for LLM generated analyses.

Morphological Analysis: This refers to tasks related to glossing of a given poem, which typically involves breaking up the individual words, adding their meanings in a more colloquial language, and tagging the relevant morphological information. In languages with word compounding (many Indian languages, for example), it is a non-trivial process to achieve the appropriate split, and since words can have multiple senses and meanings in context, the task of mapping a word to its right meaning also would involve some form of reasoning. There may also be additional information provided in such glosses, such as person/number/tense information etc. While there are no existing NLP datasets for this task, there are several publicly accessible texts for some languages (e.g., Telugu and Sanskrit) with gloss and plain text summaries for classical metrical poems, which can be utilized to build datasets for this task. There is an already existing body of work on glossing in NLP (Ginn et al., 2023), and evaluation measures from that research can be easily adapted to this task.

[1]For Telugu, a listing of 379 meters is at: https://chandamu.github.io/ChaMdOraajaM.html

[2]https://chandamu.github.io/

[3]https://sanskrit.iitk.ac.in/jnanasangraha/chanda/

Other Analysis: In digital humanities, as well as in NLP, it is not uncommon to see research studying sentiment, lexical/syntactic/stylistic patterns, authorship attribution and so on, both for prose and verse. Hence, it is natural to study these tasks even with metrical poetry, to understand the capabilities of LLMs in this area. However, while evaluation may be straightforward as such analysis may easily fit into a standard text classification task framework, compilation of relevant datasets for each of these tasks would require substantial human expertise. The role of LLMs in supporting humans in building high quality datasets for this kind of problems should also be explored in this context.

3.2. Retrieval

We refer to tasks related to identifying the right (existing) poems based on user queries as retrieval, which are listed below:

1. Retrieving the poem given its starting words or the first verse or words that appear in the middle or at the end
2. Retrieving the poem from its plain text description, including cross-lingual scenarios (Jagadeeshan et al., 2026).
3. Retrieving the poem(s) that matches in meaning, meter etc.

All of these tasks are similar to search and information retrieval tasks. Considering the large body of classical poetry based literature already available online, collecting datasets at least for the first two of these retrieval tasks should be relatively straightforward. Existing compilations of poem-summary pairs for some languages such as Telugu,[4] can be utilized as a starting point for creating a larger scale dataset, potentially with synthetically generated paraphrased versions of summaries, which can be useful for evaluation and fine-tuning purposes. In terms of evaluation, the standard retrieval based evaluation measures such as precision/recall/F-score can be used.

3.3. Generation

Generation tasks involve some form of textual generation based on a given description. We identify three generation tasks, described below:

1. **Poem summarization:** Poem to prose translation of a metrical poem, typically written in the classical literary form of the language into plain text. This can be viewed as similar to text summarization which is well-studied in NLP.
2. **Poem generation:** generating a novel poem given a textual description and a specified meter. Poem generation under such linguistic constraints was explored in the past in NLP research.
3. **Poem style transfer:** This is a challenging variation to generating a novel poem, where the input is a poem in one meter, and the output is the same content adapted to another meter. This task aligns with other existing research on style transfer in NLP, but in the context of poetry. This kind of metre style transfer is observed in the writings of classical and modern Telugu poets, and hence, can be seen as an advanced text generation task.

For poem to prose summarization, it is easier to create larger scale datasets by tapping into available resources, but for the remaining generative use cases which explore novel and creative content generation, one option is to look at synthetic data generated from LLMs followed by human evaluation. The effectiveness of generation can be evaluated through standard text generation metrics and metrical adherence can be checked through rule based checkers. But, human ratings are a must for other factors such fluency, coherence, adherence to the theme and style, creativity and aesthetics.

3.4. Support

Support tasks explore the role of LLMs in offering support to poets and students learning to write metrical poetry. We identify three main tasks as a starting point.

1. **Error detection and correction:** This refers to the process of identifying metrical, lexical or grammatical errors in the user written poem and offering corrections. This is most similar to the grammatical error detection/correction tasks and word-level translation quality estimation tasks, that have a long history in NLP research.
2. **Vocabulary suggestion:** This task, as the name indicates, offers vocabulary suggestions, but aligning with the metrical constraints of the context. There is perhaps no equivalent existing NLP task as these suggestions need both semantic and metrical compliance.
3. **Descriptive Feedback:** This refers to giving explanation to the user on the text they created and offering suggestions for rewriting. The more recently introduced Grammatical Error Explanation task (Song et al., 2024) is potentially the closest existing NLP task.

All the three support tasks, while being specific to metrical poetry generation, also share some com-

[4] https://huggingface.co/datasets/SuryaKrishna02/aya-telugu-poems

Paper	Lang	Task	MetricalARGS
Kao and Jurafsky (2012)	en	PA	Analysis
Walsh et al. (2024)*	en	PF	Analysis
Valença and Calegario (2025)*	pt	PS	Analysis
Pan et al. (2023)	zh	PA	Analysis
Kurzynski et al. (2024)	zh	PP	Analysis
Jagadeeshan et al. (2026)	sa	PG	Generation
Ghazvininejad et al. (2016)	en	PG	Generation
Belouadi and Eger (2023)*	en	PG	Generation
ElOraby et al. (2022)	ar	PG	Generation
Koziev and Fenogenova (2025)*	ru	PG	Generation
Genzel et al. (2010)	en	ST	Generation
Wang et al. (2024)*	en	ST	Generation

Table 1: Mapping existing metrical poetry works with the proposed MetricalARGS tasks. Lang: Languages supported; PA: Poetry Analysis; PF: Poetic Form; PS: Poetic Scansion; PP: Poem Parllelism; PG: Poem Generation; ST: Style Transfer; * - Indicates the works used LLMs.

monalities with other relevant research on educational applications of NLP, and data creation and evaluation approaches for the related topics can be adapted for these use cases as well. However, it is important to note that they are inherently more challenging tasks than the other ARGS tasks discussed in this paper so far.

From this discussion, it is clear that most of the MetricalARGS tasks can benefit from existing NLP research in related tasks, while introducing challenging new variations. Overall, the MetricalARGS taxonomy demonstrates that there are a wide range of complex tasks related to metrical poetry generation, where LLMs may be relevant and can be further studied. While the taxonomy is created with Telugu meter as its basis (owing to the authors' familiarity with it), these tasks are not specific to Telugu and can be studied for other Indian languages as the poetic traditions share some commonalities. We expect this task taxonomy to be improved as needed as NLP researchers adapt the taxonomy to poetic traditions in other languages and language families in future.

Figure 2 summarizes the different MetricalARGS tasks and their relation to other standard NLP tasks. Table 1 shows how some of the existing research maps into this taxonomy. Most of the past work appears to have focused on Analysis and Generation tasks, often addressing only a subset of sub-tasks within each category, while Retrieval and Support remain largely unexplored.

4. Applying MetricalARGS for Telugu

We demonstrate the use of MetricalARGS taking Telugu as the test language in this section. Telugu is recognized as one of the classical languages of India (Press Information Bureau, 2024) and has a centuries old literary tradition. Earliest known description of Telugu poetic meters and rules of prosody are from a 6th or 7th century text (Ramakrishna et al., 1983, pp.164–165). Telugu poetic meter, while sharing a lot of patterns with Sanskrit poetic tradition has several other native metrical patterns as well. Considering the agglutinative nature of the language, tasks such as breaking up of the syllable sequences into individual words for glossing and summarizing the meaning too offer a range of language processing and reasoning related challenges, along with other tasks around metrical poetry analysis and generation.

To illustrate the use of MetricalARGS taxonomy for Telugu metrical poetry, we curated a dataset of approximately 20 samples for each task (169 samples in total). The intention of using such a small dataset is not to establish a benchmark, but to showcase how current models handle the ARGS tasks and illustrate how to build benchmark datasets using this taxonomy across languages in future. Although modest in size, the dataset covers representative examples of each task. These samples were collected from the official Grade 7–10 Telugu textbooks published by the Andhra Pradesh state government in India. The dataset was prepared and annotated by two native Telugu speakers. Collection of a larger scale dataset covering all the described tasks in the taxonomy is a larger effort beyond the scope of this paper and we hope this pilot study will lead into such benchmarking efforts in future across many languages including Telugu.

Evaluation We considered two proprietary LLMs for output generation: GPT-5 and Gemini-2.5-Pro. For the analysis tasks (syllabification, syllable classification, morphological segmentation, and meter validation) as well as one generation sub-task (summarization), gold references were available. For the remaining categories, such as retrieval, poem generation, and style transfer, multiple valid outputs are possible. In both the cases, we adopted an LLM-as-a-judge approach for evaluation and used Gemini-2.5-Pro as the judge model, supplying it with a gold output for comparison where it is available. The percentage of responses the judge model scores as correct is considered as the measure of performance (with a scale of 0–1, the more the better). All experiments were conducted using the Inspect evaluation framework[5] in a zero-shot setting, using Telugu prompts, with temperature set to zero. For the subset of cases where there is no single gold standard output, we also conducted a human evaluation in which authors, who are native Telugu speakers reviewed a sample of outputs to verify both the correctness of model predictions and the validity of the LLM-based evaluations [6].

[5] https://inspect.aisi.org.uk/

[6] The dataset and the human evaluation data are both accessible at: https://huggingface.co/datasets/TeluguLLMResearch/MetricalARGS

Category	SubCategory	# Q	Accuracy	
			GPT	Gemini
Analysis	SC	20	0.60	0.20
	MA	20	0.20	0.65
	MD	20	0.40	0.50
Retrieval	FV	6	0.00	0.00
	MRV	6	0.00	0.00
	LV	6	0.00	0.00
Generation	PS	20	0.70	**0.85**

Table 2: Accuracy across different MetricalARGS tasks with gold-output, using LLM-as-a-Judge. SC: Syllabification and Syllable Classification, MA: Morphological Analysis, MD: Meter Detection, FV: Retrieval from First Verse, MRV: Retrieval from Middle/Random Verse, LV: Retrieval from Last Verse, PS: Poem Summarization. #Q indicates the number of questions per task. GPT: GPT-5, Gemini: Gemini-2.5-Pro.

Category	SubCategory	# Q	Accuracy	
			GPT	Gemini
Retrieval	MM	2	0.50	0.00
Generation	PFS	7	0.71	0.29
	RPFW	8	**1.00**	0.75
	PFP	5	0.40	0.20
	ST	20	0.00	0.00
Support	EDC	10	0.10	0.00
	VS	19	0.47	0.26

Table 3: Accuracy for tasks without gold outputs, evaluated using an LLM-as-a-judge. MM: Retrieval using meaning, PFS: Poem from Summary, RPFW: Riddle Poem from Word, PFP: Poem from a Problem (Samasya in Avadhanam), ST: Style Transfer, EDC: Error Detection and Correction, VS: Vocabulary Suggestion. #Q indicates the number of questions per task. GPT: GPT-5, Gemini: Gemini-2.5-Pro.

5. Results

We separated the evaluation into two groups depending on whether a gold standard answer was available or not in the dataset. Table 2 presents the results for the tasks where gold references are available. Both the models performed well for the poem summarization task. While the GPT-5 model did better with syllabification, Gemini model did better with morphological analysis. Meter detection was more challenging for both models. The models often misclassified syllable length, leading to downstream errors in meter identification. Metrical rules, such as identifying short and long syllables, are strict and rule-based. LLMs, as next-token predictors, potentially miss fine-grained distinctions that require precise symbolic reasoning, resulting in such errors.

In morphological analysis, both the models generally captured meanings with multiple words instead of single-word glosses, resulting in mismatches with gold annotations, indicating the need for better evaluation measures for that task. Surprisingly, retrieval proved to be a challenging task. Neither model was able to successfully retrieve the complete poem given only the first, last, or a random line as input, resulting in an accuracy of 0 for this task. GPT-5 avoided retrieval altogether by producing follow-up questions (see Figure 3), while Gemini generated paraphrased versions of the poems instead of the actual poems. This could be due to potential coverage issues in the training data, or the models trying to avoid verbatim reproduction of text.

Table 3 reports the results for the tasks where the LLM judge evaluated the generated outputs without a gold answer. GPT-5 appears to be better than Gemini for these tasks, despite the judge being Gemini. Both models achieve near-zero scores for style transfer and error detection and correction. For the generation task in which the models were asked to compose a riddle-style poem, GPT-5 achieved a score of 1.00, while Gemini-2.5-Pro achieved 0.75. In the task of generating a poem from a summary, GPT-5 obtained a score of 0.71 compared to 0.29 for Gemini-2.5-Pro. For style transfer, both models received a score of 0.00. Overall, GPT-5 outputs received higher scores than Gemini-2.5-Pro outputs, even though Gemini-2.5-Pro was used as the judge in both cases. These results suggest that the models are able to produce acceptable outputs for some creative generation tasks, while style transfer remained unsuccessful for both.

Overall, these results provide preliminary insight into the capabilities of LLMs in different MetricalARGS tasks for Telugu, highlighting areas of relative strengths (semantic-level tasks) and weaknesses (retrieval and structural-based tasks). Note that this evaluation covered all the tasks listed in the taxonomy, with two exceptions: "Other Analysis" subset of tasks require expert data curation, and "Descriptive Feedback" is assessed through its components (error correction and vocabulary suggestion). Hence, they are excluded from this case study.

5.1. Qualitative Analysis

To better understand the results, we examine two representative tasks focusing on retrieval and generation respectively. Figure 3 illustrates a retrieval question task where the query provides a first verse and the model is asked to retrieve the complete poem, while the gold standard answer has the expected poem. GPT-5 responded with a follow-up question rather than attempting a retrieval; this behavior was consistent across all samples in this

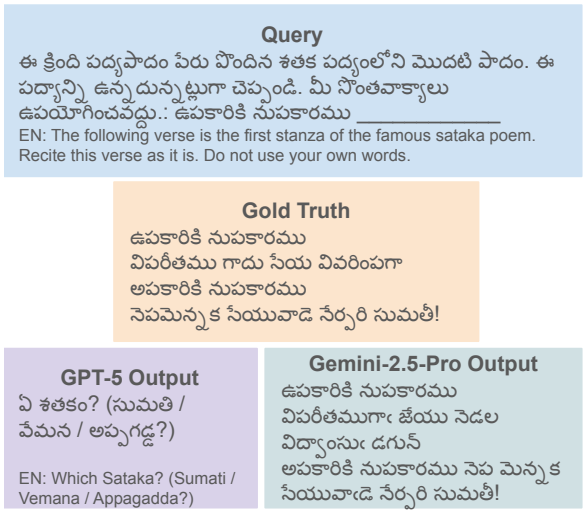

Figure 3: Example from the Retrieval QA task. GPT-5 responded with a clarification question instead of retrieval, while Gemini-2.5-Pro produced text with partial overlap, but did not match the gold reference.

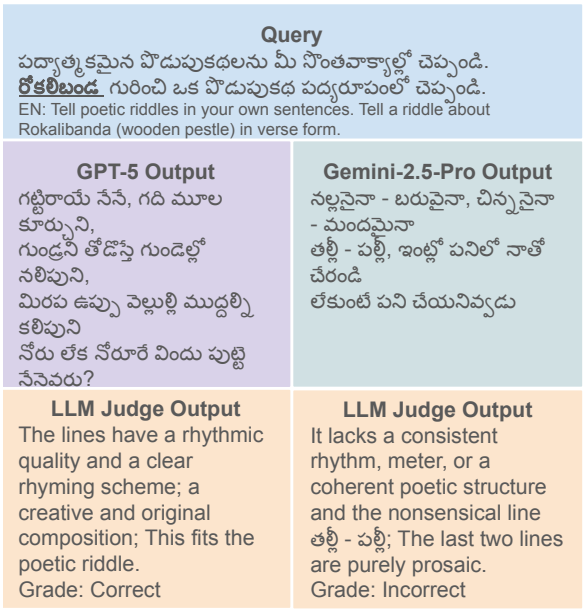

Figure 4: Example from the Riddle-style Poem Generation task. The query asks the model to compose a riddle in verse form about *Rokalibanda* (a wooden pestle).

task. Gemini-2.5-Pro produced outputs with some lexical overlap with the gold poem but did not match the target verse, and therefore did not satisfy the exact-retrieval criterion.

Our analysis revealed consistent pattern in how the Gemini-2.5-Pro model used the input verse. When given the first or last verse of the poem, the model retained the input at the appropriate position and attempted to generate surrounding verses. However, the same behavior extended even to the random (middle) verse case, where the input fragment was incorrectly treated as the opening line, and the model generated a continuation around it. This suggests that rather than recalling the original poem, the model interprets the task as one of conditional generation. This task probes the model's ability to match patterns and recall memorized sequences. Given that LLMs generate text through next-token prediction, one might expect them to reproduce a familiar verse when prompted with part of it. However, in our experiments, neither model succeeded, indicating unreliable recall of longer structured texts. Possible reasons include safety guardrails that prevent verbatim reproduction or a limited Telugu training coverage, warranting further tests in other languages.

Figure 4 presents the model outputs and the LLM-judge's reasoning criteria for a poem generation question. In this example, the model is asked to generate a poem in the form of a riddle for a given word. To succeed, the model must first understand the meaning of the word and identify its characteristic features, and then compose a poem that conveys those features indirectly, without explicitly naming the word. This combination of semantic understanding, abstraction, and structured poetic composition makes the task challenging (even for humans). While the LLM-as-a-judge marked the GPT-5 output as correct, human inspection revealed a mismatch: the verse was rhythmically consistent but contained lexical errors and ungrammatical phrasing in Telugu. This highlights both the complexity of the task and a discrepancy between LLM-based and human judgments of fine-grained linguistic accuracy. Taken together, these examples illustrate that while the LLMs can produce outputs that appear fluent or partially aligned with expectations, they often fail to meet stricter criteria such as exact retrieval or grammatical accuracy, underscoring the need for human verification in addition to LLM-based evaluation.

In the Retrieval–Matching Meaning task, lower human scores for GPT-5 stem from its failure to retrieve an existing, popular poem with similar meaning. The model reproduced the input poem verbatim with only the closing verse (last line) changed (see Figure 5), making it a hallucination. The LLM judge, however, marked this as correct, increasing the score relative to human assessments.

In the Generation–Poem from Summarization task, differences between judge and human scores arise from contrasting emphasis on meaning versus form. GPT-5 outputs generally preserve semantic content but often lack poetic coherence or structure, resembling prose or using contextually misplaced words (see Figure 6). The judge, focusing on semantic similarity, overlooks these stylistic deficiencies and assigns higher scores.

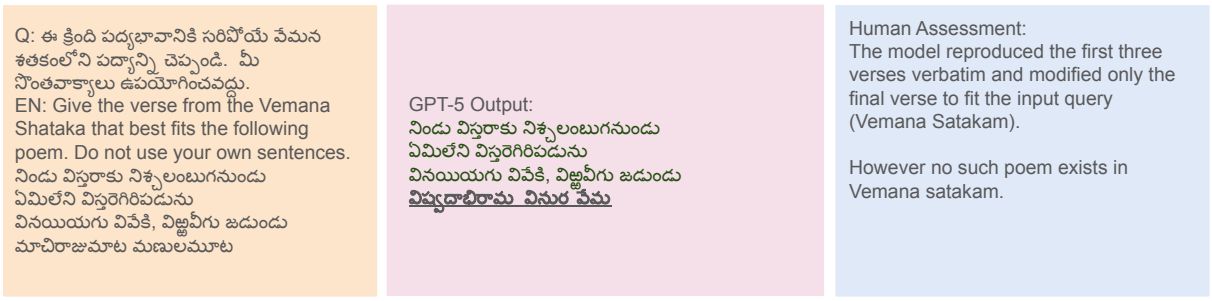

Figure 5: Example of GPT-5 output for the Retrieval–Matching Meaning task, showing the model's verbatim reproduction of input verses with minor modification.

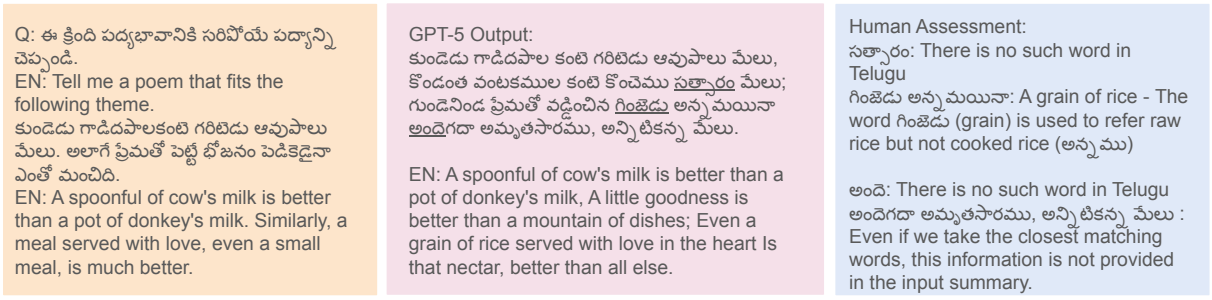

Figure 6: Example of GPT-5 output for the Generation–Poem from Summarization task, showing semantic alignment but lexical and contextual inaccuracies in Telugu usage.

A similar pattern appears in the Generation–Poem (riddle) task, where GPT-5 responses frequently lack logical or poetic structure (see Figure 7) but are still rated correct by the judge. This behavior likely results from the judge evaluating outputs primarily through translation and meaning comparison rather than structural or creative alignment. Consequently, outputs that align in surface meaning but fail in form are treated as correct, while human evaluators apply stricter criteria. In style transfer tasks, model outputs often achieve partial alignment with the target style, sometimes exceeding 70% similarity (see Figure 8).

Overall, these findings suggest that the LLM-as-a-judge primarily emphasizes semantic similarity while neglecting stylistic, structural, and contextual aspects that human evaluators recognize. This tendency results in higher scores for outputs that align in meaning but lack linguistic or creative quality, highlighting the need for evaluation frameworks that account for contextual and stylistic depth.

5.2. Human Evaluation of Model Outputs and Judge Scores

This observation lead us to conduct a human evaluation of LLMs as both generators and judges for these tasks. The authors, both native Telugu speakers, independently evaluated the outputs of both models for the tasks without gold output (i.e., the Tasks in Table 3, which cover about half of the dataset) and marked whether each output was correct. These computed scores are then compared with the scores assigned by the LLM-judge in order to assess both the correctness of the model outputs and the reliability of the LLM-as-a-judge framework. Table 4 shows a summary of this comparison.

Category	Sub-Category	JS	A1	A2
GPT-5				
Retrieval	MM	0.50	0.00	0.0
Generation	PFS	0.71	0.29	0.43
	RPFW	**1.00**	**0.00**	**0.00**
	PFP	0.40	0.20	0.20
	ST	0.00	0.00	0.00
Support	EDC	0.10	0.00	0.00
	VS	0.47	0.26	0.26
Gemini-2.5-Pro				
Retrieval	MM	0.00	0.00	0.0
Generation	PFS	0.29	0.14	0.14
	RPFW	**0.75**	**0.63**	**0.50**
	PFP	0.20	0.20	0.20
	ST	0.00	0.00	0.00
Support	EDC	0.00	0.00	0.00
	VS	0.26	0.11	0.16

Table 4: A Comparison of Human and LLM judge evaluations. MM: Retrieval using meaning, PFS: Poem from Summary, RPFW: Riddle Poem from Word, PFP: Poem from a Problem (Samasya in Avadhanam), ST: Style Transfer, EDC: Error Detection and Correction, VS: Vocabulary Suggestion.

Overall, the LLM-as-a-judge (JS) reports higher scores than human annotators (A1 and A2) across

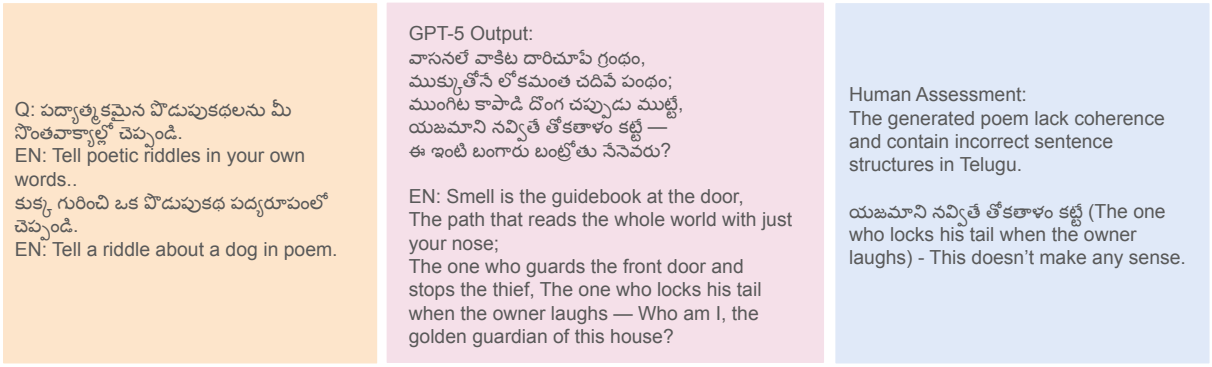

Figure 7: Example of GPT-5 output for the Generation–Poem (Riddle) task, illustrating incoherent verse construction and incorrect sentence structures in Telugu.

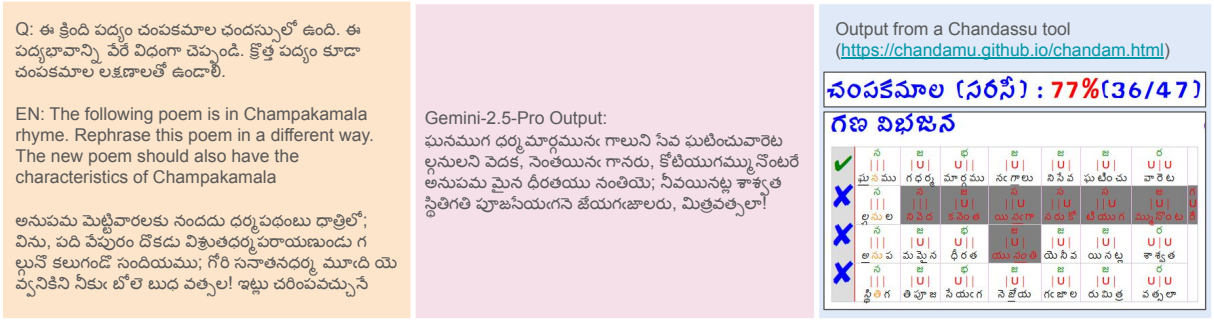

Figure 8: Example of Gemini-2.5-Pro output for the Generation–Style Transfer task, showing partial adherence to the target Champakamala rhyme scheme with approximately 77% metrical accuracy as verified using the Chandam tool.

most sub-categories (see Table 4), especially for GPT-5. In the Retrieval–MM task, for example, GPT-5 reproduced the input poem from Vemana Satakam with only the closing line altered, rather than retrieving a semantically similar poem (see Figures 5). The judge incorrectly labeled this as correct, while both the human evaluators marked it as incorrect. In the Generation–Poem from Summarization and Poem (riddle) tasks, GPT-5 outputs often preserved meaning but lacked poetic structure, coherence, or riddle form (see Figure 6 and 7). The judge, focusing mainly on semantic similarity, labeled it as correct, unlike the human annotators.

In style transfer tasks, outputs achieved partial alignment (upto 70% similarity for some questions) with the target style (see Figure 8), and occasional deviations, such as GPT-5 producing Sanskrit or Gemini-2.5-Pro generating multiple responses, showing inconsistencies in task adherence. The judge also penalized a correct output for not reproducing a well-known poem, despite the prompt not specifying such a requirement. It is important to note that there is a high level of agreement between human judges in terms of scoring the model outputs (Cohen's Kappa in the range of 0.93 for both models), but less agreement with the LLM judges. Specifically, the agreement between humans and LLM Judge was much lesser for GPT (0.29-0.33) compared to Gemini (0.74 in both cases). We notice that LLM-as-a-judge relies heavily on meaning-based evaluation, underestimating stylistic, structural, and contextual alignment, highlighting the need for richer, context-aware evaluation frameworks for this kind of tasks. More experiments with judge prompts, and potentially few-shot judge prompts and fine-tuned judge models may be needed for a reliable automated evaluation, which can in turn support large scale dataset creation and evaluation for all the MetricalARGS tasks in Telugu as well as other languages.

6. Road Ahead

In this paper, we created a taxonomy of tasks MetricalARGS for metrical poetry analysis with LLMs, and demonstrated how this can be applied to one language, Telugu, by exploring the capabilities of today's LLMs across these tasks. The case study demonstrates both the potential and challenges with using LLMs as generators and judges for these tasks. Task-specific in-context learning and fine-tuning may hold the potential for improvements, though. Creation of high quality evaluation datasets and exploring mechanisms to create larger scale

fine-tuning datasets across different tasks is an obvious next step to pursue in this direction. Building a larger dataset covering all tasks to support a more comprehensive evaluation is an obvious next step in this direction. Evaluating (and extending) the taxonomy for other Indian languages, and potentially other world languages with similar poetic traditions should be considered in future extensions of this line of research. Consider the ongoing interest in Sanskrit poetry in the NLP community (Jagadeeshan et al., 2026; Terdalkar and Bhattacharya, 2023; Sandhan et al., 2025), establishing a MetricalARGS benchmark for Sanskrit would also be a worthwhile direction to pursue.

Overall, we propose that positioning metrical poetry as a testbed opens up new ways to assess and enhance LLM understanding of form-constrained language, while also supporting application areas such as learning tool development and digital humanities research. We hope that this paper serves as a starting point for further research on exploring the relation between LLMs and metrical poetry, and investigating the role of LLMs in understanding other structured linguistic systems like meter across the world languages.

Limitations

We identify two important limitations to this work:

1. The proposed taxonomy used Telugu metrical poetry tradition as the basis, and the coverage of tasks may not be comprehensive enough to cover the poetic traditions across many languages. Additionally, it is possible to imagine tasks such as metrical poem translation between languages (*e.g., translate a given poem in Telugu meter T-A into Chinese meter C-X*), generating or retrieving a poem given an image (*e.g., generate a poem in Telugu meter T-A, or retrieve a poem in Chinese meter C-X, based on the input image*), which this taxonomy does not currently cover. This can be perceived as a limitation, but we intend this paper to be a starting point to raise further discussions on the topic, and hence, we would expect to see improvements and additions to this taxonomy of tasks in near future.

2. Our empirical study was done based on zero-shot prompting using a small dataset of 169 samples (which do cover most of the tasks in the proposed taxonomy), and two state of the art models, and hence, cannot be considered a large scale, comprehensive evaluation of the abilities of LLMs on MetricalARGS tasks. However, we view this evaluation as an essential starting point to identify requirements for a larger study. For example, our human evaluation would not have been feasible with a large scale dataset, and it did help us identify the short comings of using an LLM-as-a-judge approach for evaluation for some of the sub-tasks. We hope to conduct a larger scale evaluation in future, by using this smaller study as the basis.

   We view these not as limitations to the research topic itself, and hope that this paper laid the necessary foundation to address the above limitations in future.

Acknowledgments

We thank Isar Nejadgholi, Krishnapriya Vishnubhotla and Gabriel Bernier-Colborne for their feedback. Dileep Miriyala's Chandam (https://chandamu.github.io/) inspired us to study metrical poetry with LLMs, and we thank him for creating the tool.

7. Bibliographical References

Manex Agirrezabal, Iñaki Alegria, and Mans Hulden. 2016. Machine learning for metrical analysis of English poetry. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 772–781, Osaka, Japan. The COLING 2016 Organizing Committee.

Manex Agirrezabal, Iñaki Alegria, and Mans Hulden. 2017. A comparison of feature-based and neural scansion of poetry. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 18–23, Varna, Bulgaria. INCOMA Ltd.

Wafa Alghallabi, Ritesh Thawkar, Sara Ghaboura, Ketan More, Omkar Thawakar, Hisham Cholakkal, Salman Khan, and Rao Muhammad Anwer. 2025. Fann or flop: A multigenre, multiera benchmark for arabic poetry understanding in llms. *arXiv preprint arXiv:2505.18152*.

Jonas Belouadi and Steffen Eger. 2023. ByGPT5: End-to-end style-conditioned poetry generation with token-free language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7364–7381, Toronto, Canada. Association for Computational Linguistics.

Tuhin Chakrabarty, Vishakh Padmakumar, and He He. 2022. *Help me write a poem*: Instruction tuning as a vehicle for collaborative poetry

writing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6848–6863, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tuhin Chakrabarty, Arkadiy Saakyan, and Smaranda Muresan. 2021. Don't go far off: An empirical study on neural poetry translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7253–7265, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Maryam ElOraby, Mohamed Abdelgaber, Nehal Elkaref, and Mervat Abu-Elkheir. 2022. Generating classical Arabic poetry using pre-trained models. In *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 53–62, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Dmitriy Genzel, Jakob Uszkoreit, and Franz Och. 2010. "poetic" statistical machine translation: Rhyme and meter. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 158–166, Cambridge, MA. Association for Computational Linguistics.

Marjan Ghazvininejad, Yejin Choi, and Kevin Knight. 2018. Neural poetry translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 67–71, New Orleans, Louisiana. Association for Computational Linguistics.

Marjan Ghazvininejad, Xing Shi, Yejin Choi, and Kevin Knight. 2016. Generating topical poetry. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1183–1191, Austin, Texas. Association for Computational Linguistics.

Michael Ginn, Sarah Moeller, Alexis Palmer, Anna Stacey, Garrett Nicolai, Mans Hulden, and Miikka Silfverberg. 2023. Findings of the sigmorphon 2023 shared task on interlinear glossing. In *Proceedings of the 20th SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 186–201.

Hugo Gonçalo Oliveira. 2017. A survey on intelligent poetry generation: Languages, features, techniques, reutilisation and evaluation. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 11–20, Santiago de Compostela, Spain. Association for Computational Linguistics.

Amitha Gopidi and Aniket Alam. 2019. Computational analysis of the historical changes in poetry and prose. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 14–22, Florence, Italy. Association for Computational Linguistics.

Bhakti Jadhav, Himanshu Dutta, Shruti Kanitkar, Malhar Kulkarni, and Pushpak Bhattacharyya. 2025. An introduction to computational identification and classification of upamā alaṅkāra. In *Computational Sanskrit and Digital Humanities - World Sanskrit Conference 2025*, pages 1–14, Kathmandu, Nepal. Association for Computational Lingustics.

Manoj Balaji Jagadeeshan, Samarth Bhatia, Pretam Ray, Harshul Raj Surana, Akhil Rajeev P, Priya Mishra, Annarao Kulkarni, Ganesh Ramakrishnan, Prathosh Ap, and Pawan Goyal. 2026. Chandomitra: Towards generating structured Sanskrit poetry from natural language inputs. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 518–534, Rabat, Morocco. Association for Computational Linguistics.

Justine Kao and Dan Jurafsky. 2012. A computational analysis of style, affect, and imagery in contemporary poetry. In *Proceedings of the NAACL-HLT 2012 Workshop on Computational Linguistics for Literature*, pages 8–17, Montréal, Canada. Association for Computational Linguistics.

Vaibhav Kesarwani, Diana Inkpen, Stan Szpakowicz, and Chris Tanasescu. 2017. Metaphor detection in a poetry corpus. In *Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 1–9, Vancouver, Canada. Association for Computational Linguistics.

Ilya Koziev and Alena Fenogenova. 2025. Generation of Russian poetry of different genres and styles using neural networks with character-level tokenization. In *Proceedings of the 9th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2025)*, pages 47–63, Albuquerque, New Mexico. Association for Computational Linguistics.

Maciej Kurzynski, Xiaotong Xu, and Yu Feng. 2024. Vector poetics: Parallel couplet detection in classical Chinese poetry. In *Proceedings of the 4th International Conference on Natural Language*

Processing for Digital Humanities, pages 200–208, Miami, USA. Association for Computational Linguistics.

Jey Han Lau, Trevor Cohn, Timothy Baldwin, Julian Brooke, and Adam Hammond. 2018. Deepspeare: A joint neural model of poetic language, meter and rhyme. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1948–1958, Melbourne, Australia. Association for Computational Linguistics.

Jingkun Ma, Runzhe Zhan, and Derek F. Wong. 2023. Yu sheng: Human-in-loop classical Chinese poetry generation system. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 57–66, Dubrovnik, Croatia. Association for Computational Linguistics.

Aitor Ormazabal, Mikel Artetxe, Manex Agirrezabal, Aitor Soroa, and Eneko Agirre. 2022. PoeLM: A meter- and rhyme-controllable language model for unsupervised poetry generation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3655–3670, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Changzai Pan, Feiyue Li, and Ke Deng. 2023. TopWORDS-poetry: Simultaneous text segmentation and word discovery for classical Chinese poetry via Bayesian inference. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3372–3386, Singapore. Association for Computational Linguistics.

Government of India Press Information Bureau. 2024. Cabinet approves conferring status of classical language to marathi, pali, prakrit, assamese and bengali languages. Accessed: 2025-10-08.

Zhan Qu, Shuzhou Yuan, and Michael Färber. 2025. Poetone: A framework for constrained generation of structured chinese songci with llms.

Gamapalahalli Ramakrishna, Nagarajarao Gayathri, and Debiprasad Chattopadhyaya. 1983. *An encyclopaedia of South Indian culture*. South Asia Books.

Velcheru Narayana Rao and David Shulman. 2020. *Classical Telugu Poetry*, volume 13. University of California Press.

Rudolf Rosa, David Mareček, Tomáš Musil, Michal Chudoba, and Jakub Landsperský. 2025. EduPo: Progress and challenges of automated analysis and generation of Czech poetry. In *Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities*, pages 524–542, Albuquerque, USA. Association for Computational Linguistics.

Jivnesh Sandhan, Amruta Barbadikar, Malay Maity, Pavankumar Satuluri, Tushar Sandhan, Ravi M Gupta, Pawan Goyal, and Laxmidhar Behera. 2025. Aesthetics of Sanskrit poetry from the perspective of computational linguistics: A case study analysis on śikṣāṣṭaka. In *Computational Sanskrit and Digital Humanities - World Sanskrit Conference 2025*, pages 15–36, Kathmandu, Nepal. Association for Computational Lingustics.

Yixiao Song, Kalpesh Krishna, Rajesh Bhatt, Kevin Gimpel, and Mohit Iyyer. 2024. GEE! grammar error explanation with large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 754–781, Mexico City, Mexico. Association for Computational Linguistics.

Hrishikesh Terdalkar and Arnab Bhattacharya. 2023. Chandojnanam: A sanskrit meter identification and utilization system. In *Proceedings of the Computational Sanskrit & Digital Humanities: Selected papers presented at the 18th World Sanskrit Conference*, pages 113–127.

André Valença and Filipe Calegario. 2025. Experimenting with large language models for poetic scansion in portuguese: A case study on metric and rhythmic structuring. In *Proceedings of the 16th International Conference on Computational Creativity*.

Tim Van de Cruys. 2020. Automatic poetry generation from prosaic text. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2471–2480, Online. Association for Computational Linguistics.

Melanie Walsh, Anna Preus, and Maria Antoniak. 2024. Sonnet or not, bot? poetry evaluation for large models and datasets. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15568–15603, Miami, Florida, USA. Association for Computational Linguistics.

Shanshan Wang, Derek Wong, Jingming Yao, and Lidia Chao. 2024. What is the best way for ChatGPT to translate poetry? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14025–14043, Bangkok, Thailand. Association for Computational Linguistics.

Chengyue Yu, Lei Zang, Jiaotuan Wang, Chenyi Zhuang, and Jinjie Gu. 2024. CharPoet: A Chinese classical poetry generation system based

on token-free LLM. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 315–325, Bangkok, Thailand. Association for Computational Linguistics.

Guo Zhipeng, Xiaoyuan Yi, Maosong Sun, Wenhao Li, Cheng Yang, Jiannan Liang, Huimin Chen, Yuhui Zhang, and Ruoyu Li. 2019. Jiuge: A human-machine collaborative Chinese classical poetry generation system. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 25–30, Florence, Italy. Association for Computational Linguistics.

Semi-automatic Approach for Tamil Discourse Relation Annotation

Frances Yung, Enosh Peter Ponraj, and Vera Demberg
Saarland University, Saarland Informatics Campus, Germany
{frances,vera}@coli.uni-saarland.de
enpe00001@stud.uni-saarland.de

Abstract

Discourse relations (DRs) specify the logical relations between text spans and are essential for modeling extended discourse. Resources annotated with DRs can help train large language models (LLMs) to recognize and generate these relations more naturally. However, there is currently no open-source DR-annotated resource for Tamil. Annotation is particularly challenging because many Tamil discourse connectives are realized as morphologically complex suffixes rather than standalone tokens, often involving phonological alternations. In this work, we present a DR-annotated dataset for Tamil based on the PDTB framework. We adopt a semi-automatic pipeline: 1) projection of automatic English discourse annotations onto Tamil in a parallel corpus; 2) lexical normalization using a morphological analyzer; and 3) manual verification of each instance. The resulting resource contains approximately $7{,}200$ explicit DR annotations and a lexicon of 450 Tamil discourse connectives. The annotated data is available for download at https://github.com/Enosh-P/Tamil-Semi-Automatic-Discourse-Relation-Dataset/

Keywords: Tamil, discourse relations, discourse annotation, PDTB, language resources

1. Introduction

Discourse refers to a coherent and structured set of sentences, such as those found in conversations and written or spoken texts, expressed in natural language. For a discourse to be meaningful and coherent, its segments should be connected through logical relations, such as cause-effect, elaboration, contrast and exemplification. These connections are referred to as coherence relations or discourse relations (DRs; Hobbs, 1978). DRs provide the structural backbone of discourse, allowing readers to interpret the overall communicative goal by connecting individual units of texts (Sanders and Spooren, 2011). Incorporating knowledge of DRs has been shown to enhance the reasoning capabilities of large language models (LLMs), improving their performance in tasks such as text generation (Guan et al., 2021; Liu et al., 2026), summarization (Xu et al., 2020; Liu and Demberg, 2024), sentiment analysis (Choi et al., 2016) and question-answering (Verberne et al., 2007; Sovrano et al., 2025).

The largest DR-annotated resources are in English, including the Penn Discourse Treebank (PDTB; Prasad et al., 2008; Webber et al., 2019) and the RST-Discourse Treebank (Carlson et al., 2003). Beyond English, discourse resources have been developed for a growing number of languages (Stede, 2004; da Cunha et al., 2011; Zhou and Xue, 2012; Synková et al., 2024). For Indian languages, notable efforts include the Hindi Discourse Relation Bank (Oza et al., 2009), the Bangla RST Discourse Treebank (Das and Stede, 2018), and other resources for Hindi, Malayalam, and Tamil created for corpus studies or model training (Rachakonda and Sharma, 2011; Gopalan et al., 2017; Sheeja S and Lalitha Devi, 2022). However, none of Indian language discourse resources are publicly accessible.

In this work, we present a DR-annotated resource for Tamil, developed following the PDTB framework. In this framework, DR annotations are anchored to explicit discourse signals, known as discourse connectives (DCs), which lexically signal the underlying relations. Applying this annotation scheme to Tamil is particularly challenging because, unlike many other languages where DCs are typically isolated tokens, Tamil often realizes DCs as morphological suffixes attached to other content words, as shown in the example below.

Example (1)

Tamil: மழை பெய்ததால் **அவன் வீட்டில் இருந்தான்**
malzhai peithatha-aal **avan veetil irunthaan**
gloss: *rain fall-PST-CAUS* 3SG.M house-LOC be-PST-3SG.M

English translation: Because **it rained**, *he stayed home.*

DR sense: CONTINGENCY.CAUSAL.REASON

Here, the DC *"because"* in English marks the DR REASON between the clauses *"he stayed home"* (called the *Argument1*) and *"it rained"* (called the *Argument2*. On the other hand, in Tamil, instead of an isolated word, the relation is marked by the suffix *"-aal"*, which is part of the word *"peithathaaal"*. Nonetheless, not all DCs in Tamil are suffixes. There are also single-token DCs, such as ஆனால் (*aana*, but) and அதேபோல் (*athepol*, similarly). While suffixal DCs also occur in the

Proceedings of WILDRE-8, pages 14–24
12 May 2026.

Turkish language and are included in the lexicon of Turkish DCs (Zeyrek and Başıbüyük, 2019), they are were not annotated in the Turkish Discourse Bank (Zeyrek et al., 2010; Zeyrek and Kurfalı, 2017). To our knowledge, the current resource is the first effort to annotate DCs that are not individual tokens.

To create a DR-annotated resource for Tamil, we propose using **annotation projection**, in which automatic DR annotations are projected from English to Tamil via a parallel corpus. DC candidates together with their arguments are first extracted from the Tamil texts based on their alignment with the English words. To identify connective suffixes, a **morphological analyzer** is applied to the extracted Tamil DC candidates. Finally, an extended series of **manual verification and normalization steps** is carried out to ensure the quality and consistency of the annotations.

Our annotation pipeline results in a corpus of 7200 explicit DR annotation, from which we also derive a lexicon of 450 Tamil DCs, including The statistics of our corpus reveal that suffixal DCs are as frequent and ambiguous as free-token DCs in Tamil, showing that it is necessary to specifically model both types of DCs. This highlights the significance of our resource for training and evaluating models of Tamil discourse processing.

2. Related Work

2.1. Tamil Language

Tamil is a Dravidian language spoken by approximately 78 million people and is one of the oldest living languages in the world today. Its script is classified as an Abugida, a writing system that lies between an alphabet and a syllabary (Sarveswaran, 2024), comprising a total of 247 distinct characters. Tamil is an agglutinative language with a complex morphological system in which grammatical and semantic information is expressed through suffixation on root nouns, verbs, adjectives and adverbs (Caldwell, 1875; Sarveswaran, 2024). In particular, Tamil DCs are often realized as verbal suffixes, though they may also appear as nominal derivatives and discourse particles.

Tamil is considered a low-resource language due to the limited availability of annotated resources. Most related to the current work, Rao et al. (2011) introduce a manually annotated Tamil corpus of 8500 sentences (352 DRs; 13 unique DCs) focusing exclusively on CAUSAL relations. Using Conditional Random Fields (CRFs) to trained on this data, they reported that identifying causal marker spans is particularly challenging, as the position of the DCs varies due to Tamil's relatively free word order. Rachakonda and Sharma (2011) extend the annotation to other DR types, resulting in a corpus of 511 sentences with 323 DRs (of which 269 are explicit) with 96 unique DCs. Finally, Gopalan et al. (2017) conducted a corpus-based study of cross-linguistic variations across Hindi, Tamil, and Malayalam based on DR annotations of the three languages, including 1341 explicit DRs in Tamil.

Unfortunately, the manual annotation efforts described above are not publicly available. We propose a more scalable approach to high-quality DR annotation for Tamil that combines automatic preprocessing and manual verification. The resulting data is released as an open resource.

2.2. PDTB framework for discourse annotation

In this work, we decided to annotate DR in Tamil based on the PDTB framework because we also want to identify the explicit connectives that trigger the annotated DR in order to construct a lexicon of Tamil DCs. In this framework, each explicit DR annotation specifies the DC, the spans of the two arguments it connects, and a sense label, as seen in Example (1). *Arg2* (in *italics*) is the clause to which the DC is syntactically attached, and corresponds to the label name of the relation, i.e. *"it rained"* is the REASON. The other argument is defined as *Arg1* (in **bold**).

The sense labels in PDTB are arranged in a three-level hierarchy (28 Level-3, 22 Level-2 labels, and 4 Level-1 labels in PDTB3.0). Since DR is an ambiguous phenomenon that even human often disagree on (Sanders et al., 1992; Spooren and Degand, 2010; Das et al., 2017; Zikánová et al., 2025; Hewett and Stede, 2025), existing DR identification tasks typically models up to the granularity of the Level-2 labels (Knaebel, 2021; Braud et al., 2025). Following this, we the current resource is also labelled with Level-2 labels.

2.3. Annotation projection

Since DR resources exist for several languages, a number of previous studies have explored annotation projection from a resource-rich language, typically English, to low-resource languages (Versley, 2010; Laali and Kosseim, 2017; Sluyter-Gäthje et al., 2020; Yung et al., 2023; Bourgonje and Lin, 2024). In particular, Bourgonje and Lin (2024) combine machine translation with word alignment, allowing English DR annotations, automatically produced by PDTB-trained parsers, to be projected to a wide range of target languages.

In contrast, our work projects discourse annotations onto human-translated Tamil text in a parallel corpus, rather than relying on machine-translated output. This choice is motivated by the poor performance of current English–Tamil machine transla-

tion systems (BLEU score of 4.35; Ramesh et al., 2020).

A key limitation of word-alignment–based annotation projection is that alignments are typically available only at the word level, whereas Tamil DCs often occur as suffixes. To address this, unlike prior approaches that rely on projection for direct annotation, we employ annotation projection only as a bootstrapping step to identify potential DC candidates. To accurately identify and extract subword-level DCs, each candidate is subsequently manually analyzed and verified with the help of a Tamil morphological parser (Sarveswaran et al., 2018). The complete processing pipeline is described in Section 4.

3. Adapting PDTB scheme for Tamil

We aim to create a Tamil discourse resource following the PDTB framework. In this contribution, we annotate the spans of the explicit DC tokens or suffixes, together with their arguments and sense labels.

In the PDTB, the DCs that are annotated include conjunctions and discourse adverbials, which are single or multiple tokens. In contrast, for Tamil, we annotate free-standing conjunctions and also suffixal connectives that are attached to verbs, normalized verbs or nouns. Multi-word connectives (as in Example 2), and multi-span connectives are also annotated, but we found only a few cases in our corpus.

Example (2)

Tamil: அவர்கள் தருவதை ஏற்றுக் கொண்டு அடங்கிப்போய் ஊக்குவிக்கின்றனர், **அதன் மூலம்** சமூக கொந்தளிப்புகளை தடுப்பதற்கு உதவுகிறார்கள்.
Avarkaḷ taruvatai ettruk koṇdu aṭaṅkippōy ūkkuvikki a ar, **ata mūlam** camūka kontaḷippukaḷai taṭuppata ku utavuki ārkaḷ.
gloss: *they give*-ACC *accept*-CVB *take*-CVB *submit*-CVB *encourage*-PRS.3PL ***through that*** *social unrest*-PL.ACC *prevent*-INF.DAT *help*-PRS-3PL

English translation: Encourage the passive acceptance of handouts, and **thereby** help prevent social explosions.

DR sense: CONTINGENCY.CAUSAL

PDTB also annotates implicit DRs, but this step requires a list of explicit discourse connectives, and is usually performed by inserting a connective from the list in between two adjacent sentences.

This is problematic in the case of Tamil, as a comprehensive lexicon of Tamil DCs is not yet available, making the choice of DCs for implicit DR annotation unclear. Second, the insertion of DCs is particularly challenging in Tamil and less intuitive than in English due to its complex inflectional morphology. Lastly, the available Tamil–English parallel corpora do not consist of continuous texts, whereas implicit DRs are usually annotated between adjacent sentences. These practical constraints make the current approach unsuitable for implicit DR annotation, and we hence focus on only annotating explicit relations.

As described in Section 2.2, arguments in the PDTB framework are labeled *Argument 1* and *Argument 2* depending on the syntactic attachment of the connective, while the DC span is annotated separately. In our Tamil corpus, we represent argument and DC spans using the notation adopted in the DISRPT shared task (Zeldes et al., 2021). This representation is more flexible and facilitates comparison across discourse frameworks, while preserving all information encoded in the original PDTB format.

Specifically, *Argument 1* and *Argument 2* are ordered according to their linear position in the text, with *Argument 1* preceding *Argument 2*. An additional tag, either *1<2* or *1>2*, indicates whether the connective is syntactically associated with *Argument 1* or *Argument 2*. The DC span is included within the argument span in its original position. This notation is particularly suitable as the Tamil suffixal DCs cannot be detached from their host tokens without rendering the text ungrammatical.

4. Methodology

4.1. Data

We chose the Tamil Samanantar Dataset (Ramesh et al., 2022) as the parallel corpus for annotation projection. The corpus contains a total of 5 million web-crawled sentence pairs taken from news, education and science domains We make use of the first $200,000$ sentence pairs of the corpus.

4.2. Annotation projection from English

Our workflow for annotating DRs on the Tamil text in the parallel corpus consists of three steps. We first apply the Discopy discourse parser (Knaebel, 2021) to obtain discourse annotations on the English side of the parallel corpus. Although the parser identifies both explicit and implicit relations, we retain only explicit relations. Sentence pairs that do not contain an explicit DC on the English side are excluded[1]. We also discard cross-sentence relations, as the order of the sentences fed to the parser does not

[1] Even though the corresponding Tamil sentence may contain an explicit DC due to *explicitation* in translation.

correspond to their original discourse order[2]. After this filtering, we obtain 37,819 sentence pairs.

We then apply AWESoME-align (Dou and Neubig, 2021) to the screened sentence pairs to compute token-level alignments between English and Tamil. Using the token spans of DCs and arguments on the English side, we extract the corresponding aligned tokens on the Tamil side. The DR labels are directly projected as the annotations for the Tamil DRs. As described in Section 3, the argument whose span begins earlier in the text is labeled *Arg1*.

Among the aligned sentence pairs, only 8,225 English DCs are aligned, sometimes jointly with other English tokens, to tokens in the corresponding Tamil sentences. This suggests that in many cases, the English DC is not translated as an explicit standalone token in Tamil; but this may also result from alignment errors.[3] We focus on the aligned DC tokens to identify explicit DRs on the Tamil side. The candidate alignments that include an English DC correspond to 3,322 unique Tamil word forms. Our next step is to identify Tamil DCs, both standalone tokens and suffixes, by separating and removing the content-bearing word segments.

4.3. Suffix-level analysis

We apply the ThamizhiFST morphological analyzer (Sarveswaran et al., 2018) to the candidate alignments to separate the suffixes from the main word stems. The alignments are then grouped on the Tamil side based on the identified suffix (if any), and each alignment is manually inspected to specify valid DC-to-DC alignments, where the Tamil DC may be realized either as a token or as a suffix, while the English DC may consist of one or multiple tokens. For example, based on the alignments shown below, it can be inferred that the Tamil DC corresponding to the English DC *"when"* is the suffix -போது (bothu).

- தேவைப்படும்**போது** *thevaipadum**bothu*** (when needed)
- சென்ற**போது** *sendra**bothu*** (when ... went)
- நகர்த்தும்**போது** *nagarthum**bothu*** (when moved)

The manual alignment post-editing is carried out by one of the authors, a native Tamil speaker. During this process, morphological segmentation errors (over-/under segmentation) and word-alignment errors in the candidate alignments are corrected simultaneously. Specifically, when a candidate alignment is incorrect, i.e., when the discourse sense expressed by the English DC is not conveyed by any part of the aligned Tamil tokens, the original sentence pair is examined to identify alternative Tamil expressions that may realize the labeled DR. If such expressions are found, the alignment is updated accordingly (as in Example (3)). If not, indicating that the DR is implicit or paraphrased in Tamil, the DR instance is discarded (as in Example (4)).

Example (3)

Tamil: அந்த விளையாட்டு ஆபத்தானது, ஏனென்றால் திருட்டும் கொலையும் சர்வசாதாரணமானது என காண்பிக்கிறது.
antha vilaiyaattu aabathaanathu, **yenendraal** thiruttu kollaium sarvasaathaaranamaanathu ena kaanbikkapadukirathu.
gloss: *that game* dangerous-PRS, ***because*** theft-AND murder-AND ordinary-PRS *be* show-PRS.

English translation: The game is considered dangerous **because** it trivializes robbery and murder.

DR sense: CONTINGENCY.CAUSE

Example (4)

Tamil: அங்கே ஒரு மளிகை கடையை திறந்து முழு குடும்பமாக **மாறிமாறி** அதை கவனித்துக்கொண்டோம்.
Ange oru malligai kadai thiranthu mulu kudumbamum **maarimaari** athai kavanithukondanar.
gloss: *There one grocery store*-ACC *open*-CVB *whole family*-ADV *alternatively*-ADV *it take-care*-1PL-NOM

English translation: We opened a grocery store there **and** our whole family took turns working in it.

DR sense: EXPANSION.CONJUNCTION

In Example (3), "because" is incorretly aligned to ஆபத்தானது *aabathaanathu* (dangerous) and is therefore revised manually to ஏனென்றால் *yenendraal* (because). In Example (4), மாறிமாறி *MaariMaari* (again and again) is aligned with "and". However, in this context, மாறிமாறி *MaariMaari*, it should be aligned to "took turns". "And" is therefore inferred implicitly in the Tamil sentence and is not included in the dataset.

After this suffix-level alignment post-editing and verification, a total of 7223 explicit DRs and 1702 unique Tamil DC candidates are collected, but 1260 of these only occur once in the corpus. This is not surprising given the highly inflectional nature

[2]We nevertheless retain the predicted DC spans and project them to Tamil, while labeling the arguments and senses as *unknown*. This part of the data can still be used for DC identification

[3]James and Krishnamurthy (2025) report an AER of 68.2 using Awesome Align for English-Tamil word alignment.

of Tamil. Many of these candidates correspond to different surface forms of the same token or suffix. Therefore, additional normalization is required to extract a list of unique Tamil DCs.

4.4. Manual normalization

As a low-resource language, Tamil has limited preprocessing tools available. Consequently, the normalization of DC candidates is performed largely manually by the same author, with support from string-matching techniques. The main goal of the normalization process is to identify and group various variants of the same DC into a standardized canonical entry.

Phonetic Variation Some Tamil DCs exhibit variation in pronunciation that may also be reflected in their written form, diverging from the standard spelling. These forms therefore represent orthographic variants of the same discourse connective. In Example (5), ஆனா *aana* is an informal spoken variant of ஆனால் *aanaal*, reflected in the phonetical spelling. Both forms correspond to the same connective, which is translated as *but* in English. Here are some more examples of phonetic variants:

- "similarly"
  standard: அதேபோல் *athepol*
  informal: அதேப்போல *atheepola*
  phonetic variation: அதபோல *athapola*
- "but"
  standard: ஆனால் *aanaal*
  phonetic variation: ஆனா *aana*

Example (5)

Tamil: இது மோசமா தெரியலாம் **ஆனா** அந்த கதவுகள்... முழுவதும் மாற்றப்போகுது... அதன் கீழ்பகுதிலயிருந்து தோட்டம் வரை
Ithu mosama theriyalaam **aana** antha kadhavugal.. muluvathum maatrapokuthu... athan keelpaguthiilirrunthu thoodam varai
gloss: *this bad*-NOM *may-look* ***but*** *those door*-PL *whole change-go*-FUT *its down-part-from*-CVB *garden untill*
English translation: I know this looks bad, **but** those patio doors are going to completely revolutionise the flow from their downstairs to their garden
DR sense: COMPARISON.CONCESSION

Since variants of the same DC typically share a common prefix, we inspect the list of DC candidates sorted alphabetically and group variants accordingly. Based on these groupings, we construct a mapping from surface variants to standard DC forms. In total, 90 candidates are identified as variants and grouped into 44 standard DCs, with each standard DC having 1 to 5 variants. Using this mapping, all DC variants (1428 out of the 7233) in the corpus are assigned a normalized standard DC on top of the raw form.

Sandhi consonant linking We also account for orthographic variants arising from sandhi consonant insertion. These are phonological alternations at word boundaries that commonly involve the insertion of a consonant to ease pronunciation when a vowel-final word is followed by a vowel-initial word. Such additional consonants could also be suffixed to a DC, as in அதற்கு**ப்** *adharkku* (therefore), இவ்விதமாய்**த்** *ivvidhamaayi* (in this manner), and உதாரணமாக**க்** *udaaranamaaka*(for example). The last consonant of these DCs are the linking consonants.

Emphasis suffix Another type of variation concerns suffixal DCs that appear in emphasized forms, as shown in the examples below.

- காட்டு *kaatu* (show)
- காட்டு**வதற்கு** *kaatuvatharkku* (in order to show)
- காட்டு**வதற்கே** *kaatuvatharkke* (in order to show; with an emphasis marker)

Since both linking consonants and emphasis suffixes are governed by strict morphological rules, their normalization is straightforward. We generate the corresponding consonant-linking and emphatic forms of each DC as variants and group all such DC candidates under a single canonical entry.

Change in DR sense in translation On top of variants identification, we also manually verify the projected DRs to identify any sense mismatches. The DR sense predicted by the discourse parser on the English texts could be incorrect. [4] Also, DRs are sometimes explicitated or implicitated (Meyer and Webber, 2013; Lapshinova-Koltunski and Carl, 2022; Yung et al., 2023), which means that a DC can be inserted or omitted, or translated to a more/-less specific connective. Our focus is not the meaning shift of the DR translation, but rather whether the projected English DR sense is valid for the Tamil DC in the translation. To verify this, each DR sense projected from English is inspected against the corresponding Tamil connective in the collected lexicon. When the projected sense is intuitively incompatible with the meaning of the Tamil DC, the original sentence pair in the corpus is examined and the sense label is revised accordingly.

In Example (6), the sense of the English DC, CONJUNCTION, is projected to the Tamil DC ஆனால்

[4] The reported accuracy of the Discopy parser on PDTB explicit relations is 78% F1 (Knaebel, 2021).

Sense	unique DC counts word-type	unique DC counts suffix-type	corpus frequency word-type	corpus frequency suffix-type	Top-3 Tamil DCs	Coverage
Expansion.Conjunction	206	91	1072	1093	-உம் (491), மற்றும் (245), மேலும் (236)	45.3%
Comparison.Contrast	93	47	1152	334	ஆனால் (794), -உம் (124), எனினும் (57)	65.6%
Temporal.Asynchronous	116	45	777	373	பின்னர் (179), பிறகு (76), -இல் (66)	27.9%
Contingency.Cause	113	59	615	329	-ஆல் (151), எனவே (90), ஆகவே (70)	33.0%
Temporal.Synchrony	88	48	245	474	-போது (174), -இல் (85), போது (76)	46.7%
Contingency.Condition	56	42	130	370	-ஆல் (190), -ஆனால் (28), -இருந்தால் (18)	47.3%
Expansion.Alternative	13	17	90	30	அல்லது (75), -ஆவிட்டால் (6), -ஆல் (4)	70.8%
Comparison.Concession	19	19	27	59	-உம் (17), -ஆலும் (9), -கூட (5)	36.0%
Expansion.Instantiation	4	1	19	1	உதாரணமாக (14), இதுபோன்று (3), உதாரணமாகும் (1)	90.0%
Contingency.Purpose	1	1	1	18	-காக (18), இருக்க (1)	100.0%
Comparison.Similarity	3	1	11	1	அதேபோல் (5), இதேபோல (4), போன்று (2)	91.7%
Expansion.Equivalence	2	1	11	1	அதாவது (10), என்றால் (1), -ஆக (1)	100.0%
Total	**327**	**127**	**4150**	**3083**	—	—

Table 1: DR sense distribution across unique Tamil DC counts and corpus frequency.

anaal (but). However, this is actually an *explicitation* of the DR; the relation is more explicitly expressed with a DC specifically for CONTRAST or CONCESSION. Since the assigned sense does not match the Tamil DC intuitively, the corresponding corpus samples are inspected and the correct sense, CONCESSION, is assigned.

Example (6)

Tamil: (இவ்வாறு) அவர்கள் சூழ்ச்சி செய்தார்கள். **ஆனால்** அவர்கள் அறியாதவாறு நாமும் சூழ்ச்சி செய்தோம்.
(ivvaaru) avargal soolchi seithaargal **anaal** avargal ariyaathavaaru naamum soolchi seithoom.
gloss: *(Thus) they maneuver do*-PST-3PL ***but*** *they unknowingly we-also maneuver do*-PST-1PL-INCL

English translation: So they plotted a plot, **and** we planned a plan, while they perceived not.

DR sense: EXPANSION.CONJUNCTION →CONTINGENCY.CONCESSION

In Example (7), *"so"* is aligned to வதற்காக *vata kāka* (for that), and the English discourse parser assigns it the label CONTINGENCY.CAUSE. However, the correct connective span should be *"so as to"*, which expresses a CONTINGENCY.PURPOSE relation. Accordingly, the sense label for the Tamil DC is manually revised from CAUSE to PURPOSE.

Example (7)

Tamil: இந்த கிராமங்களில் தலித் மற்றும் பொது வீடுகளுக்கு இடையிலான ஏற்றத்தாழ்வைக் குறைக்க 50 சமூக-பொருளாதார நிலை குறிகாட்டிகளை மேம்படுத்து**வதற்காக** இப்போது இது மறுவடிவமைப்பு செய்யப்பட்டுள்ளது.
Inta kirāmaṅkaḷil talit ma um potu vīṭukaḷukku iṭaiyilā a ē attā vaik ku aikka 50 camūka-poruḷātāra nilai ku ikāṭṭikaḷai mēmpaṭuttu**vata kāka** ippōtu itu ma uvaṭivamaippu ceyyappaṭṭuḷḷatu.
gloss: *This village*-PL-LOC *Dalit and general house*-PL-DAT *between inequality*-ACC *reduce*-INF 50 *socio-economic status indicator*PL *improve*-**INF-PURPOSE** *now this redesign do*-PASS-*pst*

English translation: This has now been redesigned to include 50 socio-economic indicators that have to be improved **so as to** reduce the inequality between Dalit and general households in these villages.

DR sense: CONTINGENCY.CAUSE →PURPOSE

During this DR sense verification process, we additionally identify four sense labels that are not identified by the parser on the English side and therefore are never projected automatically. These labels are incorporated into the Tamil DC lexicon for the corresponding DCs: EXPANSION.INSTANTIATION, EXPANSION.EQUIVALENCE, CONTINGENCY.PURPOSE, and COMPARISON.SIMILARITY.

5. Results

After all verification and normalization steps, the final dataset contains 7233 explicit DR annotations involving 454 unique Tamil DCs. Our **Tamil Discourse Relation Bank** is freely downloadable from https://anonymous.4open.science/r/Tamil-Semi-Automatic-Discourse-Relation and the data structure is shown in Listing 1.

Listing 1: Tamil Discourse Relation Bank data structure

```
"tamil_ann": {
    "line": "string",
    "arg1": {
      "raw": "string",
      "span": [int, int, ...]},
    "arg2": {
      "raw": "string",
      "span": [int, int, ...]},
    "rel_direction": "1>2 or 2<1",
    "connective": {
      "raw_text": ["string", "string"],
      "canonical": ["string", "string"],
      "type": "word/suffix",
      "morphology": {
          "stem": "string",
          "suffix": "string"},
      "span": [
        [int, int, ...],
        [int, int, ...]]}
    "relation": {
      "type": "Explicit",
      "sense": "Contingency.Cause"}
}
```

The sense distributions of both the annotated DRs and the unique DCs are presented in Table 1. We observe that suffixal DCs are prevalent in Tamil discourse, accounting for 25% unique DC types and 40% of the DC occurrences in the corpus. Tamil DCs are also highly ambiguous. To quantify the ambiguity, , we compute the normalized entropy of the sense-label distribution for each unique DC. A DC that is consistently annotated with a single sense label has an entropy of 0, indicating minimal ambiguity. In contrast, a DC whose sense labels are evenly distributed will result in an entropy of 1, indicating maximal ambiguity and difficulty in predicting its sense.

Figure 1 presents the distribution of entropy values for word type and suffixal DCs. The results show that word-type DCs are generally less ambiguous than suffixal DCs, as most have entropy values close to 0, indicating that they are associated with a limited range of senses or display a strongly biased distribution toward a dominant sense. In contrast, more than half of the suffixal DCs exhibit higher entropy values, meaning that they can be interpreted in different ways with similar chance.

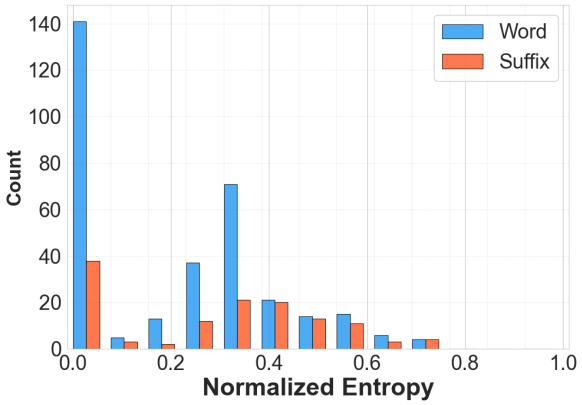

Figure 1: Distribution of sense entropy per DC

As a simple baseline, assigning each DC of the entire corpus its most frequent sense based solely on the DC lexicon yields an accuracy of 58% compared with the actual annotation. In a more realistic setting, assigning the top sense acquired from 80% of the data to the rest of the unseen data yields an accuracy of 27% only. This highlights the difficulty of explicit DR classification in Tamil. Our dataset therefore provides valuable training data for models that aim to predict the correct sense of Tamil DCs in context.

6. Conclusion

In this work, we propose a semi-automatic pipeline for creating a discourse-annotated resource for Tamil. Our approach automatically projects discourse annotations from a resource-rich language, English, and extends this projection by incorporating morphologically segmented Tamil DC suffixes, since Tamil expresses discourse mainly through verb suffixes participial constructions. To make sure these morphological elements align well with English tokens, the projection based on automatic alignment is followed by a series of manual post-editing steps. Each sample of the final 7233 explicit DRs has gone through manual verification to ensure the reliability of the DR annotations. Based on the corpus data, we also constructed the first lexicon of Tamil DCs, which are either isolated tokens or suffixes.

7. Limitation

The main limitation of the current work is its focus on intra-sentential explicit relations, due to the constraint of the parallel corpus (i.e., non-consecutive sentences). Furthermore, coherence shifts occur in manual translation (Blum-Kulka, 1986). On the one hand, the current dataset excludes explicit Tamil connectives that are implicitated or originally implicit in the English texts, as we only align the explicit connectives identified by the discourse parser. On the other hand, subtle translation nuance can unintentionally alter the coherence relations in the target language text. These limitation may affect the accuracy of the annotation projections, and lead to a biased picture of the distribution of Tamil DRs in general. As future work, we plan to develop annotation guidelines based on the DC lexicon and to conduct fully manual annotation on a subset of the current data as well as on a monolingual Tamil corpus. This manually annotated dataset will serve as an additional evaluation benchmark for Tamil shallow discourse parsing.

8. Bibliographical References

Shoshana Blum-Kulka. 1986. Shifts of cohesion and coherence in translation. Interlingual and intercultural communication: Discourse and cognition in translation and second language acquisition studies, 17:35.

Peter Bourgonje and Pin-Jie Lin. 2024. Projecting annotations for discourse relations: Connective identification for low-resource languages. In Proceedings of Workshop on Computational Approaches to Discourse, pages 39–49, St. Julians, Malta. Association for Computational Linguistics.

Chloé Braud, Amir Zeldes, Chuyuan Li, Yang Janet Liu, and Philippe Muller. 2025. The DISRPT 2025 shared task on elementary discourse unit segmentation, connective detection, and relation classification. In Proceedings of the Shared Task on Discourse Relation Parsing and Treebanking, pages 1–20, Suzhou, China. Association for Computational Linguistics.

Robert Caldwell. 1875. A comparative grammar of the Dravidian or South-Indian family of languages. Trübner.

Eunsol Choi, Hannah Rashkin, Luke Zettlemoyer, and Yejin Choi. 2016. Document-level sentiment inference with social, faction, and discourse context. In Proceedings of the Meeting of the Association for Computational Linguistics, pages 333–343, Berlin, Germany. Association for Computational Linguistics.

Iria da Cunha, Juan-Manuel Torres-Moreno, Gerardo Sierra, Luis-Adrián Cabrera-Diego, Brenda-Gabriela Castro-Rolón, and Juan-Miguel Rolland Bartilotti. 2011. The RST Spanish treebank on-line interface. In Proceedings of the International Conference Recent Advances in Natural Language Processing, pages 698–703, Hissar, Bulgaria. Association for Computational Linguistics.

Debopam Das, Manfred Stede, and Maite Taboada. 2017. The good, the bad, and the disagreement: Complex ground truth in rhetorical structure analysis. In Proceedings of the Workshop on Recent Advances in RST and Related Formalisms, pages 11–19, Santiago de Compostela, Spain. Association for Computational Linguistics.

Zi-Yi Dou and Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora. In Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics, pages 2112–2128, Online. Association for Computational Linguistics.

Sindhuja Gopalan, Lakshmi S, and Sobha Lalitha Devi. 2017. Cross linguistic variations in discourse relations among Indian languages. In Proceedings of the International Conference on Natural Language Processing, pages 402–407, Kolkata, India. NLP Association of India.

Jian Guan, Xiaoxi Mao, Changjie Fan, Zitao Liu, Wenbiao Ding, and Minlie Huang. 2021. Long text generation by modeling sentence-level and discourse-level coherence. In Proceedings of the Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing, pages 6379–6393, Online. Association for Computational Linguistics.

Freya Hewett and Manfred Stede. 2025. Disagreements in analyses of rhetorical text structure: A new dataset and first analyses. In Proceedings of the Linguistic Annotation Workshop, pages 35–47, Vienna, Austria. Association for Computational Linguistics.

Jerry R Hobbs. 1978. Why is discourse coherent. Technical report.

Antony Alexander James and Parameswari Krishnamurthy. 2025. POS-aware neural approaches for word alignment in Dravidian languages. In Proceedings of the Workshop on Challenges in Processing South Asian Languages, pages

154–159, Abu Dhabi, UAE. International Committee on Computational Linguistics.

René Knaebel. 2021. discopy: A neural system for shallow discourse parsing. In Proceedings of the Workshop on Computational Approaches to Discourse, pages 128–133, Punta Cana, Dominican Republic and Online. Association for Computational Linguistics.

Majid Laali and Leila Kosseim. 2017. Improving discourse relation projection to build discourse annotated corpora. In Proceedings of the International Conference Recent Advances in Natural Language Processing, pages 407–416.

Ekaterina Lapshinova-Koltunski and Michael Carl. 2022. Using translation process data to explore explicitation and implicitation through discourse connectives. In Proceedings of the Workshop on Computational Approaches to Discourse, pages 42–47, Gyeongju, Republic of Korea and Online. International Conference on Computational Linguistics.

Dongqi Liu and Vera Demberg. 2024. RST-LoRA: A discourse-aware low-rank adaptation for long document abstractive summarization. In Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 2200–2220, Mexico City, Mexico. Association for Computational Linguistics.

Dongqi Liu, Hang Ding, Qiming Feng, Jian Li, Xurong Xie, Zhucun Xue, Chengjie Wang, Jiangning Zhang, and Yabiao Wang. 2026. Discorag: Discourse-aware retrieval-augmented generation. arXiv preprint arXiv:2601.04377.

Thomas Meyer and Bonnie Webber. 2013. Implicitation of discourse connectives in (machine) translation. In Proceedings of the Workshop on Discourse in Machine Translation, pages 19–26, Sofia, Bulgaria. Association for Computational Linguistics.

Akshai Ramesh, Venkatesh Balavadhani parthasa, Rejwanul Haque, and Andy Way. 2020. Investigating low-resource machine translation for English-to-Tamil. In Proceedings of the Workshop on Technologies for MT of Low Resource Languages, pages 118–125, Suzhou, China. Association for Computational Linguistics.

Ted JM Sanders and Wilbert Spooren. 2011. Communicative intentions and coherence relations. In Coherence in Spoken and Written Discourse: How to create it and how to describe it., pages 235–250. John Benjamins Publishing Company.

Ted JM Sanders, Wilbert PM Spooren, and Leo GM Noordman. 1992. Toward a taxonomy of coherence relations. Discourse processes, 15(1):1–35.

Kengatharaiyer Sarveswaran. 2024. Morphology and syntax of the tamil language. arXiv preprint arXiv:2401.08367.

Kengatharaiyer Sarveswaran, Gihan Dias, and Miriam Butt. 2018. Thamizhifst: A morphological analyser and generator for tamil verbs. In International Conference on Information Technology Research (ICITR), pages 1–6. IEEE.

Henny Sluyter-Gäthje, Peter Bourgonje, and Manfred Stede. 2020. Shallow discourse parsing for under-resourced languages: Combining machine translation and annotation projection. In Proceedings of the Language Resources and Evaluation Conference, pages 1044–1050, Marseille, France. European Language Resources Association.

Francesco Sovrano, Monica Palmirani, Salvatore Sapienza, and Vittoria Pistone. 2025. Discolqa: zero-shot discourse-based legal question answering on european legislation. Artificial Intelligence and Law, 33(2):323–359.

Wilbert Spooren and Liesbeth Degand. 2010. Coding coherence relations: Reliability and validity. Corpus Linguistics and Linguistic Theory, 6(2):241–266.

Manfred Stede. 2004. The Potsdam commentary corpus. In Proceedings of the Workshop on Discourse Annotation, pages 96–102, Barcelona, Spain. Association for Computational Linguistics.

Pavlína Synková, Jiří Mírovský, Lucie Poláková, and Magdaléna Rysová. 2024. Announcing the Prague discourse treebank 3.0. In Proceedings of the Joint International Conference on Computational Linguistics, Language Resources and Evaluation, pages 1270–1279, Torino, Italia. ELRA and ICCL.

Suzan Verberne, Lou Boves, Nelleke Oostdijk, and Peter-Arno Coppen. 2007. Evaluating discourse-based answer extraction for why-question answering. In Proceedings of the annual international ACM SIGIR conference on Research and development in information retrieval, pages 735–736.

Yannick Versley. 2010. Discovery of ambiguous and unambiguous discourse connectives via annotation projection. In Proceedings of Workshop on Annotation and Exploitation of Parallel Corpora, pages 83–82.

Jiacheng Xu, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Discourse-aware neural extractive text summarization. In Proceedings of the Annual Meeting of the Association for Computational Linguistics, pages 5021–5031, Online. Association for Computational Linguistics.

Frances Yung, Merel Scholman, Ekaterina Lapshinova-Koltunski, Christina Pollkläsener, and Vera Demberg. 2023. Investigating explicitation of discourse connectives in translation using automatic annotations. In Proceedings of the Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 21–30, Prague, Czechia. Association for Computational Linguistics.

Amir Zeldes, Yang Janet Liu, Mikel Iruskieta, Philippe Muller, Chloé Braud, and Sonia Badene. 2021. The DISRPT 2021 shared task on elementary discourse unit segmentation, connective detection, and relation classification. In Proceedings of the Shared Task on Discourse Relation Parsing and Treebanking, pages 1–12, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Deniz Zeyrek and Kezban Başıbüyük. 2019. TCL - a lexicon of Turkish discourse connectives. In Proceedings of the International Workshop on Designing Meaning Representations, pages 73–81, Florence, Italy. Association for Computational Linguistics.

Deniz Zeyrek, Işin Demirşahin, Ayişiği Sevdik-Çalli, Hale Ögel Balaban, İhsan Yalçinkaya, and Ümit Deniz Turan. 2010. The annotation scheme of the Turkish discourse bank and an evaluation of inconsistent annotations. In Proceedings of the Linguistic Annotation Workshop, pages 282–289, Uppsala, Sweden. Association for Computational Linguistics.

Deniz Zeyrek and Murathan Kurfalı. 2017. TDB 1.1: Extensions on Turkish discourse bank. In Proceedings of the Linguistic Annotation Workshop, pages 76–81, Valencia, Spain. Association for Computational Linguistics.

Yuping Zhou and Nianwen Xue. 2012. PDTB-style discourse annotation of Chinese text. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 69–77, Jeju Island, Korea. Association for Computational Linguistics.

Šárka Zikánová, Anna Nedoluzhko, Jiří Mírovskỳ, and Eva Hajičová. 2025. Gold data and multiple understanding of discourse relations. In International Conference on Text, Speech, and Dialogue, pages 250–262. Springer.

9. Language Resource References

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2003. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In Current and new directions in discourse and dialogue, pages 85–112. Springer.

Debopam Das and Manfred Stede. 2018. Developing the Bangla RST Discourse Treebank. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan. European Language Resources Association (ELRA).

Sindhuja Gopalan, Lakshmi S, and Sobha Lalitha Devi. 2017. Cross linguistic variations in discourse relations among Indian languages. In Proceedings of the 14th International Conference on Natural Language Processing (ICON-2017), pages 402–407, Kolkata, India. NLP Association of India.

Umangi Oza, Rashmi Prasad, Sudheer Kolachina, Dipti Misra Sharma, and Aravind Joshi. 2009. The Hindi discourse relation bank. In Proceedings of the Third Linguistic Annotation Workshop (LAW III), pages 158–161, Suntec, Singapore. Association for Computational Linguistics.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse TreeBank 2.0. In Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), Marrakech, Morocco. European Language Resources Association (ELRA).

Ravi Teja Rachakonda and Dipti Misra Sharma. 2011. Creating an annotated Tamil corpus as a discourse resource. In Proceedings of the 5th Linguistic Annotation Workshop, pages 119–123, Portland, Oregon, USA. Association for Computational Linguistics.

Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan AK, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Mahalakshmi J, Divyanshu Kakwani, Navneet Kumar, Aswin Pradeep, Srihari Nagaraj, Kumar Deepak, Vivek Raghavan, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh Shantadevi Khapra. 2022. Samanantar: The largest publicly available parallel corpora collection for 11 Indic languages. Transactions

of the Association for Computational Linguistics, 10:145–162.

Pattabhi RK Rao, Sobha Lalitha Devi, et al. 2011. Automatic identification of cause-effect relations in tamil using crfs. In International Conference on Intelligent Text Processing and Computational Linguistics, pages 316–327. Springer.

Kumari Sheeja S and Sobha Lalitha Devi. 2022. Automatic identification of explicit connectives in Malayalam. In Proceedings of the WILDRE-6 Workshop within the 13th Language Resources and Evaluation Conference, pages 74–79, Marseille, France. European Language Resources Association.

Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind Joshi. 2019. The PDTB 3.0 annotation manual. Philadelphia, University of Pennsylvania.

Konkani Daan: A Community-Driven Culturally Grounded Speech Corpus for Low-Resource ASR

Milind M. Shivolkar, Vaibhav Gawas, Jyoti D. Pawar
CST-GBS Goa University, Vidyapati Lab-Goa University
Goa, India
milind.shivolkar@unigoa.ac.in, jrat-vidyaapati@unigoa.ac.in, jdp@unigoa.ac.in

Abstract

Culturally grounded speech remains underrepresented in existing Automatic Speech Recognition (ASR) resources for Indian languages. We introduce *Konkani Daan*, a community-driven initiative to collect culturally representative Goan Konkani speech via a participatory web platform, currently comprising over 43.9 hours of 16 kHz recordings with region-specific metadata. To evaluate real-world language coverage, we test a strong Indian multilingual ASR model, AI4Bharat IndicConformer-600M, in zero-shot mode on the Konkani Daan development set (379 utterances), obtaining a Word Error Rate (WER) of 46.46% and Character Error Rate (CER) of 15.47%, indicating substantial domain and cultural mismatch despite nominal support for Konkani. We additionally evaluate a previously developed Konkani ASR model in a cross-corpus setting and conduct a qualitative error analysis of outputs from both systems, identifying recurring challenges including compound-word segmentation, digit-word normalisation, named-entity distortion, and orthographic variation. These findings highlight the need for culturally informed resource design and normalisation-aware evaluation for low-resource Indian languages.

Keywords: Automatic Speech Recognition, Low-Resource Languages, Konkani Corpus, Multilingual Speech Models, Cultural Linguistic Variation, Word Error Rate, Domain Adaptation

1. Introduction

Indian languages are characterised by rich morphology, dialectal diversity, oral traditions, and culturally embedded lexicons shaped by regional history, religion, governance, and everyday social practice. Despite recent advances in multilingual Automatic Speech Recognition (ASR) that have significantly expanded language coverage across Indian languages (Baevski et al., 2020; Gulati et al., 2020; Anand et al., 2023), most large-scale speech models are trained on curated or domain-constrained corpora. Prior work in low-resource speech processing has shown that domain mismatch and limited in-domain data substantially affect recognition performance (Besacier et al., 2014). As a result, culturally dense and community-level speech varieties, particularly those reflecting localised named entities, compound constructions, mixed numeric formats, and contact-induced vocabulary, remain underrepresented in existing training data.

Konkani, the official language of Goa and a constitutionally recognised language of India, presents a compelling case for culturally grounded speech resource development. Goan Konkani exhibits significant intra-state variation, morphological richness, and lexical influence from Portuguese and neighbouring Indo-Aryan languages. Everyday speech frequently includes temple names, administrative terminology, historical references, and region-specific expressions that are rarely captured in generic multilingual corpora. Consequently, nominal language support in multilingual ASR systems does not necessarily imply adequate cultural or domain coverage.

In this work, we introduce *Konkani Daan*, a community-driven initiative for collecting culturally representative speech data through a participatory web platform developed at the Vidyapati Lab, Goa University. The corpus currently comprises over 43.9 hours of 16 kHz speech recordings in Devanagari script, enriched with region-specific metadata within Goa. By preserving authentic transcription practices, including compound segmentation variability and digit word alternation, the dataset intentionally reflects real-world linguistic usage rather than aggressively normalised forms, aligning with calls for culturally informed data documentation and resource design (Bender and Friedman, 2018).

To assess the impact of culturally grounded data on recognition performance, we evaluate a strong multilingual model, AI4Bharat IndicConformer 600M (Anand et al., 2023), in zero-shot mode using Connectionist Temporal Classification (CTC) decoding (Graves et al., 2006) on the Konkani Daan development set (379 utterances). The model achieves a Word Error Rate (WER) of 46.46% and a Character Error Rate (CER) of 15.47%, indicating substantial domain and cultural mismatch despite nominal support for Konkani. Through qualitative error analysis, we show that recognition discrepancies frequently arise from compound boundary variation, differences in numeric normalisation, distortion of culturally specific named entities, and orthographic flexibility typical of community-generated text.

This work foregrounds cultural specificity as a core design principle in corpus creation and evalu-

Proceedings of WILDRE-8, pages 25–32
12 May 2026.

ation for low-resource Indian languages. Our contributions are threefold: (i) we introduce a culturally grounded, community-collected speech corpus for Konkani; (ii) we provide baseline evaluation using a strong multilingual ASR model; and (iii) we analyse culturally driven error patterns to motivate normalisation-aware and culturally informed evaluation strategies for low-resource Indian languages.

Paper Structure: The remainder of this paper is organised as follows. Section 2 introduces the Konkani Daan initiative and describes the corpus design, data-collection methodology, and quality-control procedures. Section 3 presents the experimental setup and baseline ASR evaluation. Section 4 provides a qualitative error analysis highlighting culturally driven recognition challenges. Section 5 discusses implications and future research directions. Section 6 presents the ethics and data governance statement. Section 7 concludes the paper.

2. Related Work

Recent advances in self-supervised learning and multilingual speech modelling have significantly improved Automatic Speech Recognition (ASR) for low-resource languages. Models such as Wav2Vec 2.0 (Baevski et al., 2020) and Conformer (Gulati et al., 2020) have demonstrated strong performance across diverse speech recognition benchmarks. Building on these approaches, the AI4Bharat IndicConformer model (Anand et al., 2023) introduced large-scale multilingual ASR for Indian languages, aiming to expand language coverage across the Indian linguistic landscape.

Despite these advances, domain mismatch remains a major challenge in low-resource speech recognition (Besacier et al., 2014). Recent work in self-supervised and multilingual ASR has shown that models trained on large, curated corpora often degrade in performance when applied to out-of-domain or conversational speech (Baevski et al., 2020; Radford et al., 2023). This limitation is particularly relevant for Indian languages, which exhibit substantial dialectal variation, orthographic flexibility, and culturally embedded vocabulary.

In parallel, the NLP community has increasingly emphasised the importance of culturally grounded data collection and documentation. Bender and Friedman (Bender and Friedman, 2018) argue for better documentation and contextualisation of language resources to ensure responsible and representative language technology development. Within the Indian-language context, several initiatives have focused on building speech and text resources; however, community-driven, culturally embedded speech corpora remain limited.

For Konkani specifically, available speech resources remain limited in scale and diversity, and existing datasets are primarily oriented toward benchmark-style ASR experimentation rather than community-driven, culturally grounded data collection. Prior work has relied largely on curated or institutionally collected corpora, whereas participatory platforms incorporating regional metadata and culturally embedded vocabulary remain relatively scarce. Konkani Daan is intended to complement these earlier efforts by emphasising decentralised collection, community participation, and representation of real-world linguistic variation in Goan Konkani.

Our work contributes to this area by introducing a participatory, culturally grounded speech corpus for Konkani and evaluating multilingual ASR systems under cross-corpus conditions. By focusing on culturally dense community speech and qualitative error analysis, we complement existing multilingual ASR research and highlight the need for normalisation-aware evaluation in low-resource settings.

3. The Konkani Daan Initiative and Corpus Description

Konkani Daan(KD) is a community-driven speech data collection initiative developed under the Vidyapati Lab at Goa University. Designed as a participatory platform, the initiative enables native speakers to contribute speech recordings through a web-based interface, thereby decentralising speech resource creation and foregrounding community involvement. Community-driven and participatory approaches to language data collection have been increasingly recognised as essential for the development of low-resource languages and for equitable NLP resource creation (Bird, 2020; Joshi et al., 2020). In contrast to centrally curated corpora, Konkani Daan explicitly seeks to capture culturally grounded, regionally diverse speech reflective of everyday communication in Goa.

Speech in Konkani Daan is collected primarily through a prompted reading interface rather than spontaneous conversation. Contributors log into the platform, view text prompts displayed on the screen, and record themselves reading randomly assigned utterances using their own devices. The prompts are drawn from a curated pool of culturally grounded textual material sourced from existing linguistic and speech resources, including example sentences from the Konkani WordNet corpus (Prabhugaonkar et al., 2017) (32,804 sentences) and the DMU dataset (AI4Bharat, 2022) developed by AI4Bharat (31,894 sentences). These sources collectively provide a diverse range of culturally relevant content, including administrative ref-

Statistic	Value
Total speech collected	43.9 hours
Total onboarded contributors	42
Active contributors (>50 recordings)	18
Highest individual contribution	1450 recordings
Lowest (within active group)	55 recordings

Table 1: Participation statistics for Konkani Daan.

erences, historical narratives, devotional expressions, region-specific place names, and commonly used named entities relevant to Goan Konkani usage. In this way, culture-specific vocabulary is incorporated directly through prompt design and selection during data collection rather than added retrospectively through post hoc annotation. The resulting dataset therefore reflects read speech with culturally embedded lexical content rather than free conversational speech.

A distinctive feature of the platform is the integration of region-specific metadata collection. Contributors are requested to indicate the region of Goa to which they belong, enabling the corpus to reflect intra-state dialectal variation. Goan Konkani exhibits phonological, lexical, and prosodic differences across regions, and associating speech samples with speaker-region information supports future dialect-aware analysis and region-sensitive ASR adaptation. Prior research has demonstrated that dialectal variation significantly impacts ASR performance and that metadata-aware modelling improves robustness (Koenecke et al., 2020; Ragni et al., 2014). By embedding regional metadata into corpus design, the initiative aligns with culturally informed language resource development.

As of the time of writing, the platform has onboarded 42 contributors and collected approximately 43.9 hours of speech. This corresponds to an average of roughly 1.05 hours of recorded speech per onboarded participant, although the distribution of contributions is uneven, with a smaller active group accounting for a substantial portion of the recordings. The corpus currently consists of prompted speech recordings rather than conversational dialogue.

Audio quality control is integrated directly into the data capture interface. The platform performs real-time environmental sound monitoring before recording and provides contributors with feedback on ambient noise levels. Capture-time screening of acoustic conditions is a recommended best practice in speech corpus development to ensure baseline recording quality in distributed data collection settings (Besacier et al., 2014). All audio is collected at a sampling rate of 16 kHz in WAV format to ensure compatibility with contemporary ASR frameworks.

Transcripts are provided in Devanagari script and reflect authentic community writing practices. The corpus content includes culturally embedded material such as administrative references, historical narratives, devotional expressions, region-specific place names, temple names, and vocabulary influenced by Portuguese and English language contact. As a result, the dataset captures morphological constructions and culturally specific lexicon that are often underrepresented in large multilingual training datasets (Bender and Friedman, 2018).

At the time of writing, the corpus comprises over 43.9 hours of collected speech data. For experimental purposes, the dataset was partitioned into training, development, and test splits. Before experimentation, preliminary structural validation was performed, including transcript presence checks and audio-text pairing verification for the evaluation subset. Comprehensive manual transcription verification remains ongoing as part of the corpus expansion process.

Notably, the Konkani Daan corpus intentionally preserves community transcription characteristics rather than enforcing aggressive normalisation. Naturally occurring orthographic variation is retained to reflect authentic usage. Such decisions align with emerging calls for documentation-aware and culturally grounded corpus design in NLP (Bender and Friedman, 2018; Bird, 2020).

We note that Konkani is written in multiple scripts across different communities, including Devanagari, Roman, Kannada, Malayalam, and Perso-Arabic. In the present work, transcription is restricted to Devanagari because it is widely used in institutional and educational contexts in Goa and provides a practical starting point for corpus development and ASR evaluation. We acknowledge that this choice does not capture the full script diversity of Konkani, and future extensions of the platform will explore multi-script support and script-sensitive interfaces.

4. Experimental Setup and Baseline Evaluation

To assess the robustness of contemporary multilingual ASR systems on culturally grounded Konkani speech, we evaluated two systems: (i) a strong multilingual zero-shot baseline, AI4Bharat IndicConformer-600M (Anand et al., 2023), and (ii) a Wav2Vec2-based model (Baevski et al., 2020) trained and adapted using available Konkani corpora and further fine-tuned with Konkani Daan data.

The evaluation was conducted on the development split of the Konkani Daan corpus, consisting of 379 utterances. All recordings were collected at 16 kHz in WAV format and evaluated in mono configuration. Recognition performance was measured using standard Word Error Rate (WER) and Character Error Rate (CER), which are widely used metrics in automatic speech recognition evaluation

(Graves et al., 2006). Only minimal normalisation (whitespace and punctuation standardisation) was applied to preserve naturally occurring orthographic variation in community transcripts.

4.1. Zero-Shot Multilingual Baseline

The AI4Bharat IndicConformer-600M multilingual model (Anand et al., 2023) was evaluated in zero-shot mode using Connectionist Temporal Classification (CTC) decoding (Graves et al., 2006) with the language code set to *kok*. Despite nominal support for Konkani within the multilingual training framework, the model achieved: **IndicConformer (zero-shot):** WER = 46.46%, CER = 15.47%.

These results indicate substantial domain and cultural mismatch between the multilingual pretraining data and the culturally dense speech patterns present in the Konkani Daan corpus. Prior work has shown that domain mismatch significantly affects ASR performance, particularly in low-resource settings (Besacier et al., 2014). The relatively high WER suggests that compound constructions, named entities, digit-word alternations, and region-specific vocabulary significantly affect recognition accuracy.

4.2. Konkani-trained Wav2Vec2 Model

We additionally evaluated a Wav2Vec2-based ASR system (Baevski et al., 2020) that had been previously developed and trained in earlier work (Shivolkar et al., 2026) on a combination of available Konkani speech corpora (Kunchukuttan et al., 2020; Srinivasan et al., 2023; Prasad et al., 2023) and subsequently adapted using domain-relevant data. Importantly, this model was **not trained jointly on the full combined dataset used in the current experimental setting**. The results reported here correspond strictly to evaluation on the Konkani Daan development split (379 utterances), without additional retraining for this specific configuration.

On this development set, the Konkani-trained model achieved: **Wav2Vec2 (Konkani-trained), KD eval only):** WER = 49.56%, CER = 15.84%.

These results reflect cross-corpus evaluation performance and should not be interpreted as performance after unified multi-corpus training, including the full KD dataset. Although the model had prior exposure to related Konkani data, recognition accuracy remains challenging under strict token-level evaluation on culturally grounded speech.

Qualitative inspection suggests that many mismatches arise from orthographic variation, differences in compound boundaries, and digit-word normalisation rather than from complete semantic misrecognition.

Model	WER (%)	CER (%)
IndicConformer (zero-shot)	46.46	15.47
Wav2Vec2 (Konkani-trained, evaluated on KD)	49.56	15.84

Table 2: Evaluation performance on the Konkani Daan development set (379 utterances). The Wav2Vec2 model was trained on previously available Konkani corpora and evaluated on Konkani Daan in a cross-corpus setting without additional fine-tuning.

4.3. Comparative Summary

The comparable CER values across systems indicate that many word-level discrepancies are attributable to token boundary variation and orthographic flexibility rather than severe character-level corruption. This observation motivates deeper qualitative analysis of culturally driven error patterns and supports the need for normalisation-aware evaluation metrics.

5. Qualitative Error Analysis and Cultural Patterns

To better understand the interaction between culturally grounded speech and ASR performance, we conducted a qualitative analysis of recognition outputs from both the zero-shot IndicConformer model and the Konkani-trained Wav2Vec2 system. Rather than focusing solely on aggregate WER values, we examined recurring error types to identify patterns specific to culturally embedded Konkani speech.

Our analysis reveals four major categories of culturally driven recognition challenges: (i) compound word segmentation and boundary variation, (ii) numeric and date normalisation differences, (iii) distortion of culturally specific named entities, and (iv) orthographic and loanword variation.

5.1. Compound Word Segmentation and Boundary Variation

Konkani frequently employs compound constructions that may be written either as single lexical units or as separated components. Community transcripts preserve this variability. For example:
Compound boundary.

REF: बोलघट्टी

HYP: बोल घट्टी

The compound form is segmented into two tokens in the hypothesis, and minor orthographic variation is observed in (“म्हाल” vs “माल”). Such dif-

ferences inflate WER under strict token-level comparison despite limited semantic divergence. This pattern suggests that culturally grounded corpora require evaluation metrics sensitive to compound variation and boundary flexibility.

5.2. Numeric and Date Normalisation

Digit-word alternation is another prominent source of mismatch. For instance:

REF: 8 ऑक्टोबर 2010

HYP: आठ ऑक्टोबर दोन हजार धा

The hypothesis represents numeric values in spoken word form rather than digit format. Although semantically equivalent, this variation results in significant WER penalties. Similar patterns were observed for administrative identifiers and year references. These findings motivate the adoption of normalisation-aware evaluation (nWER) that accounts for acceptable numeric representation variants.

5.3. Culturally Specific Named Entities

The corpus contains region-specific temple names, place names, and administrative references that are deeply embedded in the Goan cultural context. Recognition errors frequently occur in such cases. For example:

REF: कल्याणेश्वर देवळा

HYP: किलाणेश्वर देवळां

Minor phonetic substitutions or vowel shifts in named entities can substantially affect word-level scoring. These entities are often absent from large multilingual training corpora, which increases recognition difficulty. This highlights the need for culturally informed lexicons and domain-aware language modelling.

5.4. Loanwords and Contact-Induced Vocabulary

Goan Konkani reflects extensive lexical borrowing resulting from centuries of Portuguese colonial administration and subsequent English influence in trade, governance, education, and media. As a result, many English-origin administrative and commercial terms are routinely used in everyday speech, phonologically nativised, and written in Devanagari script. These borrowed forms often lack standardised orthographic conventions, leading to variation across speakers and transcription practices.

Such contact-induced vocabulary poses challenges for multilingual ASR systems, which may not adequately model the phonetic realisation of these adapted loanwords. For example:

REF: इम्पोर्ट

HYP: एम्रट

In this example, the English loanword " “इम्पोर्ट” " (“import”) is recognised as " एम्रट ", reflecting partial phonetic matching but substantial lexical distortion. This type of error indicates that the model captures coarse acoustic patterns but struggles to correctly map them to culturally specific borrowed vocabulary that is underrepresented in multilingual training corpora.

These observations highlight the importance of incorporating culturally informed lexicons and domain-adapted language models when working with community-collected speech that reflects real-world multilingual contact.

5.5. Orthographic and Minor Morphological Variation

A substantial portion of mismatches arises from small orthographic shifts, spacing differences, or morphological inflexion changes rather than semantic misrecognition. The similarity of CER values across systems (approximately 15.47%) supports this observation, suggesting that many errors occur at the word boundary or tokenisation level rather than at the character sequence level.

5.6. Implications for Culturally Grounded ASR Evaluation

Taken together, these patterns demonstrate that standard WER conflates multiple sources of variation, including acceptable orthographic alternatives, numeric format differences, and culturally specific lexical items. While such variability presents genuine modelling challenges, it also reflects the authenticity of community-driven corpora. Therefore, culturally grounded speech resources require evaluation methodologies that incorporate normalisation strategies and culturally informed lexicons to distinguish between semantic errors and orthographic divergence.

These patterns collectively suggest that culturally grounded corpora challenge not only acoustic modelling but also the orthographic standardisation assumptions embedded in multilingual ASR systems.

6. Discussion and Future Directions

The findings of this study highlight the critical importance of culturally grounded resource design in

low-resource speech technology. Although large multilingual ASR systems such as IndicConformer (Anand et al., 2023) nominally support a wide range of Indian languages, zero-shot evaluation on the Konkani Daan corpus reveals substantial performance degradation when confronted with region-specific vocabulary, compound constructions, digit-word alternations, and culturally embedded named entities. Prior work has shown that domain mismatch significantly affects ASR performance in low-resource contexts (Besacier et al., 2014). These results suggest that language coverage alone does not guarantee cultural coverage.

The comparatively high WER observed across both evaluated systems underscores the intrinsic complexity of community-driven speech data. Unlike curated broadcast or read-speech corpora, Konkani Daan reflects authentic linguistic usage, including orthographic flexibility, morphological richness, contact-induced lexical forms, and informal punctuation practices. Such variation aligns with broader concerns regarding representational bias and under-documentation in NLP datasets (Joshi et al., 2020; Bender and Friedman, 2018). While these characteristics pose challenges under strict token-level evaluation, they are essential for capturing real-world speech variability. In this sense, the corpus intentionally prioritises representational authenticity over aggressive normalisation.

The similarity in Character Error Rate (CER) across systems further suggests that many recognition discrepancies occur at token boundaries or involve minor orthographic variations rather than complete lexical substitutions. This observation motivates the exploration of orthography-aware normalisation and evaluation strategies. Recent discussions in ASR research have emphasised the limitations of raw WER as the sole evaluation metric and advocate for linguistically informed alternatives (Morris et al., 2004). Future work will investigate a normalisation-aware Word Error Rate (nWER) that accounts for digit-word equivalence, compound segmentation variation, and standardised spelling mappings. Such approaches may provide a more linguistically grounded assessment of recognition performance in culturally diverse contexts.

Another promising direction involves incorporating culturally informed lexicons and expanding language models. Named entities specific to Goan regions, temple names, administrative identifiers, and contact-influenced vocabulary could be integrated into decoding frameworks to improve robustness. Additionally, leveraging the regional metadata collected through the Konkani Daan platform may enable dialect-aware modelling or region-sensitive adaptation strategies, an approach shown to improve ASR fairness and robustness across speaker communities (Koenecke et al., 2020).

Beyond ASR performance, the broader contribution of Konkani Daan lies in its participatory methodology. Community-driven data collection has been increasingly recognised as essential for equitable and sustainable language technology development in low-resource settings (Bird, 2020). By engaging speakers across Goa and capturing region-specific linguistic variation, the initiative demonstrates a scalable model for culturally anchored speech resource development.

In summary, the results presented in this paper demonstrate that culturally dense, community-collected speech corpora reveal limitations in current multilingual ASR systems while simultaneously offering opportunities for linguistically informed modelling innovations. Addressing these challenges requires not only architectural improvements but also evaluation frameworks and resource design principles that centre on cultural variability as a primary consideration.

7. Ethical Considerations and Data Access

Konkani Daan is a community-driven speech data collection initiative designed with voluntary participation and transparency as core principles. Contributors register on the platform and provide informed consent before submitting recordings. Participation is optional, and contributors may choose the content they wish to record.

During registration, limited demographic and regional metadata are collected to support research on dialectal variation and fairness in ASR systems. These may include the region within Goa and basic speaker information. The collection of such metadata is intended solely for research purposes and to enable dialect-aware and bias-aware modelling. No personally identifiable information is released with the dataset.

The platform performs real-time environmental noise monitoring to ensure recording quality at the time of capture. The dataset is intended strictly for academic and research use in low-resource speech technology.

To protect contributor privacy and ensure responsible usage, the Konkani Daan dataset is **not publicly downloadable**. Access to the data is provided **on a request basis**, subject to review and authentication at the administrative level of the hosting institution. Approved users must agree to ethical and non-commercial usage guidelines before receiving access to the data.

8. Conclusion

In this paper, we introduced Konkani Daan, a community-driven, culturally grounded speech cor-

pus developed under the Vidyapati Lab at Goa University. The corpus, comprising over 43.9 hours of Goan Konkani speech with region-specific metadata, captures authentic linguistic variation including compound constructions, digit-word alternations, culturally embedded named entities, and contact-influenced vocabulary.

Through zero-shot evaluation of a strong multilingual ASR model (IndicConformer-600M) and comparison with a Konkani-trained Wav2Vec2 system, we demonstrated that culturally dense speech presents significant recognition challenges. Despite nominal language coverage, the zero-shot multilingual model achieved a WER of 46.46% on the Konkani Daan development set, indicating substantial domain and cultural mismatch. Qualitative analysis revealed that many mismatches stem from differences in compound segmentation, numeric normalisation, named-entity distortion, and orthographic variation, rather than purely semantic recognition failure.

Our findings underscore the importance of culturally informed resource development and evaluation methodologies for low-resource Indian languages. Future ASR research should incorporate normalisation-aware metrics, culturally grounded lexicons, and dialect-aware modelling strategies to more accurately assess and improve recognition performance in real-world settings.

By centring community participation and regional diversity, Konkani Daan provides both a practical speech resource and a conceptual framework for integrating cultural specificity into language technology development. The Konkani Daan corpus will be progressively released to the research community following ongoing validation and ethical review.

9. Bibliographical References

AI4Bharat. 2022. Digital multilingual utterances (dmu) dataset. https://ai4bharat.org.

Pratyush Anand, Ashish Gupta, Anshuman Goyal, Anoop Kunchukuttan, Animesh Prasad, et al. 2023. Indicconformer: A multilingual speech recognition model for indian languages. In *Proceedings of Interspeech*.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Laurent Besacier, Etienne Barnard, Alexey Karpov, and Tanja Schultz. 2014. Automatic speech recognition for under-resourced languages: A survey. *Speech Communication*, 56:85–100.

Steven Bird. 2020. Decolonising speech and language technology. *Proceedings of COLING*.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd International Conference on Machine Learning (ICML)*.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, and Yonghui Wu. 2020. Conformer: Convolution-augmented transformer for speech recognition. In *Proceedings of Interspeech*.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the nlp world. *ACL*.

Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Meredith Quartey, Zeresenay Mengesha, Colin Toups, John R. Rickford, Dan Jurafsky, and Sharad Goel. 2020. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*.

Anoop Kunchukuttan, Animesh Prasad, Mitesh M. Khapra, et al. 2020. Open speech corpora for indian languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*.

Andrew Morris, Viktoria Maier, and Phil Green. 2004. From wer and ril to mer and wil: Improved evaluation measures for connected speech recognition. In *Proceedings of Interspeech*.

Manisha Prabhugaonkar, Prashant Bhandari, and Sangeeta Kamat. 2017. Building the konkani wordnet. In *Proceedings of the Global WordNet Conference*.

Animesh Prasad et al. 2023. Shrutilipi: Building speech recognition datasets for indian languages. In *Proceedings of the International Conference on Speech and Computer (SPECOM)*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision.

Anton Ragni, Mark JF Gales, Sibsankar Rath, and Yongqi Wang. 2014. Data augmentation for low resource languages. *Interspeech*.

Milind Shivolkar et al. 2026. Cross-domain generalisation in low-resource asr: A multi-corpus investigation of konkani using wav2vec2. In *Proceedings of the International Conference on Intelligent Computing and Communication (ICICI)*. Accepted.

Anirudh Srinivasan et al. 2023. Indicvoices: A large-scale multilingual speech dataset for indian languages. In *Proceedings of Interspeech*.

Is Literal Annotation Enough? Building an Annotation Framework for Metonymic Named Entities in Marathi

Pratibha Dongare
The English and Foreign Languages University
pratibhaphdlandp22@efluniversity.ac.in

Abstract

Named Entity Recognition (NER) has been a core task of natural language processing (NLP) since the Message Understanding Conferences (MUCs). Data annotation plays a crucial role in this task. However, existing annotation studies often rely on the literal sense of entities. Such annotations may lead to inconsistencies, while resolving ambiguity introduced by figurative tropes like metonymy. For example, in *India won the series*, *India* refers to a sports team instead of a geographic location. Understanding such non-literal senses is crucial for various NLP applications such as Question Answering, Information Extraction, etc. By addressing this gap, this study presents an annotation framework and detailed guidelines for annotating metonymic readings of named entities in Marathi, an Indo-Aryan language spoken in the central-western region of India. The study uses news corpus from various domains. It presents a two-tiered annotation framework for annotating conventional metonymies in Marathi language. Further, it describes the annotation framework applied to a corpus of 1,279 Marathi sentences. The result shows the inadequacy of literal-only annotation as 53.6% of named entity spans have metonymic readings. This study makes a crucial contribution for resource development for low-resource languages that share similar linguistic structures and cultural contexts. The paper describes the framework with necessary examples, challenges and concludes with a future scope.

Keywords: metonymy detection, named entities, Marathi language

1. Introduction

Named Entity Recognition (NER) has been a core task in NLP since the Message Understanding Conferences (MUC) (Grisham and Sundheim, 1996; Nadeau and Sekine, 2007) and the CoNLL shared task (Tjong Kim Sang and De Meulder, 2003) . The primary objective of NER is to identify and classify proper nouns that refer to real-world entities. Conventional categories include Person (PER), Location (LOC), and Organization (ORG). However, annotations often rely on the literal sense of entities, which can lead to inconsistencies when annotators attempt to resolve ambiguity introduced by figurative tropes such as metonymy. Metonymy is a cognitive and linguistic phenomenon, a figurative trope in one entity is used to refer to another entity associated with it (Johnson and Lakoff, 1980). For instance, in example (1), ***India***, a geographic location, refers metonymically to a sports team.

(1) ***India*** *won the series.*

In another example (2) (McShane and Nirenburg, 2021), ***the spiky hair*** refers to a particular person having spiky hair.

(2) ***The spiky hair*** *just smiled at me.*

Several studies have focused on the metonymy resolution task (Markert and Nissim, 2002a,b; Markert and Hahn, 2002; Markert and Nissim, 2007, 2009; Gritta et al., 2017; Gritta, 2019). The majority of this foundational work focused on English, with a limited attention given to German (Markert and Hahn, 2002) and French (Poibeau, 2006). While annotating named entities, metonymic senses of entities need to be annotated since such instances frequently occur in the text. Markert & Hahn (Markert and Hahn, 2002) noted 17% of metonymic instances in German magazines. This makes the nature of metonymy regular, prevalent, and productive (Markert and Nissim, 2002b).

The present study considers Marathi, an Indo-Aryan language spoken in the central-western region of India. Although several studies have focused on the foundational task of NER for the Marathi language (Patil et al., 2016, 2020; Litake et al., 2022, 2023), specific metonymic readings of named entities are rarely explored. This study addresses this critical gap by presenting annotation guidelines currently being applied to construct a comprehensive framework for metonymy detection in Marathi.

2. Dataset

For the analysis, 1,279 sentences from Marathi news across politics, finance, sports, and travel domains, entirely in Devanagari script were used. The corpus follows the structure of the SemEval 2007 shared task dataset (Markert and Nissim, 2007) and expands the scope of metonymic entity types to include PER and MISC categories in addition to LOC and ORG (Nissim and Markert, 2003; Gritta et al., 2017). Existing Marathi NLP resources such as L3Cube-MahaBERT and MahaCorpus (Litake et al., 2023, 2022) offer strong general-purpose representations for Marathi text. However, these

Proceedings of WILDRE-8, pages 33–38

resources were not adopted for the present corpus. Three key reasons motivate this decision. First, existing Marathi NER datasets follow conventional literal annotation schemes. They do not account for metonymic readings of named entities. Second, this study requires a controlled, domain-specific corpus from contemporary news text. Such a corpus better captures the metonymic patterns relevant to information extraction tasks. Third, the tagset used here extends standard NER categories with an explicit Literal/Metonymic distinction. This distinction is absent from all existing Marathi corpora. The annotated data contains 1,741 named entity spans (2,363 NE tokens). Of these, 933 spans (53.6%) are metonymic and 808 spans (46.4%) are literal. Table 1 presents the span-wise distribution across the four NE categories. Fig. 1 shows the span count by entity type. At the sentence level, 418 sentences (32.7%) contain no named entities. 292 (22.8%) contain only literal NEs, 320 (25.0%) contain only metonymic NEs, and 249 sentences (19.5%) contain both literal and metonymic NEs within the same sentence. This shows that the two readings co-occur frequently in Marathi text. Fig. 2 presents the sentence-level distribution of these senses.

NE Type	Literal	Metonymic	Total
LOC	212	580	792
MISC	79	20	99
ORG	23	313	336
PER	494	20	514
Total	**808**	**933**	**1741**

Table 1: Span-wise distribution of entities.

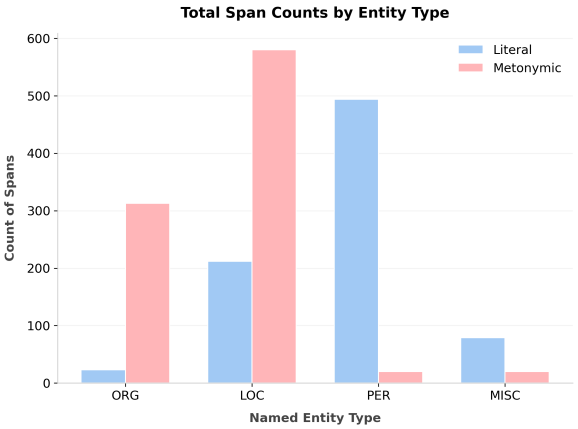

Figure 1: Total span count by entity type

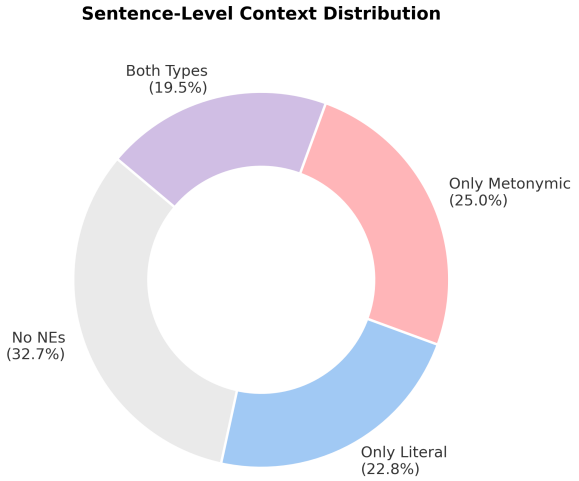

Figure 2: Sentence level distribution of senses

3. Annotation Guidelines and Framework

To achieve high consistency and reliability in metonymy annotation, a principled framework is necessary. Among the previous studies, (Markert and Nissim, 2002b) provided a foundational framework for metonymy annotation in English. This study presents a two-tiered annotation system: first, a set of general principles and second, a specific framework for metonymic interpretations across categories.

3.1. General Principles

These general principles ensure precision in corpus creation, minimizing subjectivity and ensuring high Inter-Annotator Agreement (IAA).

- **Consistency and Uniformity:** *Annotate similar patterns in a similar way across all domains and categories.*
  Apply the same annotation rule uniformly regardless of the position of the entity in the sentence or domain of the text. For example, in (1), if *India* (LOC) referring to its sports team is annotated as a metonymic case of LOC in a sports article, then other entities referring to its team must also be annotated as LOC-metonymy in a different article. If abbreviated forms (e.g., *ISRO*, *UP*) are annotated, then other acronyms (e.g., *UNESCO*, *NATO*) must be annotated using the same logic.

- **Economy:** *Prioritize practicality and efficiency.*
  Avoid fine grained or overlapping labels that increase cognitive load on the annotator(s) resulting in inconsistencies. For instance, this principle is the justification for the No Nested

Entities rule, as a coarse-grained annotation structure is more practical for both annotators and computational models. For example, in (1), *India* is marked as LOC with a metonymic sense and not for the pattern of metonymy (Location-for-organization).

- **Minimal Span:** *Select the smallest possible sequence of tokens that fully identifies the named entity.*
  In a span, include all essential components of the proper name, but exclude contextual modifiers. For example, in the phrase *the former Prime Minister Indira Gandhi*, the minimal span is *Indira Gandhi*. The modifiers *the former*, and the title *Prime Minister*, are all excluded as they are not part of the proper name itself (ref to section 3.2.1).

- **Single Span:** *An entity mention must be annotated as a single, continuous unit*.
  Only the continuous text spans are annotated. Exclude split mentions or discontinuous entities. Split mentions are annotated as separate entities. Example *Students of Mumbai and Pune University* requires two tags. *Pune University* is tagged as an organization, while *Mumbai* is tagged as location metonymy as it is a split mention.

- **Nested Entities:** *Only the outermost, most salient entity in a nested construction is annotated.*
  When an entity is embedded within a larger entity (e.g., *Maharashtra* within *Government of Maharashtra*), annotate only the outermost entity. This prevents ambiguous, overlapping spans and directly supports the Economy and Single Span principles.

- **Default to Literal:** *When in doubt, always annotate the entity as literal.*
  If an entity's usage is ambiguous or could plausibly be interpreted as either literal or metonymic, the entity is tagged as literal. This ensures that the metonymic instances in the final corpus represent high-confidence. For example,

  (3) The name ***Tata*** *is enough.*

  Here, ***Tata*** is ambiguous. It could refer to the person, the organization, or metonymically to the reputation associated with the name. The context alone is insufficient to determine the correct reading. Following the Default to Literal principle, it is tagged as literal PER or ORG depending on the contextual judgment.

- **No Metaphors and Focus on Conventional Metonymy:** *The framework's scope is strictly limited to metonymy within Named Entities.*
  This study targets metonymic sense in named entities (PER, LOC, ORG, MISC). The framework explicitly excludes all other figurative language, such as metaphor. For example,

  (4) *He is the* ***Sachin Tendulkar*** *of his team.*

  The entity ***Sachin Tendulkar*** could mean a metaphor for a great batsman. Non-NE metonymy (unconventional metonymy), such as in example (2) (a Part-for-Whole metonymy using a common noun) is excluded from the annotation.

- **Script & Form Neutrality:** *Annotation decisions are independent of the script, spelling, or form of the entity.*
  This principle helps in annotating a mixed-script (Devanagari/Roman) corpus. For example, the Romanized and Devanagari of acronym *ISRO* (Indian Space Research Organization) are both annotated with the same label (e.g., ORG), as they refer to the identical real-world entity.

- **Data Handling and Annotation Platform:** *Raw text data is annotated using the INCEpTION platform to ensure technical consistency and applicability.*
  The raw, unprocessed news text provided in *.txt* format. To capture the raw, real-world data and linguistic variations, this approach is used. For annotation, INCEpTION (Klie et al., 2018), an open-source, multi-layer annotation platform is used. The platform is user-friendly, offers layer customization, supports for diverse languages and scripts, including Devanagari. This technical choice improves the interoperability and future applicability of the corpus. Fig. 3 shows an example of annotated sentence using this platform. The sentence translates to: *Various international agreements have been criticized for being unfair to the United States.* Here, the *United States* is annotated as metonymic entity.

Figure 3: Example of annotated sentence using INCEpTION platform.

3.2. Category Specific Annotation Scheme

The present annotation framework adopts standard Named Entity categories, accounting for metonymy

across Location, Organization, Person, and Miscellaneous entities.

3.2.1. Entity Span Guidelines

These guidelines refine the general principles to handle complex cases where modifiers are present.

- **Pre/Post Modifiers (Trigger Words):** Modifiers are generally excluded from the entity span as per the Minimal Span principle (e.g., *the*, *former*). However, there are two key exceptions:
  1. **Category-changing Names:** If a modifier combines with an entity to form a new, distinct entity of a different category, the entire phrase is annotated as the new category. For example, in *Temples of Bharat*, *Bharat* is a LOC, but the entire phrase is the proper name of a book (a MISC entity). Similarly, *Shanghai summit* is annotated as a single MISC (event) entity, not as a LOC-metonymy for *Shanghai.*
  2. **Organizational Proper Names:** When a person's name functions as a standard part of an organization entity (e.g., *Modi government*, *Trump administration)*, the entire, continuous span (*Modi government*) is annotated as a single entity.
- **Ambiguity Resolution:** All instances of high ambiguity where the context is insufficient for a clear decision are resolved by applying the Default to Literal principle.

3.2.2. Specific Metonymic Interpretations

The following are the representative examples of category-wise metonymic readings.

- **Location Names:** A location entity is used to refer to an associated institution, an event, or its inhabitants. It includes place-for-government /team. For example,

  (5) ***India*** *hosted the G-20 conference.*

  Here, ***India*** (a location) metonymically refers to the Indian government or organizing body (an organization).
- **Organization Names:** An organization entity is used to refer to its associated personnel, product, or actions. For instance,

  (6) ***France 24*** *posted a video about the event.*

  France 24 (an organization) refers to the employees (a person or a group) of that organization.
- **Personal Names:** A personal entity is used to refer to their associated role, office, or a creation (work, brand, etc.). For example,

  (7) ***Armani*** *still owns the show.*

  Armani (a person) refers to the fashion brand (organization) created by that person.
- **Miscellaneous (MISC):** This category includes metonymic uses of proper nouns that do not fit other three categories. This includes awards, legislation, acts, etc. A common case is an award named after a person, where the entity refers to the object, not the person. For example, (8) *He received the* ***Dadasaheb Phalke Award***. The entity ***Dadasaheb Phalke Award*** is annotated as MISC. It refers to the award itself, which is named after the person *Dadasaheb Phalke*.

The distribution across NE categories reveals striking asymmetries that highlight the need for metonymy annotation. LOC entities are metonymic in 73.2% of spans (580 of 792), and ORG entities in 93.2% of spans (313 of 336). This suggests that a literal-only pipeline would mislabel the large majority of these entities in Marathi news text. PER entities, by contrast, are predominantly literal (494 of 514 spans, 96.1%). Fig. 4 shows the metonymy vs literal proportion per entity class.

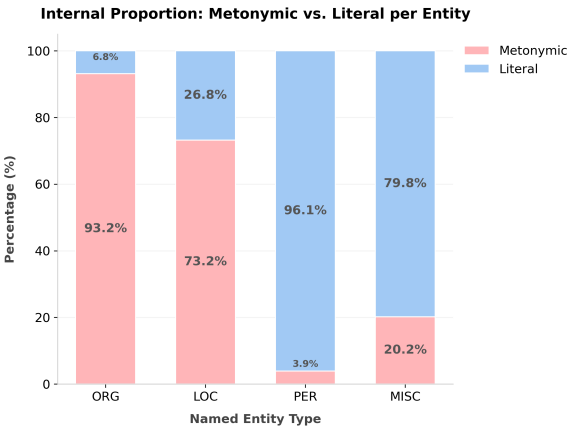

Figure 4: Internal proportion of metonymy vs. literal sense per NE

4. Challenges

The annotation process, even when guided by explicit guidelines, is challenging and complex. This study encountered several challenges, primarily related to linguistic ambiguity, resource constraints specific to Marathi.

4.1. Ambiguity

One of the main challenges is resolving the inherent ambiguity of metonymic usage. While the framework employs the Default to Literal principle to handle confusing cases, contexts often arise where

the metonymic shift is subtle. For instance, in a political domain, deciding whether a location entity like *Delhi* refers to the literal place, the government, or something else requires significant contextual understanding. Subjectivity can lead to inconsistencies and reduced Inter-Annotator Agreement. The Default to Literal principle provides a deterministic resolution strategy in such cases, ensuring that only high-confidence metonymic instances are retained in the final corpus.

4.2. Overlapping of Principles

Creating an effective framework requires balancing detailed instruction with the principle of Economy. Defining too many principles or overlapping guidelines can confuse annotators, leading to errors. For example, the precise interaction between the Minimal Span principle and the handling of pre or post modifiers (trigger words) must be defined carefully. A complex framework, even if theoretically sound, is practically difficult to deploy. This tension was resolved iteratively during the annotation process, with the Economy principle taking precedence when guidelines conflicted.

4.3. Resource and Data Constraints

Working with raw, unprocessed news data in Marathi presents several difficulties. The data is often subject to:
1. **Script and Orthographic Variation:** Spelling variations in the use of script, particularly in the representation of names and acronyms. Such variations and errors affect the annotation process.
2. **Unprocessed Text:** While standard preprocessing usually removes punctuation, this corpus retains the raw text. This results in merged tokens (e.g., punctuation attached to words), which creates significant orthographic noise and complicates the identification of precise entity boundaries. the presence of numerous punctuations and symbols creates noise and complicates the task. The Script and Form Neutrality principle directly addresses orthographic variation. This ensures that annotation decisions remain consistent regardless of spelling or script differences in the raw data.

5. Conclusion and Future Scope

This study presented a principled two-tiered annotation framework for identifying metonymic named entities in Marathi, along with a corpus of 1,279 annotated sentences. The framework establishes eight general principles and category-specific guidelines covering LOC, ORG, PER, and MISC entities. Applied to a Marathi news corpus, the framework reveals that 53.6% of named entity spans are metonymic, with particularly high metonymic rates in ORG (93.2%) and LOC (73.2%) categories. To the best of our knowledge, no prior work has explored an annotation scheme and corpus for metonymy detection in Marathi. The framework is designed to be extensible to other Indo-Aryan languages with similar linguistic structures. Future work will involve completing corpus annotation and training baseline models to evaluate the efficacy of the guidelines. As the corpus scales, intra-annotator agreement will be computed on a re-annotated subset to validate the consistency and reliability of the framework.

6. Bibliographical References

R Grisham and B Sundheim. 1996. Message understanding: a brief history. In *Proceedings of the Sixth Message Understanding Conference*.

Milan Gritta. 2019. *Where are you talking about? advances and challenges of geographic analysis of text with application to disease monitoring*. University of Cambridge (United Kingdom).

Milan Gritta, Mohammad Taher Pilehvar, Nut Limsopatham, and Nigel Collier. 2017. Vancouver welcomes you! minimalist location metonymy resolution. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1248–1259.

Mark Johnson and George Lakoff. 1980. *Metaphors we live by*, volume 1. University of Chicago press Chicago.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart De Castilho, and Iryna Gurevych. 2018. The inception platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th international conference on computational linguistics: system demonstrations*, pages 5–9.

Onkar Litake, Maithili Sabane, Parth Patil, Aparna Ranade, and Raviraj Joshi. 2023. Mono versus multilingual bert: A case study in hindi and marathi named entity recognition. In *Proceedings of 3rd International Conference on Recent Trends in Machine Learning, IoT, Smart Cities and Applications: ICMISC 2022*, pages 607–618. Springer.

Onkar Litake, Maithili Ravindra Sabane, Parth Sachin Patil, Aparna Abhijeet Ranade, and Raviraj Joshi. 2022. L3cube-mahaner: A marathi named entity recognition dataset and bert models. In *Proceedings of the WILDRE-6*

Workshop within the 13th Language Resources and Evaluation Conference, pages 29–34.

Katja Markert and Udo Hahn. 2002. Understanding metonymies in discourse. *Artificial intelligence*, 135(1-2):145–198.

Katja Markert and Malvina Nissim. 2002a. Metonymy resolution as a classification task. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pages 204–213.

Katja Markert and Malvina Nissim. 2002b. Towards a corpus annotated for metonymies: the case of location names. In *LREC*.

Katja Markert and Malvina Nissim. 2007. Semeval-2007 task 08: Metonymy resolution at semeval-2007. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 36–41.

Katja Markert and Malvina Nissim. 2009. Data and models for metonymy resolution. *Language Resources and Evaluation*, 43(2):123–138.

Marjorie McShane and Sergei Nirenburg. 2021. *Linguistics for the Age of AI*. Mit Press.

David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26.

Malvina Nissim and Katja Markert. 2003. Syntactic features and word similarity for supervised metonymy resolution. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 56–63.

Nita Patil, Ajay Patil, and BV Pawar. 2020. Named entity recognition using conditional random fields. *Procedia Computer Science*, 167:1181–1188.

Nita Patil, Ajay S Patil, and BV Pawar. 2016. Issues and challenges in marathi named entity recognition. *International Journal on Natural Language Computing (IJNLC)*, 5(1):15–30.

Thierry Poibeau. 2006. Dealing with metonymic readings of named entities. *arXiv preprint cs/0607052*.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Bengali-English and Hindi-English Code Mixed Speech Data with Disfluencies

Anuran Mitra*, **Tapabrata Mondal***, **Anirvan Chakravarty***, **Sivaji Bandyopadhyay***
*Jadavpur University, Kolkata, India
{anuranm3, tapabratamondal, anirvanchakravarty39, sivaji.cse.ju}@gmail.com

Abstract

Spontaneous speech in multilingual communities such as India frequently combines code-switching (CS) and disfluencies, yet existing Bengali–English and Hindi–English speech corpora largely consist of fluent or scripted utterances. This limits their suitability for developing and evaluating automatic speech recognition (ASR) systems intended for real conversational settings, particularly in micro-resource scenarios. We introduce BEHE-CMDisfl, a synthetic speech corpus that explicitly integrates disfluency phenomena within Bengali–English and Hindi–English code-mixed (CM) utterances. The textual content was generated using prompting strategies with large language models (LLMs) to encourage controlled switching and varied disfluency patterns, including filled pauses, repetitions, and restarts. The utterances were subsequently synthesized using Indic Parler text-to-speech (TTS) system. To demonstrate usability, we establish a reproducible GMM–HMM baseline for Bengali–English ASR using Kaldi on a 1.3-hour subset of the corpus. In our experiments, improvements were mainly observed after ensuring consistency in the pronunciation lexicon and applying phonetic normalization, with the best setup reaching a word error rate (WER) of 37.74%. A closer look at the decoded transcripts suggests that filled pauses and repetitions are not automatically collapsed, but appear in the output, indicating that the disfluency cues present in the synthetic speech are captured during recognition.

Keywords: code-switching (CS), disfluencies, automatic speech recognition (ASR), code-mixing (CM), large language models (LLMs), text to speech (TTS), word error rate (WER)

1. Introduction

Multilingual communication is a defining feature of the South Asian linguistic landscape. In daily communication, speakers frequently blend languages such as Bengali, Hindi, and English within a single utterance, a phenomenon commonly referred to as code-mixing or code-switching (Auer, 2013; Moyer, 2002). ***Code-mixing (CM)*** refers to the mixing of linguistic units such as morphemes, words, phrases, or clauses from two grammatical systems within a sentence, whereas ***code-switching (CS)*** denotes alternation across sentences or discourse segments. CM is intra-sentential, whereas CS is inter-sentential, but quite often both are used interchangeably (Kim, 2006). This linguistic blending is particularly prevalent in multilingual societies such as India (Harya, 2018), where conversational speech naturally exhibits both cross-lingual alternation and spontaneous irregularities.

Alongside multilingual blending, spontaneous speech exhibits disfluencies such as filled pauses ("uh," "um"), repetitions, hesitations, and restarts, which are natural by-products of human planning and self-monitoring during speech production (Shriberg, 2001; Adell et al., 2006). However, most existing speech corpora for Indic languages consist of scripted, fluent recordings (Saranya et al., 2025; Diwan et al., 2021), and contemporary ASR systems often treat non-lexical tokens such as filled pauses as noise to be removed (Kundu et al., 2022; Zayats et al., 2016). This sanitization strategy simplifies decoding but limits the applicability of ASR systems for conversational AI, sentiment analysis, and speech-based behavioral modeling, where disfluencies carry meaningful information. While several text-based code-mixed resources exist, such as BnSentMix (Alam et al., 2024) and SentMix-3L (Nishat Raihan et al., 2023) but they do not capture acoustic variability. Similarly, studies focused on disfluencies have advanced toward multilingual or synthetic augmentation settings (Bhat et al., 2023; Romana et al., 2024; Amann et al., 2024), yet remain largely English-centric or text-bound. Speech-level corpora that jointly represent code-mixing and disfluency for Indic languages remain scarce. Even in Hindi-English ASR (Sitaram et al., 2019), Bengali-English disfluent speech has received limited attention.

Collecting naturalistic disfluent code-mixed speech is costly and difficult, especially in micro-resource scenarios (approximately one hour of data). Large-scale end-to-end ASR models require hundreds of hours to generalize effectively (Hannun et al., 2014; Watanabe et al., 2018), rendering them unsuitable for cold-start settings. In contrast, Gaussian Mixture Model–Hidden Markov Model (GMM-HMM) systems remain practical and interpretable for severely data-constrained environments (Besacier et al., 2014; Mohri et al., 2008), and the Kaldi toolkit

Proceedings of WILDRE-8, pages 39–48

provides a robust experimental framework for such regimes (Povey et al., 2011). Recent advances in Large Language Models (LLMs) and multilingual neural text-to-speech (TTS) systems such as Indic Parler TTS (Lacombe et al., 2024; Lyth and King, 2024) offer an alternative pathway. Synthetic speech generation has been shown to support low-resource ASR research when real data collection is infeasible (Yeo et al., 2026). Rather than aiming to replace natural corpora, synthetic pipelines provide a controlled environment for modeling specific linguistic phenomena. Building on these advances, we construct **BEHE-CMDisfl** by generating disfluent Bengali–English and Hindi–English code-mixed utterances using structured LLM prompts using OpenAI's **ChatGPT**[1], followed by speech synthesis with **Indic Parler TTS**[2]. The corpus explicitly incorporates filled pauses, repetitions, and restarts within multilingual speech. Beyond resource creation, we evaluate its practical utility in a micro-resource setting by establishing GMM-HMM baselines on a 1.3-hour Bengali–English subset using Kaldi. This study therefore pursues two complementary goals: to release a disfluency-aware code-mixed speech corpus, and to examine its suitability for low-resource ASR experimentation.

1.1. Our Contribution:

- We develop **BEHE-CMDisfl**, a Bengali–English (BE) and Hindi–English (HE) speech dataset generated through a controlled LLM–TTS pipeline that explicitly integrates filled pauses, repetitions, and restarts within multilingual utterances.
- We document the prompt design strategy, linguistic constraints, and synthesis pipeline to ensure reproducibility and clarity in the data creation process.
- We establish reproducible GMM-HMM baselines (Monophone, Triphone, LDA-MLLT) for Bengali–English disfluent code-mixed ASR using Kaldi on a 1.3-hour subset.
- We demonstrate the impact of transliteration normalization and lexicon consistency on recognition performance, achieving a best WER of 37.74%.
- We show that disfluency markers are retained in decoding outputs when explicitly modeled in the lexicon, supporting their inclusion rather than removal in conversational ASR research.

[1] https://chatgpt.com/

[2] https://huggingface.co/ai4bharat/indic-parler-tts

2. Related Works

Research on code-mixed language processing has expanded in recent years, particularly for multilingual contexts such as South Asia. Early work primarily focused on textual datasets for *Hinglish*, *Banglish*, and *Tamil–English*, targeting sentiment analysis and sequence labeling tasks (Patra et al., 2018; Singh et al., 2018; Alam et al., 2024; Nishat Raihan et al., 2023). On the speech side, datasets such as *MUCS 2021* (Diwan et al., 2021) and *Prabhupadavani* (Sandhan et al., 2022) support multilingual ASR and speech translation. More recent efforts include *MediBeng* (Ghosh, 2025) and *Switchlingua* (Xie et al., 2025). While these resources advance code-switched speech research, they generally focus on fluent speech and do not explicitly incorporate disfluency phenomena. Disfluency research has been well established in English corpora such as *Switchboard* (Godfrey and Holliman, 1997) and *FluencyBank* (Romana et al., 2024). Studies on automatic disfluency detection and correction (Amann et al., 2024) and synthetic augmentation for Indic languages (Bhat et al., 2023) highlight the importance of modeling spontaneous irregularities. However, speech-level corpora that jointly represent code-mixing and disfluency for Indic languages remain limited (Saranya et al., 2025).

Recent advances in neural TTS and large language models enable scalable synthetic corpus creation. Indic Parler TTS (Lacombe et al., 2024; Lyth and King, 2024) supports multilingual speech generation, and synthetic pipelines have been used to support low-resource ASR research (Thai et al., 2019; Yeo et al., 2026). *MixFluent* (Paul et al., 2025) has successfully demonstrated the feasibility of generating synthetic code-mixed disfluent text using LLMs, it remains strictly limited to the textual modality. MixFluent provides valuable linguistic benchmarks for Bengali-English code-mixing, but its lack of acoustic realization renders it unsuitable for training end-to-end speech systems or modeling the prosodic features of disfluency. Our dataset, *BEHE-CMDisfl*, directly extends this foundational research by bridging the gap between text and speech. We not only adapt the text generation methodology for broader coverage incorporating both Bengali-English and Hindi-English pairs, but crucially, we project these textual disfluencies into the acoustic domain using the Indic Parler TTS framework.

In ASR research, Kaldi has been widely used for low-resource and code-switched settings (Kullmann, 2016; Yılmaz et al., 2016). Although deep models perform well in high-resource scenarios (Panayotov et al., 2015; Bu et al., 2017), their effectiveness in micro-resource conditions remains lim-

ited (Dar and Pushparaj, 2026; Dhasmana et al., 2026). Indian code-mixed ASR work has largely focused on Hindi–English datasets of moderate size (Pandey et al., 2018), leaving disfluent Bengali–English speech under-explored. Taken together, existing resources either emphasize code-mixing without disfluency, disfluency without multilingual speech, or text without audio. Our dataset fill this gap that integrates Bengali–English and Hindi–English code-mixing with explicit disfluency modeling, alongside baseline ASR experiments in a micro-resource setting.

3. Dataset Development Methodology

The *BEHE-CMDisfl* dataset was constructed through a two-stage synthetic pipeline that combines (i) *ChatGPT* based generation of code-mixed disfluent text and (ii) *Indic Parler TTS* synthesis to produce the corresponding speech. Bengali and Hindi are treated as matrix languages, with English as the embedded language. The design prioritizes controlled disfluency injection, realistic code-switching behavior, and reproducibility of the generation process.

3.1. Text Corpus Design and Preparation

To ensure a generalized domain, we curated reference text prompts covering some of the major conversational contexts:

- Daily life and informal communication (e.g., greetings, personal anecdotes, job oriented discussions etc.)
- Task-oriented dialogs (e.g., customer–service interactions)
- Education and classroom discussions among teachers and students
- Healthcare and teleconsultation contexts
- Social media–style opinions, commentary and discussions about sports and entertainment

This step is a way to explore the intersection of code-mixing and disfluency in bilingual speech and text, with a focus on understanding how LLMs handle code-mixed disfluent utterances. One of the primary objectives was to explore LLMs' ability to generate code-mixed disfluent sentences and to address the lack of high-quality code-mixed disfluent corpora, particularly for Indic languages.

3.2. LLM-Driven Text Generation

To generate real and natural code-mixed disfluent text, we used a LLM, specifically, ChatGPT (**GPT-5 model**) using zero shot and few shot prompting techniques. A typical prompt used to generate a BE-CM disfluent textual conversational data is shown below:

> You are a dialog generation assistant. Generate realistic, informal conversations between multiple speakers in a South Asian context (Bengali-English), incorporating the following instructions:
>
> 1. Speaker Roles and Identity
>
> - Each line should start with the speaker's name followed by a colon, e.g., "Arjun: ..."
> - Maintain consistent personality for each speaker:
>   - Speaker A: friendly, informal, slightly humorous
>   - Speaker B: polite, thoughtful, sometimes hesitant
> - Include a variety of speakers with different age, gender, and occupation backgrounds.
>
> 2. Code-Mixing Requirements
>
> - Blend Bengali with English naturally within sentences.
> - Code-mixing should occur at the word level, phrase level, or mid-sentence.
> - Ensure frequent switching for realism, but do not overuse English.
>
> 3. Disfluency Requirements
>
> - Insert natural disfluencies like:
>   - Filled pauses: "uh", "umm", "মানে", "আচ্ছা", "you know", "hmm", "err", etc.
>   - Repetitions: repeating words or phrases
>   - Hesitations: restarting sentences, partial words
> - Include disfluencies in roughly 15
> - Disfluencies should appear randomly but contextually plausible.
>
> 4. Conversation Context
>
> - Base dialogs on everyday scenarios:
>   - Catching up with friends

- Work or office discussions
- Academic or student-related interactions
- Social plans or small talk about sports, fils, tours, etc.

- Keep conversations casual and natural.

5. Output Formatting

- Format each utterance as: SpeakerName: <utterance>
- Avoid narration or meta-comments.
- Maintain proper punctuation.
- Use Unicode Bengali characters where applicable.
- Do not translate English words unnecessarily; retain natural code-mixing.

6. Length

- Each conversation should contain 15-25 lines of dialog.

Example Instruction:
"Generate a casual conversation between two friends, Kabir and Aditi, where they catch up about work and weekend plans. Include filled pauses and spontaneous repetitions. Use Bengali-English code-mixing naturally."

We used this structured prompt in both zero-shot and few-shot settings, adding examples to stabilize speaker consistency and turn-taking. Explicit constraints on roles, code-mixing, and disfluency show a balanced structural control with natural flow. We performed no manual rewriting beyond basic formatting checks (e.g., script consistency and structural validation). All generated dialogs are available in the repository[3]. A brief portion of the structural output of our prompting strategy is shown in the below example, showcasing how the model realizes conversational disfluencies within a code-mixed Bengali–English context.

> **Arjun:** Ohh, hi... hi অদিতি, umm কেমন আছো? মানে, এখানে দেখা হবে ভাবিনি, actually…
> **Aditi:** Hey, hey অর্জুন! আমি আছি umm, আমি ভালো- I mean, আমার ভালোই যাচ্ছে এখন. তুমি- err, you?
> **Arjun:** Oh, আমি? আমি তো... মানে, আমি ঠিকই আছি. কাজের চাপটা- uh, কাজের চাপ বেশি এখন তবে... but it is okay.
> **Aditi:** Exactly… exactly, বুঝি. আমার-ও... আমার-ও actually আগের week-এ deadline ছিল- ওই যে... um, sorry, কি বলছিলাম? হ্যাঁ, deadline-টা...একদম impossible লাগছিল কিন্তু হয়ে গেলো somehow.

3.3. Speech Synthesis with Indic Parler TTS

The generated dialogs were converted into speech using Indic Parler TTS developed by AI4Bharat[4], a multilingual neural TTS framework supporting multiple Indic languages. The model was selected due to its support for Indic scripts and its ability to process mixed-script inputs, which is essential for Bengali–English and Hindi–English code-mixed text. Speaker profiles were defined through textual descriptions and supplied to the TTS model to simulate inter-speaker variability across dialog turns. This allowed consistent voice characteristics within a conversation while preserving multi-speaker structure. For instance, a profile might be:

> **Example Speaker Profile (Aditi):** "A young adult female from West Bengal with a friendly, natural voice. Speaks in a casual tone typical of a Kolkata student. Bengali words sound native and informal, while English terms are clear with a Bengali accent. Incorporates natural pauses and fillers (*um*, *hmm*, *achha*) at switching boundaries to maintain a steady, expressive conversational rhythm."

Each generated text files (as described in the previous section) was processed into speech according to Algorithm 1. Utterances were segmented into chunks of at most 25 words prior to synthesis. This chunking strategy was adopted to improve waveform stability and prevent degradation in longer inputs. The synthesized segments were then concatenated to form a single audio file per dialog. No manual post-editing was performed on the generated waveforms beyond structural validation and file consistency checks. The generated speech data are available on this repository[5].

3.4. Statistical Analysis of the Dataset

In this section we discuss about the statistical analysis of both the textual data and the corresponding speech data and introduce some of the evaluation metrics used.

3.4.1. Evaluation Metrics

Some of the evaluation metrics used are:

[3] https://github.com/anonrpd/BEHE-CMDisf/tree/main/text_data

[4] https://ai4bharat.iitm.ac.in/

[5] https://github.com/anonrpd/BEHE-CMDisf/tree/main/speech_data

Algorithm 1 Text-to-Speech Generation for Disfluent Bengali-English/Hindi-English Code-Mixed dialogs

Require: `input_folder`: directory of dialog `.txt` files; `speaker_profiles`: speaker descriptions
Ensure: `output_folder`: generated `.wav` audio files

for each file F in `input_folder` **do**
  Initialize `audio_segments` list
  Read all lines from F
  for each line in F (up to MAX_LINES) **do**
    Parse line $\rightarrow$ `speaker_name`, `utterance_text`
    if `speaker_name` not in `speaker_profiles` **then**
      skip line
    end if
    `speaker_desc` = `speaker_profiles[speaker_name]`
    Split `utterance_text` into chunks ≤ 25 words
    for each chunk **do**
      Generate audio using `model.generate(tokenized_speaker_desc, tokenized_chunk)` $\rightarrow$ `audio_waveform`
      Append `audio_waveform` to `audio_segments`
    end for
  end for
  if `audio_segments` not empty **then**
    Concatenate `audio_segments` $\rightarrow$ `final_audio`
    Save `final_audio` as `.wav` in `output_folder`
  end if
end for

1. **Mean Filler Percentage (%)**: Proportion of tokens in an utterance that are filler words (e.g., *uh*, *um*, *ah*).
2. **Mean Repetition Percentage (%)**: Percentage of tokens that are repeated immediately or within a short span.
3. **Mean Restart Percentage (%)**: Fraction of utterances exhibiting restart phenomena
4. **Mean Code-Switch Percentage (%):** Ratio of words from the embedded (non-matrix) language to total words, reflecting the degree of code-switching.
5. **Speech Ratio:** Proportion of the total duration of the audio that contains active speech segments (excluding silence).
6. **Speech Ratio Range:** Difference between the maximum and minimum speech ratio across utterances in a corpus.

3.4.2. Text Data Analysis

The corpus includes 77 Bengali–English (BE-CM) and 40 Hindi–English (HE-CM) code mixed disfluent text dialog files, covering a range of conversational contexts. As shown in Table 1, Bengali–English dialogs contain an average of $\sim$13.6 words per utterance, while Hindi–English dialogs average $\sim$11.8 words. These values indicate moderately sized conversational turns across both subsets, with comparable structural organization.

Feature	BE–CM Text Data	HE–CM Text Data
Total Files	77	40
Mean Utterances per File	31.62	44.32
Mean Words/File	392.42	520.05
Mean Words per Utterances	13.59	11.77
Mean Filler %	0.25	0.24
Mean Repetition %	0.37	0.14
Mean Restart %	1.35	1.22
Mean CS %	30.94	34.46

Table 1: Summary statistics of the Bengali–English and Hindi–English disfluent code-mixed text corpora.

Disfluency markers are present at controlled levels throughout the corpus. Filled pauses account for approximately 0.24

Figure 1 illustrates distinct code-mixing behaviors between the datasets. The Bengali-English data exhibits high variability (4.68%-62.93%), suggesting context-dependent heterogeneity, whereas the Hindi-English dataset shows a narrower, more consistent range (22.44%-55.4%). This contrast highlights the differential nature of the generated text, likely reflecting the underlying distribution of the LLM's training data.

3.4.3. Speech Data Analysis

The Bengali–English code mixed (BE-CM) speech corpus comprises 77 recordings with mean duration of 195.8 s, while the Hindi–English code mixed (HE-CM) set contains 40 recordings with mean duration of 224.4 s. Table 2 clearly shows that both subsets exhibit high speech ratios (>0.90), indicating minimal silence and stable conversational flow. Average speech rates fall within 2.26–2.34 words

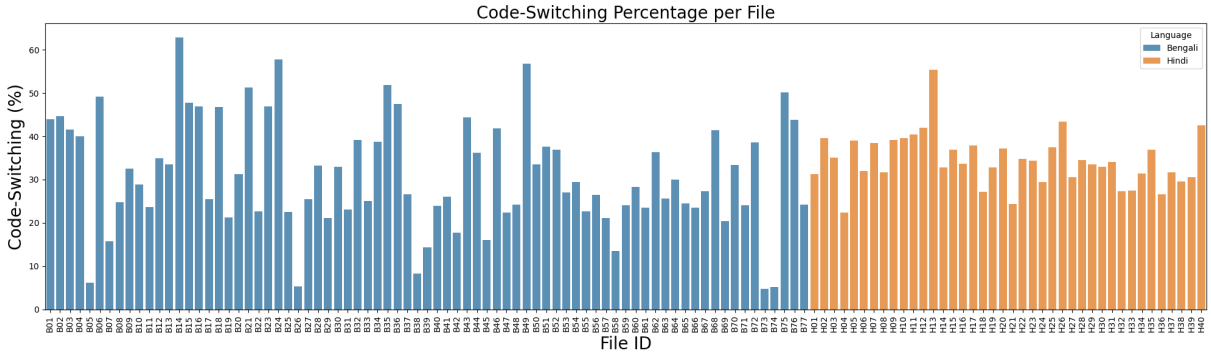

Figure 1: Code-Mixing Percentage Per Text File for Both the Languages.

per second, consistent with natural bilingual conversational tempo. The relatively low standard deviation in speech rate suggests uniform articulation speed across synthesized speakers.

Feature	BE-CM Speech Data	HE-CM Speech Data
Total recordings	77	40
Duration range (s)	108–304	136–329
Average duration (s)	195.8	224.4
Mean speech ratio	0.90	0.94
Speech ratio range	0.82–0.95	0.91–0.97
Estimated words per file	438.5 ± 110	525.4 ± 120
Speech rate (words/s)	2.26 ± 0.07	2.34 ± 0.05

Table 2: Comparative summary of Bengali-English and Hindi-English speech corpora.

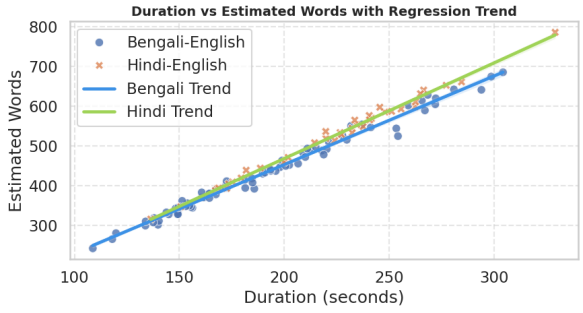

Figure 2: Relationship between speech duration and estimated word count for BE and HE datasets.

The plot shown in figure 2 shows the scatter plot of *Duration vs Estimated Words* that complements this by showing a consistent relationship between speech duration and word output for both the code-mixed disfluent datasets, with regression trends highlighting differences in speaking rate and verbosity. Together, these results indicate that our dataset maintains controlled disfluency injection at the textual level, stable acoustic realization at the speech level, and meaningful variation across the two code-mixed language pairs. This combination makes the dataset appropriate for low-resource ASR benchmarking and exploratory studies on disfluency-aware modeling.

4. Experimental Validation

To assess the efficacy of the BEHE-CMDisfl corpus in downstream tasks, we established a baseline ASR system using the Kaldi toolkit. Our objective was to determine if a model trained on this purely synthetic, disfluent data could successfully learn to decode code-mixed speech and, crucially, preserve disfluency markers rather than treating them as noise. We selected a 1.3 hour subset from the BE-CM disfluent data for this study to rigorously evaluate the pipeline's effectiveness under severe data constraints typical of cold-start scenarios.

4.1. Experimental Setup

We utilized a GMM-HMM pipeline, which remains a robust choice for micro-resource regimes where deep learning models often fail to generalize. The dataset was partitioned into a training set (12,006 words) and a held-out test set (1,236 words), maintaining an approximate 90/10 split to ensure rigorous evaluation. We employed a standard feature extraction pipeline (MFCCs + deltas + delta-deltas) and trained a progression of Monophone, Triphone, and LDA-MLLT models.

4.2. Results and Analysis

Table 3 summarizes the WER performance across different acoustic modeling stages.

Model Configuration	WER (%)
Monophone	51.14
Triphone (Tri1)	39.23
LDA-MLLT (Tri2b)	38.33
Wide-Beam Decoding (17.0)	**37.74**

Table 3: ASR performance (WER %) on the Bengali-English disfluent code-mixed subset using the Kaldi GMM-HMM baseline. The system achieves a best-case WER of 37.74% with wide-beam decoding.

The results indicate that the synthetic data possesses sufficient phonetic consistency to train a viable acoustic model from scratch. Notably, a simpler Triphone model achieved a WER of 39.23%, suggesting that for micro-resource synthetic data, lexicon consistency is often more critical than model complexity. The best performance (WER of 37.74%) was achieved by widening the decoding beam to account for the higher perplexity at code-switching points.

Achieving a WER of 37.74% is particularly significant given the micro-resource regime (1.3 hours) where deep learning benchmarks typically struggle to generalize. For context, recent studies on similar low-resource Indic languages using advanced Wav2Vec2 architectures reported WERs as high as 47.10% even with significantly more data (10 hours) (Dar and Pushparaj, 2026). Our results demonstrate that for extremely scarce, disfluent code-mixed data, a rigorously normalized GMM-HMM pipeline performs better than the theoretical baselines of deep neural models which are prone to severe overfitting in cold-start scenarios.

4.3. Disfluency Retention

A key objective was to evaluate if the model could capture synthetic disfluencies. Unlike standard systems that often filter filled pauses, our GMM-HMM baseline successfully retained these tokens. Table 4 highlights two examples where the model correctly transcribed filled pauses (*"um"*), repairs (*"i mean"*), and repetitions (*"oh oh"*). This confirms that the synthetic data provides sufficient spectral evidence to treat disfluencies as valid lexical units rather than background noise.

4.4. Impact of Normalization

We also observed that rigorous text normalization was essential. Initial experiments with raw, inconsistent transliterations (e.g., **'ar'** vs **'aar'** for the Bengali word **আর**) yielded a higher WER of 42.74%. Standardizing the lexicon improved this to 38.33%, highlighting that the quality of the synthetic text prompts is as important as the audio quality itself.

Case	Type	Transcript Sequence
1	Ref	*... **um i mean** ekta choto reunion*
	Hyp	*... **um i mean** ekta choto reunion*
	Status: Correctly identified repair & filled pause	
2	Ref	*acha **oh oh***
	Hyp	*acha **oh oh***
	Status: Correctly identified repetition	

Table 4: Successful recognition of disfluent markers in the BE-CM disfluent synthetic test set.

5. Applications and Uses

The BEHE-CMDisfl corpus is designed to address the scarcity of training data for conversational Indic speech technologies. Based on our experimental validation, we identify three primary applications:

- **Robust ASR Training in Low-Resource Regimes:** As demonstrated by our Kaldi baseline results, the corpus serves as a critical resource for bootstrapping ASR systems in micro-resource scenarios ($<$2 hours). It enables the training of acoustic models that do not merely treat disfluencies as noise but learn to transcribe them as valid lexical tokens (e.g., filled pauses, repetitions), which is essential for accurate transcriptions of spontaneous dialogue.

- **Disfluency Detection and Repair:** The corpus provides explicit, aligned examples of disfluent phenomena (restarts, hesitations) in code-mixed contexts. This structured data supports the development of supervised disfluency detection modules that can be integrated into end-to-end ASR pipelines to improve readability and downstream processing.

- **Conversational TTS Modeling:** The paired text–speech data, generated via the Indic Parler TTS, offers a controlled environment for studying prosodic modeling at code-switching points. Researchers can use this dataset to analyze and improve the naturalness of synthetic voices when navigating the complex prosody of bilingual sentence structures.

6. Conclusion and Future Work

We presented BEHE-CMDisfl, a synthetic, reproducible Bengali-English and Hindi-English code-mixed speech corpus with explicit annotations for disfluency phenomena. Our experimental validation using a Kaldi GMM-HMM baseline demonstrates that the dataset can successfully train disfluency-aware acoustic models, achieving a WER of 37.74% and correctly identifying filled pauses and repetitions for the BE-CM disfluent data. This confirms that synthetic data, when generated with consistent lexicons, is a viable stopgap for low-resource Indic speech research. While our synthetic pipeline effectively captures disfluency markers for ASR training, we view it as a targeted computational framework for modeling specific linguistic phenomena. We acknowledge that synthetic audio serves to approximate conversational patterns and may not fully replicate the complex acoustic variety of natural human speech. Future iterations of this study will include statistical significance testing and cross-validation across multiple training seeds to further verify the stability of our WER findings.

Our immediate road map focuses on validating and extending this resource:

- **Real-Speech Validation:** To evaluate the generalizability of our findings, we plan to curate a 'gold-standard' dataset consisting of 3-5 hours of natural conversational speech. This benchmark will allow us to formally quantify the 'Sim-to-Real' gap and assess how effectively our synthetic-trained models generalize to noisy, real-world environments.
- **Speech-to-Speech Translation:** We plan to explore using this disfluent data to train translation models that can normalize disfluent code-mixed speech into fluent English speech.
- **Broader Language Coverage:** The pipeline will be extended to other morphologically rich Indic languages, such as Tamil, Marathi, and Telugu, to test cross-lingual generalization. Also we will train and test on the Hindi-English disfluent code-mixed dataset as well.

7. Summary

7.1. Summary of Contributions

We present **BEHE-CMDisfl**, a synthetic, reproducible Bengali–English and Hindi–English code-mixed speech corpus with explicit annotations for disfluency phenomena and language segments. The corpus fills an important gap for multilingual, disfluent speech modeling in Indic contexts and supports ASR fine-tuning, disfluency detection, TTS research, and sociolinguistic studies.

7.2. Final Remarks

We believe **BEHE-CMDisfl** provides a practical, reproducible resource for the community while emphasizing transparency and ethical safeguards. Researchers using the corpus should validate findings on real conversational speech before deployment in safety-critical settings.

8. Ethical Considerations and Limitations

We recognize both the benefits and potential risks of generating synthetic, code-mixed, disfluent speech. Key ethical aspects and mitigations are summarized below:

Transparency: All generation steps—including LLM prompts, TTS model versions, and post-processing scripts—should be fully documented to ensure reproducibility and accountability.

Misuse Risks: Synthetic speech can be exploited for voice spoofing or misinformation. To mitigate such misuse, dataset releases must include clear usage guidelines, watermarking procedures, and, where possible, benchmarks for deepfake detection.

Bias and Representativeness: Both text generation and TTS systems inherit demographic and linguistic biases. We report speaker coverage, avoid mimicking real individuals, and caution against overgeneralization across underrepresented dialects.

Evaluation Limits: Synthetic disfluencies may not fully capture natural conversational patterns; results obtained from synthetic data should be validated on real-world speech to ensure generalization.

Privacy and Consent: As no real voices are used, privacy risks are minimal. However, the potential for misuse (e.g., impersonation) warrants controlled access and restrictive licensing.

References

Jordi Adell, Antonio Bonafonte, and David Escudero. 2006. Disfluent speech analysis and synthesis: A preliminary approach. In *Proc. 3rd Int. Conf. on Speech Prosody*, pages 1–4.

Sadia Alam, Md Farhan Ishmam, Navid Hasin Alvee, Md Shahnewaz Siddique, Md Azam Hossain, and Abu Raihan Mostofa Kamal. 2024. Bnsentmix: A diverse bengali-english code-mixed dataset for sentiment analysis. *arXiv preprint arXiv:2408.08964*.

Robin Amann, Zhaolin Li, Barbara Bruno, and Jan Niehues. 2024. Augmenting automatic speech recognition models with disfluency detection. pages 224–231.

Peter Auer. 2013. *Code-switching in conversation: Language, interaction and identity*. Routledge.

L. Besacier, E. Barnard, A. Karpov, and T. Schultz. 2014. Automatic speech recognition for under-resourced languages: A survey. *Speech Communication*, 56:85–100.

Vineet Bhat, Preethi Jyothi, and Pushpak Bhattacharyya. 2023. Adversarial training for low-resource disfluency correction. *arXiv preprint arXiv:2306.06384*.

Hui Bu, Jiayu Du, Xingyu Na, Bengu Wu, and Hao Zheng. 2017. Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline. In *Proc. O-COCOSDA*, pages 1–5.

M. A. Dar and J. Pushparaj. 2026. A wav2vec2 model-based automatic speech recognition system for low-resource kashmiri language. *International Journal of Speech Technology*, 29(1):2.

A. Dhasmana, A. Srivastava, and D. Chiang. 2026. Dialect matters: Cross-lingual asr transfer for low-resource indic language varieties. *arXiv preprint arXiv:2601.04373*.

Anuj Diwan, Rakesh Vaideeswaran, Sanket Shah, Ankita Singh, Srinivasa Raghavan, Shreya Khare, Vinit Unni, Saurabh Vyas, Akash Rajpuria, Chiranjeevi Yarra, et al. 2021. Multilingual and code-switching asr challenges for low resource indian languages. *arXiv preprint arXiv:2104.00235*.

Promila Ghosh. 2025. Medibeng (revision b05b594).

J Godfrey and E Holliman. 1997. Switchboard-1 release 2: Linguistic data consortium. *Switchboard: a user's manual*.

A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, et al. 2014. Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*.

T. D. Harya. 2018. Sociolinguistics: Code switching and code mixing. *Lentera: Jurnal Ilmiah Kependidikan*, 11(1):87–98.

Eunhee Kim. 2006. Reasons and motivations for code-mixing and code-switching. *Issues in EFL*, 4(1):43–61.

E. Kullmann. 2016. Speech-to-text for swedish using kaldi. Master's thesis, KTH Royal Institute of Technology.

Rohit Kundu, Preethi Jyothi, and Pushpak Bhattacharyya. 2022. Zero-shot disfluency detection for indian languages. In *Proceedings of the 29th international conference on computational linguistics*, pages 4442–4454.

Y. Lacombe, V. Srivastav, and S. Gandhi. 2024. Parler-tts. GitHub repository.

D. Lyth and S. King. 2024. Natural language guidance of high-fidelity text-to-speech with synthetic annotations. *arXiv preprint arXiv:2402.01912*.

M. Mohri, F. Pereira, and M. Riley. 2008. Speech recognition with weighted finite-state transducers. In *Springer Handbook of Speech Processing*, pages 559–584. Springer.

Melissa G Moyer. 2002. Bilingual speech: A typology of code-mixing.

Md Nishat Raihan, Dhiman Goswami, Antara Mahmud, Antonios Anastasopoulos, and Marcos Zampieri. 2023. Sentmix-3l: A bangla-english-hindi code-mixed dataset for sentiment analysis. *arXiv e-prints*, pages arXiv–2310.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An asr corpus based on public domain audio books. In *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, pages 5206–5210.

A. Pandey, B. M. L. Srivastava, and S. Sitaram. 2018. Adapting monolingual resources for code-mixed hindi-english speech recognition. In *Proc. IEEE ICASSP*.

Braja Gopal Patra, Dipankar Das, and Amitava Das. 2018. Sentiment analysis of code-mixed indian languages: An overview of sail_code-mixed shared task@ icon-2017. *arXiv preprint arXiv:1803.06745*.

Aryan Paul, Tapabrata Mondal, Dipankar Das, and Sivaji Bandyopadhyay. 2025. Generating and analyzing disfluency in a code-mixed setting. *Recent Advances in Natural Language Processing (RANLP)*, pages 915–924.

D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, et al. 2011. The kaldi speech recognition toolkit. In *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*.

Amrit Romana, Minxue Niu, Matthew Perez, and Emily Mower Provost. 2024. Fluencybank timestamped: An updated data set for disfluency detection and automatic intended speech recognition. *Journal of Speech, Language, and Hearing Research*, 67(11):4203–4215.

Jivnesh Sandhan, Ayush Daksh, Om Adideva Paranjay, Laxmidhar Behera, and Pawan Goyal. 2022. Prabhupadavani: A code-mixed speech translation data for 25 languages. *arXiv preprint arXiv:2201.11391*.

S Saranya, B Bharathi, S Gomathy Dhanya, and Aishwarya Krishnakumar. 2025. Real-time continuous tamil dialect speech recognition and summarization. *Circuits, Systems, and Signal Processing*, 44(4):2855–2881.

Elizabeth Shriberg. 2001. To 'errrr'is human: ecology and acoustics of speech disfluencies. *Journal of the international phonetic association*, 31(1):153–169.

Vinay Singh, Deepanshu Vijay, Syed Sarfaraz Akhtar, and Manish Shrivastava. 2018. Named entity recognition for hindi-english code-mixed social media text. In *Proceedings of the seventh named entities workshop*, pages 27–35.

S. Sitaram, K. R. Chandu, S. K. Rallabandi, and A. W. Black. 2019. A survey of code-switched speech and language processing. *arXiv preprint arXiv:1904.00784*.

Bao Thai, Robert Jimerson, Dominic Arcoraci, Emily Prud'hommeaux, and Raymond Ptucha. 2019. Synthetic data augmentation for improving low-resource asr. In *2019 IEEE Western New York Image and Signal Processing Workshop (WNYISPW)*, pages 1–9.

S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, et al. 2018. Espnet: End-to-end speech processing toolkit. In *Proc. Interspeech*.

Peng Xie, Xingyuan Liu, Tsz Wai Chan, Yequan Bie, Yangqiu Song, Yang Wang, Hao Chen, and Kani Chen. 2025. Switchlingua: The first large-scale multilingual and multi-ethnic code-switching dataset. *arXiv preprint arXiv:2506.00087*.

Y. H. Yeo, Y. Hu, S. Gopal, Y. Peng, H. Liu, and E. S. Chng. 2026. Improving code-switching speech recognition with tts data augmentation. *arXiv preprint arXiv:2601.00935*.

Emre Yılmaz, M. Andringa, S. Kingma, J. Dijkstra, F. van der Kuip, H. van de Velde, et al. 2016. Longitudinal speaker clustering and identification for frisian-dutch code-switching. In *Proc. Interspeech*, pages 3668–3672.

Vicky Zayats, Mari Ostendorf, and Hannaneh Hajishirzi. 2016. Disfluency detection using a bidirectional lstm. *arXiv preprint arXiv:1604.03209*.

Konkani Wordnet Resources

Hanumant Redkar[1], Mahadev Gawas[2], Anjali Desai[1], Jyoti Pawar[1]
[1]Discipline of Computer Science and Technology, Goa Business School, Goa University, India
[2]State Higher Education Council, Directorate of Higher Education, Govt. of Goa, India
[1]hanumantredkar@unigoa.ac.in, [2]gawas-dhe.goa@gov.in, [1]desai.anjali.ad@gmail.com, [1]jdp@unigoa.ac.in

Abstract

Konkani is a low-resource Indo-Aryan language spoken along the western coast of India, characterized by significant dialectal variation, multi-script usage, and limited standardized computational resources. This paper presents a consolidated and analysis-ready lexical resource derived from the Konkani Wordnet, built under the IndoWordNet framework. The resource comprises **32,370** synsets, 37,719 unique lexical entries, 32,370 glosses, and 33,318 example sentences, enriched with pronunciations, semantic relations, and illustrative examples. We describe the systematic extraction, normalization, and structural integration of wordnet data, resolving identifier inconsistencies and ensuring semantic coherence across distributed lexical files. To demonstrate the practical utility of this resource, we present an API-based bilingual vocabulary exercise generation system that leverages shared synset identifiers to automatically produce semantically aligned Hindi–Konkani word pairs for e-learning applications. The resulting resource enhances accessibility, reproducibility, and computational readiness for NLP tasks, while providing a foundational infrastructure for developing technology-driven teaching and learning tools for Konkani.

Konkani, Language Resources, Konkani Wordnet, Wordnet, IndoWordNet, Low-Resource Language, Synset, Lexical Resources, Corpus, KWN

1. Introduction

The expansion of digital learning platforms has increased the need for scalable tools that support multilingual education, particularly for Indian languages where structured learning resources are limited. Vocabulary acquisition remains central to language learning, and interactive formats such as match-the-pairs exercises are commonly used to reinforce word associations (Redkar et al., 2018).

WordNet provides a structured lexical framework in which words are organized into synsets representing language-independent concepts (Miller et al., 1990; Fellbaum, 1998). IndoWordNet (Bhattacharyya, 2010) extends this model to multiple Indian languages, including Hindi and Konkani, using a shared synset structure (Walawalikar et al., 2012; Kashyap et al., 2016). This shared conceptual layer enables cross-lingual alignment based on semantic equivalence rather than direct translation (Sarma et al., 2012; Dash et al., 2017).

This paper presents a consolidated Konkani Wordnet resource and demonstrates its application in automatically generating bilingual Hindi–Konkani match-the-pairs vocabulary exercises. The approach uses synset identifiers as a pivot: words from both languages belonging to the same synset are paired to form semantically aligned vocabulary questions. The system is implemented as an API-based module that retrieves synset–word mappings, performs cross-lingual synset matching, and generates word pairs for use in e-learning applications (Redkar et al., 2017b,a, 2018).

1.1. Contributions

The primary contributions of this work are as follows:

- **Dataset Extraction and Normalization:** A systematic consolidation of the Konkani Wordnet comprising **32,370** synsets, 37,719 unique lexical entries, 32,370 glosses, and 33,318 example sentences. The resource resolves structural redundancies, harmonizes identifier mappings, and ensures semantic consistency across distributed lexical files, making it analysis-ready for NLP tasks.
- **Cross-lingual Pairing System:** An API-based bilingual exercise generation system that uses shared IndoWordNet synset identifiers to automatically produce semantically aligned Hindi–Konkani word pairs without relying on direct translation dictionaries.
- **Educational Application Demonstration:** A working demonstration of how structured lexical databases can be operationalized into scalable, concept-based vocabulary generation systems suitable for integration with e-learning platforms and language learning applications.

2. Literature Review

Work on building lexical and computational resources for Konkani has moved in steps — first laying down basic word lists, then expanding them using real language data, creating visualization tools,

Proceedings of WILDRE-8, pages 49–54

adding new features like audio, and testing these resources in actual applications. Konkani, being a low-resource language with multiple dialects and scripts, presents challenges in standardizing vocabulary, maintaining consistency, and ensuring interoperability. Below we trace the development of Konkani WordNet within the broader context of Indian language wordnet development.

2.1. Princeton WordNet and the WordNet Model

The foundation of all modern wordnet development was laid by Miller et al. (1990) with the creation of Princeton WordNet for English. WordNet organizes lexical knowledge into synsets — groups of synonymous words representing a single underlying concept — connected through semantic relations such as hypernymy, hyponymy, meronymy, and antonymy (Fellbaum, 1998). This model proved highly influential, inspiring the development of wordnets for dozens of languages worldwide. The core insight — that meaning is best captured through conceptual groupings rather than individual word definitions — remains the guiding principle for all IndoWordNet development.

2.2. IndoWordNet and Indian Language Wordnets

Bhattacharyya (2010) introduced IndoWordNet as a multilingual lexical database linking wordnets of major Indian languages under a shared synset framework. The shared synset structure means that languages do not need direct translation dictionaries — instead, words from different languages are linked through a common concept identifier, enabling cross-lingual alignment based on semantic equivalence (Sarma et al., 2012; Dash et al., 2017).

Hindi WordNet, developed at IIT Bombay, is the most mature and largest Indian language wordnet and serves as the pivot language for IndoWordNet development (Kashyap et al., 2016). It has been used extensively for educational applications, including the Hindi Shabdamitra tool for vocabulary learning (Redkar et al., 2017b,a, 2018). Other Indian languages represented in IndoWordNet include Bengali, Marathi, Telugu, Tamil, Gujarati, Punjabi, and Assamese, each developed using either the expansion approach — translating from Hindi synsets — or the merge approach — building independently then linking. Sanskrit WordNet has also been developed and applied for educational purposes (Kulkarni et al., 2019). The breadth of IndoWordNet demonstrates both the scalability of the shared synset model and the growing importance of structured lexical resources for Indian NLP.

2.3. Building Konkani WordNet

Walawalikar et al. (2010) initiated Konkani WordNet[1] development using the expansion approach pioneered for Hindi WordNet at IIT Bombay. They systematically translated and adapted existing synsets into Konkani, producing around 1,969 core synsets. This process required careful handling of Konkani's dialectal diversity — including Goan and Saraswat varieties — multi-script usage across Devanagari, Roman, and Kannada scripts, and numerous vocabulary items without straightforward equivalents in other languages. Their work established Konkani's place within the broader IndoWordNet network and enabled data sharing with other Indian language wordnets (Desai et al., 2010, 2016). Further, Konkani Wordnet Dictionary, KWN-Dict, has been developed as one of the applications of Konkani Wordnet by (Redkar et al., 2026).

2.4. Surveying Konkani NLP Resources

Rajan et al. (2020) conducted a comprehensive survey of computational resources for Konkani, cataloguing tools such as the ILCI parallel corpus (approximately 50,000 sentences), POS taggers, morphological analyzers, and early speech datasets. Konkani WordNet was identified as a key resource, though the survey highlighted persistent gaps — small corpora, insufficient annotated data, few benchmarks, and limited integration between existing tools. The authors called for dataset standardization and stronger connections between lexical databases and real NLP applications.

2.5. Corpus-Based Enhancement

Manerkar et al. (2022) moved beyond simple synset expansion by using the Shabdarth crowdsourcing platform to identify gaps through real language data analysis. They identified 572 missing words and added 71 new synsets, improving coverage in certain domains by up to 27%. This work marked a methodological shift — from top-down list expansion to bottom-up, data-driven enrichment — and demonstrated the value of community participation in resource building for low-resource languages.

2.6. Visualization and Learning Tools

As the Konkani WordNet database grew beyond 32,000 synsets, concerns around usability and accessibility emerged. Gawde et al. (2024a) addressed this by developing a tree-based visualizer that allows users to explore semantic relationships such as hypernyms and hyponyms interac-

[1] https://konkaniwordnet.unigoa.ac.in/

tively. This tool improved the transparency of the resource, supported concept-based teaching, and helped identify structural inconsistencies within the network.

2.7. Multimodal Enrichment

The Shabdocchar project (Gawde et al., 2024b) extended Konkani WordNet beyond text by adding audio pronunciation recordings through a gamified crowdsourcing mechanism. This multimodal enrichment is particularly significant for Konkani where dialectal pronunciation differences are substantial. The audio data supports development of speech recognition and text-to-speech systems, pushing Konkani WordNet toward becoming a truly multimodal lexical resource.

2.8. Application-Focused Studies

Ghosarwadkar et al. (2024) demonstrated an NLP application using zero-shot transfer learning from Marathi for Konkani sentiment analysis, exploiting the linguistic proximity between the two languages to compensate for limited Konkani training data. Their work underscores the importance of structured lexical resources as a foundation for downstream NLP tasks and highlights the broader potential of cross-lingual transfer for low-resource Indian languages.

3. Statistical Analysis of the Consolidated Resource

3.1. Core Resource Statistics

The consolidated Konkani WordNet resource comprises **32,370** synsets and 37,719 unique lexical entries. In addition, the dataset contains 32,370 glosses and 33,318 example sentences. These figures reflect the scale of the resource following normalization, identifier harmonization, and structural integration across distributed lexical files.

Component	Count
Synsets	32,370
Unique Lexical Entries	37,719
Glosses	32,370
Example Sentences	33,318

Table 1: Statistics of Konkani Wordnet

3.2. Derived Quantitative Metrics

To assess semantic coverage and structural density, quantitative metrics were computed relative to the total number of synsets.

Gloss Coverage:

$$\frac{32{,}370}{32{,}370} \times 100 = 100\% \tag{1}$$

Every synset contains a definitional description, ensuring complete semantic coverage.

Example Coverage:

$$\frac{33{,}318}{32{,}370} \times 100 = 102.9\% \tag{2}$$

Some synsets contain more than one example sentence, resulting in a coverage ratio exceeding 100%.

Lexical Density:

$$\frac{37{,}719}{32{,}370} = 1.17 \tag{3}$$

This ratio reflects the presence of polysemy and shared lexical realizations across conceptual nodes.

Average Words per Synset:

$$\frac{55{,}530}{32{,}370} \approx 1.72 \tag{4}$$

Each synset contains 1.73 lexical items on average, indicating a semantically rich but compact lexical organization.

3.3. Sample Synset Entries

To illustrate the structure and richness of the consolidated resource, Table 2 presents representative synset entries from the Konkani Wordnet across different parts of speech.

These examples demonstrate several key properties of the resource. First, multiple synonymous words are grouped under a single synset, capturing the synonymy relation. Second, glosses provide definitional clarity in natural language. Third, example sentences ground each concept in authentic Konkani usage, supporting both language learners and NLP researchers. The structured synset representation further enables bilingual applications such as the exercise generation system described in Section 4.

4. Konkani Wordnet Resources

The Konkani WordNet used in this work is part of the larger IndoWordNet initiative, which has taken inspiration from the Princeton WordNet model by organizing lexical knowledge into synsets representing language-independent concepts. Each synset consists of a set of synonymous words and is linked to other synsets through semantic and lexical relations such as hypernymy, hyponymy, and meronymy. In the database schema, lexical

Synset ID	POS	Konkani Words	Gloss	Example Sentence
99	Noun	गिन्यान, ज्ञान (*ginyaana, nyaana*) - knowledge	मनाक वा विचारांक जाता अशी वस्तूंची वा विशयांची माहिती (*manaaka vaa vichaaraanka jaataa ashee vastoonchee vaa vishayaanchee maahitee*)	ताका संस्कृताचें बरें गिन्यान आसा (*taakaa sanskrutaacheM bareM ginyaana aasaa*)
86	Adjective	हजर, उपस्थीत (*hajara, upastheeta*) - present	लागीं बशिल्लें, मुखार वा लागीं आयिल्लें (*laageeM bashilleM, mukhaar vaa laageeM aayilleM*)	आज वर्गांत हजर आशिल्ले विद्यार्थी उणें आसले (*aaja vargaanta hajara aashille vidhyaarthee uNeM aasale*)
2078	Verb	पियेवप, पिवप, घोंटप (*piyevapa, pivapa, ghoMTapa*) - to drink	पातळ द्रव, रोस, उदक, दूद, बी पदार्थ पोटांत घालप (*paataLa drava, raosa, udaka, dooda, bee padaartha poTaanta ghaalapa*)	उदक पियेवप भलायकेक बरें (*udaka piyevapa bhalaayakeka bareM*)
123	Adverb	सारकें, प्रमाणें (*saarakeM, pramaaNeM*) - like that	कोणायच्या मतान वा नदरेन (*koNaayachyaa mataana vaa nadarena*)	तो म्हजे सारकें काम करपाक सोदिना (*to mhaje saarakeM kaama karapaaka sodinaa*)
28106	Adjective	बरें, उत्कृश्ट, उत्तम (*bareM, utkrushTa, uttama*)	जो बरे तरेन परिणामाच्या रुपान आसा वा येता असो (*jo bare tarena pariNaamaachyaa rupaana aasaa vaa yetaa aso*)	हांव थंय नाशिल्लों ही बरी गजाल (*haaMva thaMya naashilloM hee bari gajaala*)

Table 2: Sample synset entries from Konkani Wordnet with Part-of-Speech information

items are stored in the `wn_word` table, while the mapping between words and concepts is maintained in the `wn_synset_words` table through the `synset_id` field (Redkar et al., 2015, 2016).

The Konkani WordNet provides structured lexical coverage across multiple parts of speech, including nouns, verbs, adjectives, and adverbs. Each word entry is associated with a unique `word_id`, and multiple words may be linked to the same `synset_id`, representing synonymy at the conceptual level.

The system operates on two IndoWordNet-compliant lexical databases, `wordnet_hindi` and `wordnet_konkani`, which follow a standardized relational schema. Both databases contain two core tables:

- `wn_word (word_id, word)` — stores lexical entries in the respective language.
- `wn_synset_words (synset_id, word_id)` — maps each word to one or more synsets representing concepts.

Cross-lingual alignment is performed by identifying synset identifiers present in both databases. These shared synsets form the semantic bridge between Hindi and Konkani. For any shared synset s, where H_s represents the set of Hindi words and K_s represents the set of Konkani words linked to s, the number of potential bilingual word pairs derived from that synset is:

$$|H_s \times K_s| \tag{5}$$

Thus, the total number of automatically generable pairs depends on the lexical coverage of the two WordNets and the extent of their synset overlap. The system is designed to operate on any dataset conforming to this schema, making it independent of specific corpus sizes and suitable for future expansion as WordNet resources grow.

5. Discussion

The proposed system highlights the usefulness of lexical knowledge bases for educational technology. By leveraging synset identifiers as language-independent concept markers, the system avoids the limitations of direct translation dictionaries, where one-to-one word mappings may ignore pol-

ysemy and contextual meaning. Instead, words are paired based on shared conceptual membership, making the learning process semantically grounded.

The modular architecture allows the system to function purely as a backend question generator, independent of any specific user interface. This separation of concerns enables integration with different e-learning platforms, learning management systems, or mobile applications.

5.1. Replicability to Other Indian Languages

A significant strength of the proposed system is its replicability across other IndoWordNet language pairs. Since all languages within the IndoWordNet framework share a common synset identifier structure, the bilingual pairing mechanism described in this work can be extended to any two IndoWordNet languages with minimal modification. For example, the same system architecture could generate vocabulary exercises for Hindi–Bengali, Hindi–Marathi, Hindi–Telugu, or Konkani–Marathi pairs simply by substituting the corresponding WordNet databases, provided they conform to the standard IndoWordNet relational schema comprising the `wn_word` and `wn_synset_words` tables.

This cross-language scalability positions the system not merely as a Konkani-specific tool but as a general-purpose vocabulary exercise generator for Indian language e-learning. As IndoWordNet resources continue to grow and improve in coverage and quality, the number of valid bilingual pairs generated by the system will increase proportionally, making the approach increasingly powerful over time. Future work may explore multi-way alignment — generating exercises across three or more languages simultaneously — further extending the educational utility of structured lexical resources in Indian language contexts.

5.2. Limitations

The system's effectiveness depends on the lexical coverage and quality of the underlying WordNet resources. Incomplete synset mappings or uneven vocabulary distribution across languages may limit the number of valid bilingual pairs. Additionally, while synset-level alignment ensures semantic equivalence at a conceptual level, differences in usage frequency, register, or regional variation are not explicitly addressed.

6. Conclusion

This paper presented a consolidated Konkani Wordnet resource and demonstrated its application in automatically generating bilingual Hindi–Konkani match-the-pairs vocabulary exercises. By using shared IndoWordNet synset identifiers as the basis for cross-lingual alignment, the system produces semantically equivalent word pairs suitable for vocabulary learning in e-learning environments. The architecture separates configuration, database access, and service logic, resulting in a modular and extensible API-based solution.

The work demonstrates how structured lexical databases can be transformed into practical educational tools that support concept-based language learning. The approach is replicable across all IndoWordNet language pairs, making it a scalable solution for Indian language vocabulary learning at large. Future extensions may include integration of additional IndoWordNet languages, incorporation of difficulty levels based on word frequency or part of speech, and the use of glosses or example sentences to further enrich learning content.

References

Pushpak Bhattacharyya. 2010. IndoWordNet. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Malta.

Niladri Sekhar Dash, Pushpak Bhattacharyya, and Jyoti D. Pawar, editors. 2017. *The WordNet in Indian Languages*. Springer Singapore.

Shilpa N. Desai, Ramdas N. Karmali, Shantaram W. Walawalikar, and Damodar Ghanekar. 2010. Tools for IndoWordNet development. In *Proceedings of the International Conference on Natural Language Processing*.

Shilpa N. Desai, Shantaram W. Walawalikar, Ramdas N. Karmali, and Jyoti D. Pawar. 2016. Insights on the Konkani WordNet development process. In *The WordNet in Indian Languages*, pages 101–117. Springer Singapore.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Sunayana R. Gawde et al. 2024a. Konkani WordNet visualizer as a concept teaching-learning tool. In *Proceedings of the 21st International Conference on Natural Language Processing (ICON)*.

Sunayana R. Gawde et al. 2024b. Shabdocchar: Konkani WordNet enrichment with audio feature. In *Proceedings of the 21st International Conference on Natural Language Processing (ICON)*.

Rohit M. Ghosarwadkar et al. 2024. Sentiment analysis for Konkani using zero-shot Marathi trained neural network model. In *Proceedings*

of the 21st International Conference on Natural Language Processing (ICON).

Laxmi Kashyap, Salil Rajeev Joshi, and Pushpak Bhattacharyya. 2016. Insights on Hindi WordNet coming from the IndoWordNet. In *The WordNet in Indian Languages*, pages 19–44. Springer Singapore.

Malhar Kulkarni, Nilesh Joshi, Sayali Khare, Hanumant Redkar, and Pushpak Bhattacharyya. 2019. Introduction to sanskrit shabdamitra: An educational application of sanskrit wordnet. In *Proceedings of the 6th International Sanskrit Computational Linguistics Symposium*, pages 117–133.

Sanjana Manerkar, Kavita Asnani, Preeti Ravindranath Khorjuvenkar, Shilpa Desai, and Jyoti D. Pawar. 2022. Konkani WordNet: Corpus-based enhancement using crowdsourcing. *Transactions on Asian and Low-Resource Language Information Processing*, 21(4):1–18.

George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. 1990. Introduction to WordNet: An on-line lexical database. volume 3, pages 235–244.

Annie Rajan, Ambuja Salgaonkar, and Ramprasad Joshi. 2020. A survey of Konkani NLP resources. *Computer Science Review*, 38:100299.

Hanumant Redkar, Sudha Bhingardive, Kevin Patel, Pushpak Bhattacharyya, Neha Prabhugaonkar, Apurva Nagvenkar, and Ramdas Karmali. 2016. WWDS APIs: Application programming interfaces for efficient manipulation of world wordnet database structure. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.

Hanumant Redkar, Mahadev Gawas, and Sherwin Pereira. 2026. Kwn-dict: An online konkani dictionary based on konkani wordnet. In *Proceedings of Research Libraries: A Source to Adapt to Digital Form Using the Indian Knowledge Systems (IKS) in Promoting Social Science Research*, pages 110–116.

Hanumant Redkar, Nilesh Joshi, Sayali Khare, Lata Popale, Malhar Kulkarni, and Pushpak Bhattacharyya. 2017a. Hindi shabdamitra: A wordnet based tool for enhancing teaching-learning process. In *Proceedings of the 14th International Conference on Natural Language Processing (ICON 2017)*.

Hanumant Redkar, Rajita Shukla, Sandhya Singh, Jaya Saraswati, Laxmi Kashyap, Diptesh Kanojia, Preethi Jyothi, Malhar Kulkarni, and Pushpak Bhattacharyya. 2018. Hindi WordNet for language teaching: Experiences and lessons learnt. In *Proceedings of the 9th Global Wordnet Conference*, pages 314–323.

Hanumant Redkar, Sandhya Singh, Nilesh Joshi, Anupam Ghosh, and Pushpak Bhattacharyya. 2015. IndoWordNet dictionary: An online multilingual dictionary using IndoWordNet. In *Proceedings of the 12th International Conference on Natural Language Processing*, pages 71–78.

Hanumant Redkar, Sandhya Singh, Meenakshi Somasundaram, Dhara Gorasia, Malhar Kulkarni, and Pushpak Bhattacharyya. 2017b. Hindi shabdamitra: A wordnet based e-learning tool for language learning and teaching. In *Proceedings of the 4th Workshop on Natural Language Processing Techniques for Educational Applications (NLPTEA 2017)*, pages 23–28, Taipei, Taiwan.

Shikhar Kr Sarma, Dibyajyoti Sarmah, Biswajit Brahma, Himadri Bharali, Mayashree Mahanta, and Utpal Saikia. 2012. Building multilingual lexical resources using wordnets: Structure, design and implementation. In *Proceedings of the 3rd Workshop on Cognitive Aspects of the Lexicon*, pages 161–170.

Shantaram Walawalikar, Shilpa Desai, Ramdas Karmali, Sushant Naik, Damodar Ghanekar, Chandralekha D'Souza, and Jyoti Pawar. 2010. Experiences in building the Konkani WordNet using the expansion approach. In *Proceedings of the 5th Global WordNet Conference*, Mumbai, India. Narosa Publishing House.

Shantaram Walawalikar, Shilpa Desai, Ramdas Karmali, Sushant Naik, Damodar Ghanekar, Chandralekha D'Souza, and Jyoti Pawar. 2012. Experiences in building the Konkani WordNet using the expansion approach. In *Proceedings of the 5th Global WordNet Conference*. Narosa Publishing House.

Development of Speech Corpus for Low-Resource Language- A case of Sanskrit

Devendr Kumar[1], Girish Nath Jha[2], Khalid Choukri[3]

Department of Language Technology and Language Engineering, Mahatma Gandhi Antarrashtriya Hindi Vishwavidyalaya, Wardha, Maharashtra[1]
School of Sanskrit and Indic Studies, Jawaharlal Nehru University, New Delhi[2]
European Language Resource Association, Paris[3]
devneed2@gmail.com[1], girishjha@gmail.com[2], choukri@elda.org[3]

Abstract

This paper presents a comprehensive framework for the development of a speech corpus for Sanskrit, designed to facilitate advances in Automatic Speech Recognition (ASR) and AI/ML research. The proposed corpus comprises over 107 hours of transcribed speech data, collected from diverse Sanskrit sources through a systematic and scalable pipeline. We detail the end-to-end methodology adopted for corpus creation, encompassing web crawling, data sanitization, audio downloading, and transcription alignment. Particular emphasis is placed on the methodological rigor applied at each stage, including source selection, preprocessing for quality assurance, transcription protocols, and forced alignment techniques. The paper further addresses the unique complexities inherent to Sanskrit, spanning its phonetic richness, intricate morphological structure, and distinctive syntactic patterns. By systematically addressing these dimensions, the resulting 107-hour corpus aims to serve as a foundational resource for speech technology research in Sanskrit.

Keywords: Sanskrit Speech Corpus, Resource creation for Sanskrit, Sanskrit speech recognition, Sanskrit Data, Sanskrit Text, aligned data.

1. Introduction

Sanskrit, one of the oldest and most grammatically sophisticated languages of the Indo-Aryan family, occupies a position of profound cultural, religious, and linguistic significance in the Indian subcontinent. Recognized as one of the 22 scheduled languages under the Eighth Schedule of the Indian Constitution, Sanskrit has served for millennia as the medium of philosophy, science, literature, and religious thought. Its grammatical framework, codified by the legendary grammarian Pāṇini in the 7th century BCE, remains unparalleled in its precision, drawing admiration from modern linguists and computer scientists alike.

Despite its rich legacy, Sanskrit today faces the paradox of reverence without widespread spoken use. Its everyday spoken application has declined significantly over the centuries, rendering it predominantly a written and liturgical language. This limited oral presence has hindered the development of modern speech technologies for Sanskrit, making the compilation of large-scale spoken language data a persistent challenge.

With the rapid advancement of Natural Language Processing (NLP) and AI/ML-driven speech technologies, languages such as English, Mandarin, and Hindi now benefit from robust speech corpora powering virtual assistants, transcription tools, and language learning platforms.

Proceedings of WILDRE-8, pages 55–60

However, Sanskrit remains largely underrepresented in this digital revolution. To address this gap, the present work describes the systematic development of a 107-hour Sanskrit speech corpus, compiled from diverse online sources through a pipeline involving web crawling, data sanitization, audio processing, transcription, and forced alignment. The corpus aims to facilitate advances in Automatic Speech Recognition (ASR), Text-To-Speech (TTS) synthesis, and broader linguistic research, while contributing to the digital preservation of the Sanskrit language.

The remainder of this paper is organized as follows: Section 2 reviews related work, Section 3 describes the data collection methodology, Section 4 details the preprocessing and alignment pipeline, Section 5 presents corpus statistics, Section 6 discusses challenges, and Section 7 concludes with future directions.

2. Related Work

There have been efforts to create Sanskrit speech corpora, and some related work has been done. The **AI4Bharat (**Javed, et. al. 2024) program has been started at IIT Madras for creating language resources for various Indian languages, including Sanskrit. They have worked on building a vertical corpus for Sanskrit, which includes both text and speech data. Sanskrit and other 11 Indian languages are represented in the **Shrutilipi** (Chadha, et. al. 2022) tagged ASR dataset which was created by mining parallel audio and text pairings at the document scale from All India Radio news bulletins. 27 hours (Gupta, et. al. 2021) of tagged speech data is available for Sanskrit. The **Vāksañcayaḥ** (Adiga, et. al. 2021) (Holla, et. al. 2022). speech corpus contains 45,953 sentences that were recorded at a sampling rate of 22 KHz over the course of 78 hours of data. Readings from diverse texts from Sanskrit literature make up the majority of the content.

3. Text Selection

Sanskrit encompasses a diverse literary tradition, spanning a vast range of genres, styles, and historical periods. At the broadest level, Sanskrit literature is categorized into two major forms: Vedic and Classical.

Vedic Sanskrit represents the earliest stratum of the language and comprises the four foundational Vedas- the Rigveda, Yajurveda, Samaveda, and Atharvaveda along with associated texts such as the Brahmanas, Aranyakas, and Upanishads. This body of literature is primarily composed of hymns, ritual prescriptions, and philosophical discourses, and stands as one of the oldest recorded literary traditions in human history.

Classical Sanskrit, on the other hand, is broadly divided into two principal literary forms: Prose and Poetry. For the purpose of the present corpus, Prose Sanskrit also commonly referred to as spoken Sanskrit was selected as the primary source of data. This decision was informed by the fact that Sanskrit Automatic Speech Recognition (ASR) research is still in its nascent stages, and spoken or prose-form text offers a more practical and representative foundation for building speech datasets.

In terms of data sourcing, openly available texts (Kumar, et. al. 2023) free from copyright restrictions were prioritized. These include content from Sanskrit Wikipedia, which provides a substantial collection of freely usable linguistic material. Additionally, a broad range of public domain literature was explored, encompassing classical texts, historical works, and other writings available across various digital repositories. Throughout this process, careful consideration was given to the ethical implications of collecting and

utilizing textual data, ensuring that all sourced material adheres to applicable legal and ethical standards.

3.1 Text Collection Tools

For large-scale text collection, the present study employed the IL Crawler[1] tool a specialized utility designed to systematically gather, scrape, and extract textual data from a broad spectrum of digital sources. These sources include websites, online documents, social media platforms, and various other repositories accessible on the internet. The tool operates by accepting URLs as input and autonomously retrieving and compiling the relevant linguistic material from the specified locations.

The primary strength of the IL Crawler lies in its ability to automate the otherwise labour-intensive and time-consuming process of data acquisition. By eliminating the need for manual data collection, the tool ensures a high degree of efficiency, consistency, and scalability in corpus construction qualities that are particularly critical when building large-scale linguistic resources such as a Sanskrit speech corpus.

4. Audio Recording

We obtained audio recordings of the collected text from a total of 181 speakers. In the initial stage, 112 speakers participated. The second stage involved 53 speakers. In the final stage, 16 speakers were engaged. The development of high-quality audio recordings requires meticulous attention to both the equipment used and the recording environment. For this purpose, we carefully selected appropriate equipment and optimized the recording conditions. The choice of microphone is particularly critical in determining recording quality. We employed clip-on microphones in conjunction with closed-back headphones, allowing for hands-free recording, real-time audio monitoring, and effective isolation from external noise.

Equally important was the selection of a recording location with minimal background interference. To ensure clarity, we avoided areas affected by traffic, electrical appliances, or other disruptive noise sources. Furthermore, recording rooms were acoustically optimized by closing windows and doors and applying weather stripping to seal gaps, thereby reducing external disturbances.

Through this rigorous setup, we ensured that the recordings achieved the necessary standard of clarity and consistency, providing a reliable foundation for subsequent speech processing tasks.

4.1 Speaker Selection and Variation

Careful attention was given to speaker selection in order to ensure a diverse and representative corpus. Speakers were recruited from a wide range of age groups, spanning from 15 to 45 years, to capture variation in voice characteristics, speech patterns, and pronunciation styles across different life stages. To promote regional diversity, participants were selected from multiple states and regions across India, representing a broad spectrum of geographical and cultural backgrounds. Furthermore, speakers with varying native language backgrounds were included, reflecting India's rich multilingual landscape and accounting for the influence of different mother tongues on Sanskrit pronunciation and intonation.

In addition to regional and linguistic diversity, gender balance was maintained throughout the selection process. An equal

[1] https://sanskrit.jnu.ac.in/download/index.jsp (Accessed on 12/04/2026)

representation of male and female speakers was ensured across all phases of recording, enabling the corpus to capture the full range of acoustic and phonetic variation associated with gender. This carefully balanced and inclusive approach to speaker selection significantly enhances the robustness, diversity, and generalizability of the corpus, making it a more reliable resource for training and evaluating Sanskrit speech recognition and synthesis systems.

4.2 Audio Recorder

To streamline the audio recording process, a dedicated web application recorder was developed using a Python Django backend server. This solution was specifically designed to address several key challenges inherent in large-scale speech data collection, including displaying text to participants, capturing audio recordings, standardizing file naming conventions, maintaining quality standards, and efficiently uploading and organizing the recorded files. The system automatically generates metadata including participant IDs and sentence IDs immediately after each recording is completed, enabling real-time tracking and organization of collected data. The recorded audio files are stored in a structured directory hierarchy, with a parent folder containing individual subfolders for each participant, ensuring clean and systematic file management.

In the context of the present corpus, this web application served as the primary recording platform for both Phase 2 and Phase 3 of data collection. The tool ensured consistent audio quality and uniform file formats across all recordings, significantly reducing the time and effort required for post-processing. The automated file naming and metadata generation further eliminated manual errors, improved data organization, and enabled faster file retrieval. By centralizing and standardizing the entire recording workflow, the web application substantially enhanced the efficiency, reliability, and scalability of the corpus development process.

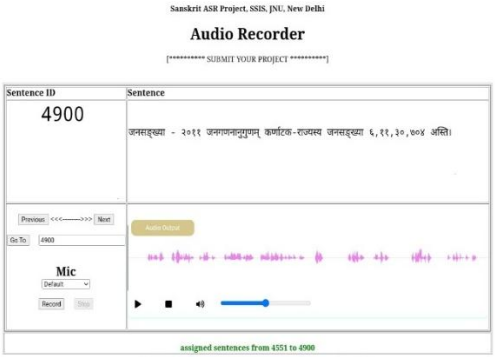

Figure 1: The Audio Recorder

4.3 Audio Editing and Formatting

The process of manipulating audio recordings plays a crucial role in improving quality, enhancing content, and preparing the data for linguistic applications. For this purpose, we employed the audio editing software Audacity along with an online audio converter, selected on the basis of functionality, reliability, and user familiarity. Essential editing tasks included adjusting the volume, panning, and timing of each track to achieve the desired sound balance and clarity. All recorded audio files were standardized by converting them into the .WAV format, a widely accepted format in automatic speech recognition research. Furthermore, parameters such as sample rate, bit depth, and encoding specifications were carefully configured in accordance with the requirements of speech processing systems. To facilitate efficient organization and categorization, comprehensive metadata was appended to each file, including details such as the recorder's name, gender, age, and region of origin. This metadata significantly enhances the dataset's utility for linguistic and sociolinguistic analysis.

Following the editing process, each audio file was subjected to critical listening to ensure that it satisfied the established

standards of clarity and accuracy. Additionally, recordings were tested across multiple playback devices to verify consistency and quality under different listening conditions. This rigorous post-processing pipeline ensured that the final dataset was both technically robust and linguistically valuable.

Phase	No. of speakers	No. of audio files	Duration (hours)
1	112	20,763	67
2	53	9,828	30
3	16	2,952	10
Total	**181**	**33,543**	**107**

Table 1: overview of JNU Data Set

4.4 Size of Speech Corpus

The size of a speech corpus is generally measured in terms of the total duration of audio recordings it contains. In the present work, a total of 33,543 audio files were recorded from 181 speakers, resulting in more than 107 hours of original in-house speech data. In addition, two publicly available Sanskrit speech corpora were collected from online sources, contributing a further 105 hours of data. Altogether, the present corpus comprises approximately 212 hours of Sanskrit speech data, making it one of the largest Sanskrit speech resources developed to date.

5. Conclusion

This paper presents the development of a comprehensive Sanskrit speech corpus aimed at advancing computational processing and digital preservation of Sanskrit. The corpus comprises a total of 212 hours of spoken Sanskrit data — of which 107 hours were recorded in-house from 181 speakers across three systematic phases at JNU, New Delhi, while the remaining 105 hours were sourced from publicly available online corpora. Together, these resources constitute one of the largest Sanskrit speech datasets developed to date.

The corpus development pipeline encompassing web crawling, text sanitization, audio recording, transcription, and forced alignment was designed with rigor and scalability in mind. The introduction of a custom-built web application recorder in Phases 2 and 3 standardized the recording process, reduced costs significantly, and improved overall data quality and consistency. The resulting corpus holds considerable promise for a wide range of applications including Automatic Speech Recognition (ASR), Text-To-Speech (TTS) synthesis, linguistic research, and digital preservation of Sanskrit manuscripts. Beyond its technological utility, this work carries deeper cultural significance helping bridge the gap between an ancient language and modern speech technology.

6. Future Work

The present corpus opens several promising directions for future research. Most immediately, the data will be used to train and evaluate Sanskrit ASR models, teaching systems to accurately recognize and transcribe spoken Sanskrit. The corpus will equally support the development of Sanskrit-capable voice assistants, enabling natural language understanding and response in Sanskrit. Beyond speech technology, linguists and sociolinguists can leverage this resource to study regional dialects, accents, and language variations among Sanskrit speakers. Expanding the corpus in terms of speaker diversity and domain coverage remains a long-term goal,

with the aim of building an increasingly robust foundation for Sanskrit speech technology research.

7. References

Javed, T., Nawale, J. A., George, E. I., Joshi, S., Bhogale, K. S., Mehendale, D., ... & Khapra, M. M. (2024). Indicvoices: Towards building an inclusive multilingual speech dataset for indian languages. *arXiv preprint arXiv:2403.01926*.

Chadha, H. S., Gupta, A., Shah, P., Chhimwal, N., Dhuriya, A., Gaur, R., & Raghavan, V. (2022). Vakyansh: ASR Toolkit for Low Resource Indic languages. *arXiv preprint arXiv:2203.16512*.

Gupta, A., Chadha, H. S., Shah, P., Chhimwal, N., Dhuriya, A., Gaur, R., & Raghavan, V. (2021). Clsril-23: Cross lingual speech representations for indic languages. *arXiv preprint arXiv:2107.07402*.

Adiga, D., Kumar, R., Krishna, A., Jyothi, P., Ramakrishnan, G., & Goyal, P. (2021). Automatic speech recognition in Sanskrit: A new speech corpus and modelling insights. *arXiv preprint arXiv:2106.05852*.

Holla, S. S., Kumar, T. M., Hiretanad, J. R., Deepak, K. T., & Narasimhadhan, A. V. (2022, March). End-to-end speech recognition for low resource language sanskrit using self-supervised learning. In *2022 International Conference on Wireless Communications Signal Processing and Networking (WiSPNET)* (pp. 148-152). IEEE.

Kumar, D. & Jha G. N. (2023). Resource Creation for Sanskrit ASR (Automatic Speech Recognition). *An International Journal of Engineering Science,* 35, 103-109.

The *shabd* portal – searchable lexical resources for Indian languages by Government of India

Mercy Lalrohluo Hmar[1], Girish Nath Jha[2], Dhananjay Singh[3]

Commission for Scientific and Technical Terminology[1]; Jawaharlal Nehru University[2]; Chairman, Commission for Scientific and Technical Terminology[3]
New Delhi, India[1], New Delhi, India[2], New Delhi, India[3]
mercylhmar.cstt@gmail.com[1] , girishjha@jnu.ac.in[2], dhananjaysinghcstt@gmail.com[3]

Abstract

The shabd portal (https://shabd.education.gov.in) of Commission for Scientific and Technical Terminology (CSTT), a subordinate office under the Ministry of Education, Department of Higher Education, Government of India (GOI) is a data server designed and developed by Prof. Girish Nath Jha, former Chairman CSTT, featuring all the standardized scientific and technical glossaries of CSTT in digital searchable mode. The aim is to launch a central repository for the terminologies prepared in Indian Languages, thus enriching the language bank of India enabling user friendly and free access to standardized terminology. This website is available in 22 Indian languages. The data covers several domains of science, humanities, engineering, medical science and agriculture subjects. The data is dynamic with regular updates in various domains. Users can search the equivalents of terms in Indian languages and submit their feedback for those equivalents prepared by CSTT. The unique feature of the search platform is that users have various options for search, based on languages, subjects, dictionary type and language pairs. The user can also choose to search in a specific glossary or the entire collection which includes about 471 glossaries having about (29,56,125 headwords).

Keywords: standardized terminology, Indian languages, domains, searchable mode, scientific and technical glossaries

About CSTT -

The Commission for Scientific and Technical Terminology of GOI

The Commission for Scientific and Technical Terminology was established with the objective to evolve technical terminology in all Indian Languages, on 1st October 1961 by the Presidential Order dated April 27, 1960, through a resolution of the Government of India (Ministry of Education), as per the recommendations of the Committee constituted under the provisions of the Clause (4) of the Article 344 of the Constitution of India. The Commission was established with the objective to evolve standardized technical terminology in all Indian Languages. Prof. Dhananjay Singh is the current Chairman of CSTT.

The main objectives of the Commission are to evolve standard terminology, propagate its use, obtaining useful feedback and distribute it widely. In the process of evolution of scientific and technical terminology and reference material in Indian Languages, the Commission collaborates with State Governments/ Universities/ Regional Text-Book Boards and State Granth Academies to ensure uniformity of terminology in Indian languages.

The Commission publish glossaries, definitional dictionaries, journals, monographs, encyclopaedia etc; to see that the evolved terms and their definitions reach the students, teachers, scholars, scientists, officers etc., to ensure proper usage necessary updating/ correction improvement on the work done (through workshops/ seminars/ orientation programmes) by obtaining useful feedback, to coordinate with all States to ensure uniformity of terminology in Hindi and other Indian languages.

Some of the Schemes of the Commission are as follows:

1. Preparation of English-Hindi and Hindi-English Technical Glossaries/Dictionaries
2. Preparation of Bilingual Technical Glossaries/Dictionaries
3. Preparation of Trilingual Technical Glossaries/Dictionaries
4. Preparation of National Technical Terminology
5. Preparation of Definitional Dictionaries
6. Preparation of Technical Encyclopedias
7. Preparation of School-Level Terminology
8. Preparation and/or Approval of Departmental Glossaries

Proceedings of WILDRE-8, pages 61–66

9. Revision and Updating of Glossaries
10. Identification and Publication of Pan-Indian Terms
11. Propagation, Expansion and Critical Review of Terms Coined and Defined via Seminars/ Conference/ Workshops.
12. Scheme of Production of University Level Books in Hindi and Regional Languages
13. Preparation and Publication of Monographs
14. Preparation and Publication of Journals (Gyan Garima Sindhu (for Humanities subjects) and Vigyan Garima Sindhu (for science subjects).
15. Sales of Publications
16. Free Distribution of Publications
17. Organizing Exhibitions

About the *shabd* portal

One of the new initiatives taken by the CSTT in the field of propagation of scientific and technical terminology is the *shabd* portal hosted at https://shabd.education.gov.in .

The *shabd* is a data server which features all the glossaries of CSTT in digital searchable mode. Other institutions/ agencies preparing dictionaries will also be able to host their work in digital form on this platform. The aim is to showcase a central repository for all the terminologies prepared in/for Indian Languages.

Through this platform the users will not only be able to search the equivalent terms of scientific and technical terminology in Indian languages but will also be able to register their feedback for the equivalents already prepared by CSTT. This aims at providing a user-friendly search environment and also in involving the users via the feedback mechanism. The unique feature of the search platform 'https://shabd.education.gov.in' is that the user will have various options for search, whether based on languages, subjects, dictionary type, or language pairs. Not only this, the user can choose to search in a specific glossary or through the entire collection which currently includes about **471** glossaries having about (**29,56,125** headwords).

One website – 22 languages- wide user base

The 'Shabd' website is accessible in 22 languages and contains the data in the form of searchable glossaries in 22 Scheduled Indian languages. The users can easily search the glossary and use the plethora of Indian language equivalents available for a wide range of subject domains.

Concept design execution, website development and programming was done by Prof. Girish Nath Jha, former Chairman, CSTT and was launched in March 2024.

The following partners of CSTT have greatly contributed to make the content available for the website:

- Udaan Project team, IIT Bombay (All Languages) and International Centre for Free and Open Source Software (ICFOSS), Kerala (Malayalam) for OCR of printed glossaries

- GIST group, CDAC Pune for UNICODE Consultation and support

- Central Institute of Indian Languages (CIIL) Mysore for Data Typing support

- Bhartiya Bhasha Samiti (BBS), Overall Guidance

- Language Division, Ministry of Education for Administrative Support

History of *shabd*

The inception

CSTT prepares the terminologies through the Expert Advisory Committees, consisting of subject and language experts along with linguists and Sanskritist who are focused on finding out the equivalent terms in the specific subject areas and language. The terminology prepared by CSTT has not only been used by Granth academies, textbook boards and publication cells to prepare textbooks but is also being used by institutions such as NCERT, NTM, AICTE and so on.

CSTT glossaries available in printed form or pdf form on the official website were not searchable and this made it less popular among the user group. The need to have easily accessible and searchable glossaries resulted in the inception of the vision for a website wherein the glossaries prepared by CSTT are available in searchable form.

Thus began the humongous task of building such a website and getting the data of the glossaries in digital form available for the task to take shape. Prof. Girish Nath Jha, designed and programmed

a server which could host the data locally. After a long and strenuous brainstorming, the initial layout of the local server took shape. Based on several trials, hits and errors, the server slowly evolved to take a stable form. The demo of the server was shown on several occasions to collect valuable inputs and suggestions from the persons of interest and such constructive suggestions were implemented to make the server more user friendly and useful.

For the data, CSTT began with the glossaries available in digital soft copy which were initially digitized as part of the agreement between CSTT and CIIL. Then it was found that many legacy publications of CSTT were not available in such digital form. Thus began the search for partners who could do this task in record time. This led to our MoU with UDAAN, IITB team led by Prof. Ganesh Ramakrishnan. The team has helped CSTT in digitizing most of the legacy documents and are continuing to do so.

Once the server was tested and checked locally, the website team helped to carry out the necessary task required to acquire the domain and host the *shabd* website in a public domain for use by the users. The site first went active in March, 2024, since then, it has had 20,66,090 hits from across the country and the world as of now.

The journey till date

Initially the website was built having an English and Hindi layout. Both the layouts were not interconnected. Changes had to be made on both pages regularly when any suggestions or bugs were identified. Thereafter the layouts were merged and programmed by Prof Jha to switch, without having to make corrections in every page and content.

The search was simple and user had to search in a specific glossary initially. Next the search was extended to all glossaries and the feedback mechanism was also introduced. The next big jump was the introduction of all 22 Indian languages. All the code strings were translated to the 22 target languages by a panel of linguists who were a part of CSTT's meetings to update the principles of terminology preparation. The linguists timely helped CSTT whenever there was any issue to give language equivalents for the various contents available on the website. The panel of linguists headed by Prof. K. L. Verma, Former Chairman, CSTT are:

1. Dr Bipasha Patgiri, Assistant Professor, Department of Linguistics and Language Technology, Tezpur University, Assam
2. Prof. Niladri Sekhar Dash, Head, Linguistic Research Unit, Indian Statistical Institute, Kolkata
3. Prof. Swarna Prabha Chainary, Professor, Dept. of Bodo, Gauhati University
4. Prof. Lalit Mangotra, President, Dogri Sanstha, Jammu
5. Dr. Baldevaanand Sagar,National President, World Sanskrit Media Council
6. Prof. Chandan Kumar,Professor, Faculty Member,Dept. Of Hindi, University of Delhi
7. Dr. Girisha Bhat A, Professor and Principal, Govt First Grade College, Siddakatte, Karnataka
8. Prof Aadil Amin Kak, Dean, School of Arts, Languages and Literatures University of Kashmir
9. Dr. Kiran Budkuley,Author, Critic& Translator, President, Aksharpath, Goa.
10. Prof. Awadesh Kumar Mishra, Professor, Chief Coordinator, Bharatiya Bhasha Samiti, Vishwakarma Bhawan, IIT Delhi, New Delhi
11. Dr. Shobha L, Member Research Staff, AU-KBC Research Centre , Madras Institute of Technology, Anna University, Chennai
12. Dr Hanjabam Surmangol Sharma, Assoc. Professor, Department of Linguistics, Manipur University, Imphal
13. Dr. Shakuntala Gawde, Head and Assistant Professor, Department of Sanskrit, University of Mumbai
14. Shri. Vishnu Bahadur Gurung, In-Charge (Rtd.), Nepali Service, External Services Division, All India Radio, New Delhi
15. Prof. Panchanan Mohanty, Dean, School of Languages and Literature/Humanities, Nalanda University
16. Dr Suman Preet, Professor , Department of Linguistics and Punjabi Lexicography, Punjabi University Patiala
17. Prof. Satyapal Singh, Professor, Department of Sanskrit, University of Delhi

18. Dr. Thakur Prasad Murmu, Assistant Professor, Department of Santali, Sidho Kanho Birsha University
19. Smt. Shalini Sagar, Senior Broadcaster All India Radio, Sindhi Language Expert, NCERT
20. Dr. S. Arulmozi, Professor & Head, Centre for Applied Linguistics and Translation Studies, University of Hyderabad
21. Dr. MC Kesava Murty, Professor, Dept. of Dravidian and Computational Linguistics, Dravidian University, Kuppam
22. Prof. Syed Imtiaz Hasnain, Retd. Prof of Sociolinguistics, Chair-Professor, Maulana Azad National Urdu University (MANUU)

Figure 1: Shabd website localization

Main features of the Shabd Website:

The various features which were added slowly as the 'shabd' website began to become popular among the users are:

- **Localization of website content**

Figure 1: Main content of the website is available in 22 Indian languages which can be manually selected

- **Website link to language information page/ website of organization working for that language**

Figure 2: Link to the major organization working in the specific language provided on the main page when language preference is selected

- **Addition of the footer with features such as -popular website, credits, contact us etc**

Figure 3: Footer has all the popular and websites along with the credit and contact details

- **Search options – language wise, subject wise, language pair wise, glossary type, etc.**

Figure 4: many options to customize the search is available for the user

- **Transliteration feature**

Figure 5: Transliteration feature for all non-Devnagari glossaries

- **Glossary details**

Figure 6: Glossary title, status of the glossary, headword count and Officer in-charge details and contact details are available for the glossaries

- **List of Expert involved in the project**

Figure 7: Expert panel involved in terminology preparation is mentioned in this section

- **Log of the user stat & feedback**

Figure 8: Feedback of user is recorded and necessary action is taken from time to time

- **User feedback mechanism**

Figure 9: User can provide feedback for every equivalent and contribute. Feedbacks are considered and placed before the Expert Advisory Committee for review

- **Popular glossaries on main page**

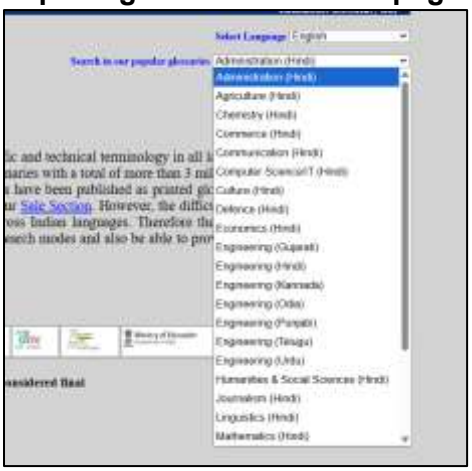

Figure 10: All popular glossaries are enlisted in this as per the frequency of user hits

- **List of glossaries in alphabetical order on main page**

Figure 11: The alphabetical list of all glossaries is available as drop-down menu on main page

The entire collection which as of now includes about **471** glossaries having about **29,56,125** headwords. This covers disciplines from Humanities, Social Science, Medical Science, Engineering, Agricultural Sciences and Science which include **more than 60 subjects** such as Agriculture, Public Administration, Chemistry, Botany, Zoology, Psychology, Physics, Economics, Ayurveda, Mathematics, Civil and Electrical Engineering, Computer Science, Political Science, Culture, Transport, Geology,

Capital Market, Cell Biology, Broadcasting, Journalism, Music and Fine arts, CSIT, AIML, Linguistics, Forestry, Entomology, Plant Pathology, Soil Science, Sports, Nematology, Sericulture, LIS, etc. Many more domains are being added as and when the terminology preparation meetings are being held across the country.

Bibliographical reference:

1. Resource available from CSTT main website (https://cstt.education.gov.in)
2. Resource available from CSTT shabd website (https://cstt.education.gov.in)

POS Tagging in Low-Resource Maithili Language: Specific Challenges and Nuances

Shivani Priya, Shruti Jha, Urmila Jha, Deepali Tiwari, Jyoti, Girish Nath Jha

School of Sanskrit and Indic Studies, Jawaharlal Nehru University,
New Delhi-110067
{priyashivani683, 17shrutijha, urmilajha006, deepali0128, girishjha}@gmail.com,

jyotiraj@mail.jnu.ac.in

Abstract

Part-of-Speech (POS) tagging is a key step in Natural Language Processing (NLP), laying the groundwork for more advanced syntactic and semantic tasks. Despite Maithili's status as an Indo-Aryan language with a rich literary tradition and official recognition in India, computational resources for it are still very limited. In this paper, the creation of an annotated corpus of 25,000 sentences drawn from the fields of health, tourism, and administration is described with the hierarchical tagset currently used for Maithili. This paper also indicates that standard tagsets, typically adapted from English or Hindi, fail to capture the linguistic nuances of Maithili. This underscores the need for a dedicated tagging framework that considers characteristics like vocative particles, verbal nuances, honorific complexities.

Keywords: Parts of Speech, Natural Language Processing, Maithili, annotation

1.Introduction

Tagging became a large part of Natural Language Processing (NLP) that attributes grammatical labels or noun, verb, adjective etc. to words within a sentence. The process is central to most NLP tasks, such as parsing, machine translation, and text-to-speech synthesis since it gives morpho-syntactic information (Hardie, 2003) (Priyadarshi et al., 2023) regarding words and their distribution in a sentence. Surface form of a text is converted into morpho-syntactic labelled form by POS annotation. It is used as an interface between raw corpus data and a more abstract language processing with the conversion of unstructured text into structure-analyzed data at the morpho-syntactic level. However, in the instance of Maithili, the Indo-Aryan language predominantly spoken on a vast, expansive and heavily populated territory, in the state of Bihar in India and in Nepal, this medial role is brought to the fore. Despite having a rich literary tradition and being given institutional acknowledgment in the Schedule VIII of the Indian Constitution, Maithili is still considered relatively poor in the area of computational resources. It is a morphologically inflectional language with gender, number, case, honorific inflexions, and complex verb forms of agreement (Kumar and Chaudhary, 2025).

In this type of a wide-ranging system, POS annotation facilitates the syntactic analysis, Part of Speech (POS) or morpho-syntactic semantic processing and creation of sophisticated NLP tools. Therefore, in Maithili, it is a preliminary move towards the development of annotated corpora and the creation of high-level language technologies. In Maithili, annotation of POS is usually performed with the help of the standardized tagsets like the Bureau of Indian Standards (BIS) tagset to provide consistency and interoperability between the languages of India (Gopal, 2011). Labelled language resources for Maithili will support many kinds of NLP applications, such as machine translation, information retrieval, morphological analysis, and parsing.

2. Maithili Linguistic features

2.1 Rich morphology

Maithili has a highly inflection morphology. It displays conjugation of verbs in terms of tense (past, present, future), person (first, second, third), mood (indicative and imperative) and aspect (perfective and imperfective). Another indication of Maithili is that non-verbal agreement is responsive to the honorific status of the subject (Jha et. al. 2018).

Example: अहाँ जाइ छी।
(You (HON) are going.)

The auxiliary 'छी' marks honorific agreement.

Proceedings of WILDRE-8, pages 67–74

2.2 Relatively free word order

Maithili has a relatively free word order language, which is frequently determined by constituent structure and discourse prominence as opposed to strict syntactic patterns. It follows that, positional factors (commonly applied to fixed word order language POS tagging) are not absolutely reliable. Therefore, categorization should be done correctly by relying on functional and morphological diagnostics as opposed to a linear position.

Example: **a) SOV-**

हम घर जाइत छी।

“I am going home.”

b) SVO-

हम जाइत छी घर।

Acceptable in spoken/discourse context

c) OSV-

घर हम जाइत छी।

Home focused, topicalization

2.3 Particles

Particles such as ‘तऽ’, ‘सेहो’, ‘तइयो’, and ‘बादो’ inclusive in the Maithili language. These aspects add pragmatic senses like emphasis, contrast or continuation and not lexical senses.

2.4 Demographic variations

Maithili has an enormous demographic variation based on the area, age, gender/socio-educational dimension and this has a direct impact on the part-of-speech annotation. The regional dialect is also observed in the realisation of the auxiliaries and particles such that the same progressive or aspectual construction can be formed with the alternative auxiliaries in the different dialect such that it leads to the existence in the possible differences in the marking of the auxiliary verbs, and the subsequent confusion of the main and auxiliaries.

Example:

ओ कहैत अछि।

ओ कहै है।

“He says.”

काज कऽ लियऽ।

काज कए लियऽ।

“Do the work.”

These examples demonstrate that demographic variation in Maithili produces multiple surface forms for the same syntactic category, requiring POS annotation guidelines that rely on functional and contextual interpretation rather than purely formal criteria.

3. Annotated data-set

In case of annotation of Maithili POS, the data set was tagged in such a way that it would include domain variety and acceptable linguistic coverage. Texts were collected from various domains, including health, tourism, and administration. This multi-domain strategy was followed so as to represent the difference in vocabulary, morphology and syntactic structures among the various registers of usage. The total number of data tagged was approximately 25k sentences, as follows-

S. No.	Domain	Data
1.	Administration	7000
2.	Law	5267
3.	Education	6310
4.	Tourism	838
5.	Health	1008
6.	Agriculture	763
7.	Technology	567

Table 1: Domains and data

4. Tagsets used for tagging

The ILCI Annotation Tool (ILCIANN) (Kumar et al., 2012) is a server-based web application developed under the Indian Languages Corpora Initiative to facilitate large-scale word-level annotation for Indian languages. It is designed to be useful in producing annotated corpora to support NLP, particularly in less-resource languages, through an annotated centralized, uniform environment. It allows annotation of POS through manual means in standardized groups of tags (Bureau of Indian Standards (BIS) tagset and a little automatic tagging in closed grammatical categories) to minimize the workload of the annotators. The tool ensures that annotations will be stored on a central server in sentence-by-sentence form, reducing the data loss and gaps.

Sl. No.	Category		Label	Annotation Convention	Examples
	Top Level	**Sub-type**			
1.	**Noun**		N	N	
1.1		Common	NN	N_NN	किताब, गाछ
1.2		Proper	NNP	N_NNP	राम, मधुबनी
1.3		Nloc	NST	N_NST	नीचाँ, आगू
2	**Pronoun**		PR	PR	
2.1		Personal	PRP	PR_PRP	आहाँ, हम
2.2		Reflexive	PRF	PR_PRF	अपना, अपना-आप
2.3		Relative	PRL	PR_PRL	जकर, एकर
2.4		Reciprocal	PRC	PR_PRC	आपस, परस्पर
2.5		Wh-word	PRQ	PR_PRQ	केकर, कतय
2.6		Indefinite	PRI	PR_PRI	कोनो,
3	**Demonstrative**		DM	DM	
3.1		Deictic	DMD	DM_DMD	ओ, एहि
3.2		Relative	DMR	DM_DMR	जकर, जे, जेना, जकाँ
3.3		Wh-word	DMQ	DM_DMQ	के, कखन
3.4		Indefinite	DMI	DM_DMI	कोनो, कोई
4	**Verb**		V	V	
4.1		Main	VM	V_VM	चलब, देखब
4.2		Auxiliary	VAUX	V_VAUX	अछि, थिक
5	**Adjective**		JJ	JJ	सुन्दर, नीक

6	**Adverbs**		RB	RB	अचानक, धीरे
7	**Postpositions**		PSP	PSP	केर, लेल
8	**Conjunctions**		CC	CC	
8.1		Co-ordinator	CCD	CC_CCD	आओर, मुदा
8.2		Subordinate	CCS	CC_CCS	जँ-तँ, जखन-तखन
9	**Particles**		RP	RP	
9.1		Classifier	CL	RP_CL	बला, टा
9.2		Default	RPD	RP_RPD	तऽ, सेहो
9.3		Interjection	INJ	RP_INJ	अरे, हे
9.4		Intensifier	INTF	RP_INTF	एतेक, बड
9.5		Negation	NEG	RP_NEG	नहि, मत
10	**Quantifiers**		QT	QT	
10.1		General	QTF	QT_QTF	खूब, कनि
10.2		Cardinals	QTC	QT_QTC	एक, दू
10.3		Ordinals	QTO	QT_QTO	पहिल, दोसर
11	**Residuals**		RD	RD	
11.1		Foreign word	FW	RD_FW	Website, Link
11.2		Symbol	SYM	RD_SYM	$, &
11.3		Punctuation	PUNC	RD_PUNC	।,?
11.4		Unknown	UNK	RD_UNK	
11.5		Echo words	ECH	RD_ECH	अलग-थलग, हुइल-माइल

Table 2: Tagset for Maithili language

Example:

- सरकारी\JJ स्कूल\N_NN मे\PSP प्रवेश\N_NN लेल\PSP 'प्रवेश\N_NN सप्ताह'\N_NNP शुरू\N_NN भऽ\V_VM गेल\V_VAUX अछि\V_VAUX ।\RD_PUNC
  "Admission week has started for admission in government schools (Priyadarshi Ankur, 2023)."

- चाउरक\N_NN क्वालिटी\N_NN नीक\JJ भेला\V_VM पर\PSP बजार\N_NN मे\PSP एकर\DM_DMD दाम\N_NN सेहो\RP_RPD नीक\JJ भेटत\V_VM\।\RD_PUNC
  "If the quality of rice is good, it will fetch a good price in the market."

5. Nuances found in the process

There were also some linguistic nuances that caused discrepancies in the process of POS annotation. The Penn Treebank tagset of English was used as a source of influence by annotators, who sometimes think of intensifiers (बड, खूब, बेसी) as general adverbs, overlooking their distinct functional role in Maithili.

Example: एहि\DM_DMD कीटनाशक\N_NN कें\PSP खेती\N_NN मे\PSP बड\RP_INTF उपयोग\N_NN अछि\V_VAUX ।\RD_PUNC

"This pesticide has many uses in farming."

Here, the word 'बड' has been tagged as an intensifier.

ओ\DM_DMD बड\RB सुन्दर\JJ बचिया\N_NN छैक\V_VAUX ।\RD_PUNC

"She is a very beautiful girl."

Here, even though the word 'बड' acts as an intensifier but the tagger has tagged it as adverb because of the influence of Penn Treebank from English.

Additionally, some words were tagged based on their lexical identity rather than the sentence context, whereas in the sentence, the word functioned as a proper noun.

Example: भारतीय\JJ जनता\N_NN पार्टी\N_NN was tagged as an adjective while the word in the context was functioning as a proper noun.

Another instance is in conjunct verbs like शुरु\V_VM कएल\V_VAUX गेल\ V_VAUX रहय\V_VAUX ।\RD_PUNC (was started), the noun 'शुरु' (start) is frequently mislabelled as a verb instead of noun. This occurs because English linguistic interference treats "started" as a single verbal unit.

Moreover, as a result of the uncertainty between Main Verbs and Light Verbs in Compound Verb Constructions (CVCs), a tagger can tag auxiliary verbs as a main verb, although the auxiliary verb loses its literal sense to give aspectual state in a sentence.

Example:

मारि\V_VM देब\V_VM (will kill). In this case, the word 'मारि' is the main verb, and 'देब' is the light verb showing the completion of the action. If 'देब' is tagged as a Main Verb, the system may interpret the sentence as having two distinct actions ("killing" and "giving") rather than one unified event.

6. Challenges

6.1 Low-resource language

Maithili is generally a low resource language in the field of computational linguistics, where there are not many annotated corpus, standardized tagsets, and pre-trained language models. The issue is also aggravated by data sparsity which can impact especially when it comes to processing rare morphological forms or dialectal variation as well as discourse particles. locative forms like घरमे, गाममे, or dialectal variants such as घरम' काज कऽ, काज कए may not appear frequently enough for the system to robustly learn the postpositional pattern '-मे/-म' 'कऽ/कए' as demographic variation.

6.2 Cliticization and particles

Not all words are morphologically simple but represent complex structures that are formed by a component (deictic, pronoun or adverb)

and enclitic particles such as 'नो' 'यो' 'हु' 'बे' in 'तखनो', 'तइयो', 'हमहु', 'हेबे करतै', creating a structural ambiguity. The tagger has to determine whether to assign a single unified tag (e.g., RB or PRP or PSP) or to recognize and segment the internal components (e.g., PRP + RP).

6.3 Honour displayed in verb forms

The rich honorific system that Maithili has also makes tagging more difficult. Pronouns such as 'तूँ', 'ई', and 'अहाँ' trigger distinct verb forms, increasing morphological variation and data sparsity. In Maithili traditions, daughter-in-laws often address their relatives-in-law in the third person, using corresponding verb forms to express respect. Eg- 'ई बैसौथ' (You "hon." sit.), 'ई लऊथ' (You "hon." take). Here, forms that function as third-person pronouns pragmatically serve as an honorific second-person address.

Another instance of this is 'अपनेक बैसियौ' (you "hon." sit), 'अपने' is reflexive but here functions as a personal pronoun that marks honour.

Treating honorificity solely as a semantic feature fails to capture its grammatical impact, leading to loss of information in POS annotation. This creates a significant ambiguity for POS tagging systems.

6.4 Intensifier and plurality

A challenge in Maithili POS tagging arises from the language's strategy of marking plurality. In the language, plurality is marked by 'सब'(all), as in 'बच्चा सब' (children), but 'सब' also acts as an intensifier when attached to 'सँ' (सबसँ) (superlative degree). The same problem is faced with the word 'बेसी' (very/many). The word 'बेसी' functions as an intensifier and a quantifier, like 'बेसी' चोट' (bad injury), 'बेसी' acts as an intensifier that shows the intensity of the injury, while in 'बेसी' कऽ आटा'(more flour) 'बेसी' here acts as a quantifier denoting the quantity of 'आटा' (flour).

6.5 Morphological case marker ambiguity

The word 'के' (of) is semantically versatile. At some places it acts as a genitive marker, like in 'सोमनाथकेँ मन्दिर' (Temple of Somanath), as an accusative marker, like in 'श्यामकेँ बजाऊ' (call Shyam), and as a wh-word, like in 'के आयल?' (who came). A tagger might default to a frequentist approach (e.g., always tagging 'के' as a Postposition), thereby missing the subtle genitive and other nuances.

6.6 Functional Overlap

Another major challenge in Maithili POS tagging is the functional overlap that bridges the gap between closed-class postpositions and open-class verbal roots. A prime example is the lexical item लेल', which exhibits functional polysemy across different syntactic environments. In phrases such as 'रामक लेल' (for Ram), 'लेल' functions as a postposition (PSP) denoting purpose or beneficence. Conversely, in the construction 'लेल गेल' (was taken/received), it functions as the main verb (VM) derived from the root "to take."

In the same way, the word 'क' also acts as a postposition denoting the accusative or genitive aspect of the word, like in the phrase 'रामक किताब' (book of Ram), whereas in the phrase 'काज कऽ लेलहुँ', 'क' is acting as a main verb form meaning "did the work".

6.7 Absence of Reduplication tag

A critical limitation in existing tagging frameworks is the absence of dedicated tags for reduplication, forcing a conflict between a word's literal lexical category and its contextual functional role. Maithili frequently employs morphological reduplication to encode distributive, frequentative, or aspectual nuances that are absent in the individual roots.

Example:

'घरे-घरे' (in every house) 'चलैत-चलैत' (while walking). Here the phrase 'घरे-घरे' seems like noun but it is reduplicated, so tagger gets confused while tagging.

7. Suggested future work

7.1 Addition of honorific tag

Maithili Honorificity has a direct influence on the agreement morphology and syntactic structure. Verbs are different in terms of amounts of respect and form paradigmatic contrasts that are grammatically inevitable in situation. The fact that all such forms are treated under one POS label distorts morpho-syntactic information. Suggested future work could explore incorporating an HON (honorific feature) within the POS framework or introducing feature-level annotation (e.g., PRON[+HON], VERB[+HON]).

7.2 Tagging of address words

Reclassification of address words like **'यौ', 'हौ', 'रौ', 'रे'** (hey!). These objects fall under the general group of interjections. But interjections usually have spontaneous emotional or expressive meaning (e.g. surprise, pain, exclamation). Conversely, these Maithili forms are mainly used as an address or vocative particle to attract attention, indicate interpersonal stance or as a part of a direct address construction. It would be possible to introduce special tags like VOC.PART (vocative particle) or ADDR (address marker) to have a more functional differentiation.

7.3 Creation of Maithili-specific Tagset

The existing model of Maithili-POS tagging has borrowed the model of Hindi and English POS tagging. These are insufficient to represent Maithili-specific characteristics, e.g. layered honorificity, vocative particles reduplication tag etc. The design of a linguistically based, Maithili-specific POS tagset, which considers morpho-syntactic and discourse-pragmatic categories, can assist in adding these characteristics to the structure.

8. Conclusion

The Maithili annotated corpus creation illustrates that language-specific frameworks are imperative in Natural Language Processing. Although this experiment was able to tag 25,000 sentences, it was found that conventional tagsets do not always reflect the structural peculiarities of Maithili, including its stratified honorific system, cliticization and functional polysemy.

The study has shown that the reduplication and vocative particle are not specially marked, which results in severe semantic loss, due to the influence of annotator bias. Such issues as the confusion of light verbs and main verbs, and the shift of postpositions when the context is changed prove that the general strategy cannot fit morphologically rich languages.

Finally, in order to enhance the accuracy of the tagging, future efforts should be focused on developing a Maithili-specific tagset. Adding such features as honorific markers and reduplication tags will make sure that the computational models capture the syntactic and pragmatic reality of the language better, which will be a more solid base of advanced Maithili language technologies.

REFERENCES

Bureau of Indian Standards (BIS). (2011). POS Tagset Guidelines for Indian Languages.

Choudhary, N. (2019). *LDC-IL* Maithili Speech Corpus Documentation. *LDC-IL*.

Choudhary, N., & Ramamoorthy, L. (2019). *LDC-IL* Raw Text Corpora: An Overview. In *Language Data Consortium for Indian Languages (LDC-IL)*.

Gopal, Madhav, & Jha, Girish Nath (2011). Tagging Sanskrit Using BIS POS Tagset. In *Information Systems for Indian Languages: International Conference, ICISIL, Communications and Computer and Information Science, Springer,* Volume-139, PP. 191-194.

Hardie, A., Archer, D., McEnery, T., Rayson, P. (2003). Developing an Automated Semantic Analysis System for Early Modern English. *In Proceedings of Corpus Linguistics*, PP. 22-31.

Jha, A. K., Singh, P. P., & Dwivedi, P. (2019, July). The Maithili text-to-speech system. In *IEEE International Conference on Electronics, Computing and Communication Technologies* (*CONECCT 2019*), PP. 1-6. IEEE.

Jha, Girish Nath and Kumari, Sangita (2010). Maithili Bhāṣā. In *Bharatīya Bhāṣā Paricaya,* Volume-2, PP. 453-478.Jha, S.K., Singh, P. P., & Kaul, V. K. (2018). VEA Model. In *Word Formation Process of Maithili MT*.

Kumar, Ritesh, Kaushik, Shiv, Nainwani, Pinkey, Banerjee, Esha, Hadke, Sumedh,

Jha, Girish Nath. (2012). Using the ILCI Annotation Tool for POS Annotation: A Case of Hindi. In *IJCLA Vol.3*, *Number 2, Jul-Dec 2012*, PP. 93-104.

Kumar, S. (2020). Named Entity Recognition in Maithili. Banaras Hindu University, Varanasi.

Kumar, Shantanu, Choudhary, Narayan (2025). Maithili Language Technology: A Survey. In *Language in India,* Volume-25(6), PP. 160-175.

Mundotiya, R. K., Gatla, Praveen, Kanwar, Nikita, Singh, Anil Kumar. (2025). Deep Learning-Based Similar Languages' POS Tagging: Experiments on Bhojpuri, Maithili, and Magahi. In *Lecture Notes in Networks and Systems, Springer*.

Priyadarshi, A., & Saha, S. K. (2023). A study on the performance of recurrent neural network-based models in Maithili part-of-speech tagging. In *ACM Transactions on Asian and Low-Resource Language Information Processing*.

Priyadarshi, A., & Saha, S. K. (2023). A study on the performance of recurrent neural network-based models in Maithili part-of-speech tagging. In *ACM Transactions on Asian and Low-Resource Language Information Processing*.

Singh, Srishti (2015), *Challenges in Automatic POS Tagging of Indian Languages- A Comparative Study of Hindi and Bhojpuri*, Master of Philosophy, Centre for Linguistics, Jawaharlal Nehru University.

Singh, Srishti, Banerjee, E. (2014). Annotating Bhojpuri Corpus using BIS Scheme. In *LREC 2014 workshop WILDRE*.

Preserving Civilisation Memory: A Digital Humanities Approach to The Ramayana

Shashank Tiwari[1], **Girish Nath Jha**[2]
School of Sanskrit & Indic Studies
Jawaharlal Nehru University
New Delhi- 110067, India
[1]tiwarishashank1104@gmail.com [2]girishjha@jnu.ac.in

Abstract

The Ramayana is one of the most important texts in world literature, with a multiplicity of textual traditions and more than 300 variants throughout South and Southeast Asia. These manuscripts have been preserved in various forms, including palm-leaf codices, inscriptions on temple walls, in highly illustrated folios and through generations of oral performance. But the corpus now faces serious challenges due to environmental destruction, material fragility, fragmentation, and the scope of modern script recognition methods. The study examines how digital conversion projects are redefining the preservation of the Ramayana within a networked heritage system spanning the world. In this paper, the author critically assesses the work of large-scale projects through the use of a qualitative research design that has been conducted between the years 2003 and 2026, including the National Mission for Manuscripts (NMM), the digital reunification of the Mewar Ramayana, and efforts by Southeast Asian countries to document adaptations, like the Reamker by Cambodia. It discusses imaging standards, metadata formatting policies, integration of optical character recognition (OCR) and digital access structures. The study addresses multilingual complexity (Grantha, Devanagari, Kawi), partial coverage of variant texts, and infrastructural inequities. The results support the view that digitization has a significant positive impact on scholarly accessibility and comparative research, but the advantages remain unexplored, especially in relation to oral traditions. To make the endeavour sustainable, preservation and interoperable standards must be adopted, script recognition using AI should be encouraged, the community should be involved, and cross-border collaboration should be institutionalised to protect the long-term cultural viability of the Ramayana.

Keywords: Ramayana, Manuscript Preservation, Digital preservation of Cultural Heritage, Digital Humanities, Textual Transmission, Palm-Leaf Manuscripts, Grantha and Kawi Scripts, Metadata Standards (METS/XML), IIIF Interoperability, AI-Assisted Philology, Intangible, Cultural Heritage, Archival Sustainability, Cultural Heritage Informatics, Open Access Repositories, Civilizational Memory.

1. Introduction

1.1 Hindu Cultural Relevance of the Ramayana

The Ramayana is one of the oldest and most influential epics in the world, which is believed to have been written by Sage Valmiki and dated to the middle of the fifth millennium before the Common Era. It has, in its Sanskrit original, seven kandas and an estimated 24,000 shlokas, and it is not just a story but a source of moral philosophy, of devotional (bhakti) feeling, and of archetypal ethics, which today still animate societies throughout Asia. It is closely associated with the Indian festival of Diwali and with theatrical performances like Ramlila, and the imagery of the epic can be found in artistic work, including the sandstone carvings of Angkor Wat and the shadow puppets of Bali. UNESCO's recognition of Ramayana traditions underscores the epic's significance as a World Intangible Cultural Heritage, highlighting its role in creating community identity and cultural continuity.[1]

However, it is not a canonical version of the text. There are more than three hundred different recitations throughout the world, each a local rearticulation of the Valmiki story. Examples are the Cambodian Reamker, performed as classical dance-drama; the Ramakien of Thailand, which intermingles with royal murals in the Grand Palace; the Javanese Kakawin Ramayana of the ninth century, which combines Hindu epic motifs with indigenous aesthetic; and the Austronesian versions in Laos, Phra Lak Phra Lam and the Philippines, Maharadia Lawana. Although it is not only the Asian context where the echoes of the Ramayana can be found, it is possible to see its universality in the Iranian recensions and oral traditions of Africa. These variants, most of which are written on perishable palm leaves, birch bark, or temple inscriptions, are a standard part of civilizational heritage and are at risk of extinction.

[1]UNESCO 2019

Proceedings of WILDRE-8, pages 75–87
12 May 2026.

1.2 Digital Preservation and Its Problems

The cultural products of the Ramayana tradition confront a plethora of existential challenges: climate change, colonialism, increasing urbanisation, and institutional decay. The Valmiki original and most of the existing copies found in Southeast Asia are written on palm-leaf manuscripts, which disintegrate with moisture and insect damage; a significant collection is more than 1,000 years old. The looting of the colonial period fragmented collections; for example, the folios of the Mewar Ramayana were scattered between the British Library and the City Palace Museum in India. In Arunachal Pradesh, recently discovered tribal manuscripts are at risk of extinction unless action is taken to preserve them. The variants have been further threatened by political turmoil in the Southeast Asian region; the destruction of Reamker records under the Khmer Rouge rule is a case of such a threat.[2] Hard statistical figures paint a bleak picture: India alone can boast of an estimated five million manuscripts, of which more than 80 per cent are never digitised and are at high risk. According to the NMM's digital programme, 3.31 crore pages from 3.16 lakh manuscripts have been digitised. Still, innumerable works of the Ramayana conception in regional scripts such as Devanagari, Grantha, and Javanese Kawi remain in the offline form. In the absence of long-term preservation, these texts can be lost, severing their connection to pre-modern epistemological traditions.[3]

1.3 The Digitisation Imperative

One transformative solution is digitisation, which involves converting physical objects into high-fidelity digital surrogates using scanners, OCR, and standardised metadata models (such as Dublin Core or METS). The technology democratises scholarly access to academic resources, giving scholars around the world, e.g., access to a Bhandarkar Institute folio without having to travel there, as well as enabling AI-driven literacy analysis, digital reconstructions, and multilingual interfaces. Projects like Project Ramayanam provide overlay translations of Sanskrit originals that can be accessed via application programming interfaces, encouraging mobile interactions. The Digital Manuscripts Library of NMM gathers content that can be searched, and it focuses primarily on epics, including the Ramayana.[4]

Digitisation aligns with Sustainable Development Goal 11, Sustainable Cities and Communities, which calls for protecting intangible heritage and encouraging inclusive access. It alleviates the threat of the digital divide by creating open-access repositories, as IGNCA's public interfaces do. There are also ethical concerns arising: who owns what, whether it is more accessible or more secure against piracy, whether institutions have a role in managing digital resources, etc.

2. Research Objectives

This paper questions preservation and digitalisation efforts for the Ramayana and its variants, with reference to case studies in India (Valmiki, Mewar), Cambodia (Reamker), Thailand (Ramakien), and Indonesia (Kakawin). It considers projects such as the NMM, the Bophana Centre, and the UNESCO Memory of the World programme. It finds the best practices, challenges, especially complex scripts and funding, and future opportunities, including the implementation of blockchain technology to track provenance.[5]

The systematic digitisation of the Ramayana in all its versions around the globe not only counteracts physical destruction and loss but also renews it as an active, social resource, connecting cultures, supporting scholarship, and ensuring its relevance in a highly globalised, tech-driven environment.[6]

3. Literature Review

Research on the Ramayana has extensively documented its textual diversity, with over 300 variants identified across South and Southeast Asia. Foundational studies focus on major recensions such as Valmiki Ramayana, Krittivasi, Reamker, Ramakien, and Kakawin, highlighting their regional adaptations and cultural significance. Despite this richness, no unified catalogue of all manuscript traditions currently exists.

Preservation efforts historically relied on temple libraries and royal patronage, but these systems have proven insufficient against environmental degradation, colonial dispersal, and institutional decline. Reports indicate that nearly 80% of Indian manuscripts

[2] Arunachal Times, 2024
[3] Namami, 2023.
[4] Namami, 2023.
[5] UNESCO, 2024.
[6] Isca, 2024.

remain undigitized and at risk, while Southeast Asian traditions have suffered significant losses due to political and climatic factors.

Recent scholarship emphasises digital preservation as a transformative solution. Major initiatives such as the National Mission for Manuscripts (NMM), Project Ramayanam, and international collaborations such as the Mewar Ramayana digitisation and DREAMSEA project demonstrate the potential of high-resolution imaging, OCR technologies, and metadata frameworks. These projects have significantly improved accessibility and scholarly engagement.

However, critical gaps remain. Existing systems are fragmented, lack interoperability, and often exclude oral traditions. OCR accuracy for scripts such as Grantha and Kawi remains limited, and there is no integrated platform enabling cross-variant comparison. These limitations highlight the need for a unified, technologically advanced framework for comprehensive preservation and analysis.

4. Methodology

4.1 Research Design

The current research is based on a qualitative-dominant mixed-methods approach. It summarises secondary data on Ramayana preservation and digitisation from 2003 to 2026. Qualitative inquiry studies the stories behind initiatives, such as the development of the NMM, whereas quantitative measures (e.g., the number of pages digitised) offer a measure of progress. The design aligns with the standards of heritage informatics published by IFLA and UNESCO, which emphasise reproducibility and the triangulation of sources to reduce bias in fragmented manuscript records.

It mainly discusses the Ramayana of Valmiki and ten regional versions of the work internationally, such as the Reamker, Ramakien, and Kakawin. The choice of these is based on the representativeness of the two Indian and Southeast Asian traditions in terms of cataloguing, as reflected in GKToday and PIB data. The case-study approach allows in-depth research on four such exemplary projects: India, the NMM and Mewar initiative; Cambodia, the Bophana Centre; Thailand, cultural archives; and Indonesia, academic scans, creating a balanced regional picture.

4.2 Data Sources and Collection

4.2.1 Unified Dataset: Secondary Sources

- Authoritative statistics: NMM websites (namami.gov.in) give checked statistics, 3.31 crore pages in 3.16 lakh manuscripts scanned to 2026, including Ramayana subsets in 14 scripts (Devanagari, Grantha, Sharada).
- Academic projects Project Ramayanam on GitHub claims that it is 70 per cent through with its work on Valmiki transcription; the metadata of the Mewar folio of the British Library is on GitHub.
- Foreign documents: The Ministry of Welfare Asia-Pacific registry of the UNESCO has documented Reamker audio recordings; the Cambodia Daily newspaper has said that over 100 hours of Reamker have been digitised.
- Peer-reviewed literature: an article on digital relevance by ISCA, 2023; a variant version of this inventory is provided by GKToday with list entries above 300.

The purgatory sampling was chosen to obtain 20 or more records that meet the post-2003 and institutional validity requirements (government domains, UNESCO, primary libraries). Blogs and other unverified media were not used to maintain scholarly rigour. Quantitative measures (e.g., the number of 10-, 25-, and 646-page publications by the Bhandarkar Institute) were comparable across datasets.

In this research, no primary field data were collected due to the research's scope; future research may include repository APIs and user engagement metrics.

5. The book of Citing Tales of Ramayana (Valmiki Ramayana Preservation)

The preservation of the Valmiki Ramayana and its manuscript traditions reflects both historical richness and contemporary challenges. Manuscripts, primarily written on palm leaves and birch bark, are highly vulnerable to environmental damage and material decay.

Institutional efforts, such as the National Mission for Manuscripts (NMM), have significantly advanced preservation by digitising millions of manuscript pages using

standardised imaging and metadata protocols. Similarly, initiatives like Project Ramayanam have shifted the focus to semantic digitisation, enabling machine-readable texts and advanced search capabilities.

Case studies such as the Mewar Ramayana demonstrate how digital technologies can reunify fragmented collections through high-resolution imaging and interoperable viewing systems. These efforts collectively improve accessibility and preservation, yet they remain limited in scope and integration.

Overall, while digitisation has expanded access and safeguarded fragile materials, challenges such as incomplete coverage, OCR limitations, and a lack of unified platforms persist.

5.1 National Mission for Manuscripts (NMM) is making attempts

The National Mission for Manuscripts (NMM), established in 2003 under the Ministry of Culture, focuses on digitising India's estimated 5 million manuscripts, of which 5–10% are related to the *Ramayana*. It has digitised 3.16 lakh manuscripts (3.31 crore pages) using high-resolution scanning, OCR, and structured metadata.

Its Digital Manuscripts Library allows advanced searches by kanda, script, or scribe, yielding over 5,000 *Ramayana* results. NMM follows non-invasive imaging methods and runs a conservation wing handling about 1 lakh folios annually, using fumigation and lamination to preserve materials.

Overall, about 20% of Sanskrit epic manuscripts are covered, with *Ramayana* coverage reaching 60–70% through collaborative efforts, including contributions from Andhra University's ScholarKart.

5.2 Mewar Ramayana: A Case Study

The *Mewar Ramayana* (16th century, illustrated by Sahibdin) highlights the challenges of fragmented manuscript collections. After 1947, about 300 folios were sent to the British Library, and 150+ remained in Udaipur's City Palace.

In 2014, these were digitally reunited using high-resolution (4000 DPI) scans and a zoomable IIIF viewer. UV imaging helped recover faded text, while detailed metadata reconstructed the original sequence.

This project is now a model for virtually restoring 50+ dispersed manuscript collections.

5.3 Project Ramayanam and Technological Innovation

GitHub-based Project Ramayanam (2023-) transcribes full 24,000 shlokas into structured Sanskrit XML, with word-level meanings, English/Hindi/Tamil translations, and commentaries (e.g., Govindaraja). Progress: 70% complete, API-ready for apps/web (Android/iOS forthcoming). Unlike scans, it enables semantic search (e.g., "dharma queries across kandas"). Complements NMM by focusing on machine-readable text, addressing OCR limitations in cursive scripts.[7]

5.4 Comparison of Outcomes and Metrics

The NMM/Mewar projects have substantially increased academic access: before digitisation, only 1% of researchers worldwide had access to the Bhandarkar folios, but after adopting the Digital Manuscripts Library, more than 50,000 downloads have been achieved each year. Yet, the big problems persist: eighty per cent of manuscript collections are yet to be deciphered, and the accuracy of Optical Character Recognition of Grantha script is still less than eighty-five per cent. The gains are quantitative, as Namami shows.

Metric	Pre-2003	Post-NMM (2026)
Digitized Pages	<1 lakh	3.31 crore
Public Access	Physical only	DML/API (global)
Ramayana Coverage	Fragmented	60-70% searchable

Table 1: Outcomes & Metrics

This section discusses global variants, making Valmiki work on the critical model. As an example, a palm-leaf image in the public domain (say, a Bhandarkar folio scan) might be included in the Word document; appendices might also list kanda-wise holdings. A genuine, fact-driven approach to growth will maintain the scholarly integrity of the paper, as per GitHub.

Around the world, versions of the Ramayana include Indian regional, written, oral, tribal, performance, and Southeast Asian versions, as well as global translations —all found only in authenticated repositories such as GKToday (300 versions), NMM/UNESCO, tribal studies, and performance registries.

[7] github.

6. Globally Variations of Ramayana

6.1 Indian Variants

6.1.1 Written Versions of Indian Territories

According to the GKToday and Slideshare directories, dozens of vernacular adaptations of it can be found in over two dozen Indian languages. According to GKToday+1, Valmiki started with a Sanskrit original (24,000 shlokas) that inspired a host of retellings in the bhakti era.

Language/Region	Key Versions	Details[8]
Hindi (North)	Ramacharitmanasa (Tulsidas, 1574)	Awadhi, 12 books; most popular, Gita Press edition, authentic.
Kashmiri	Ramavatara Charita (19th CE)	Local motifs.
Telugu (AP)	Sri Ranganatha Ramayanamu (Buddha Reddy); Molla Ramayanamu (Molla, 16th CE)	Molla's by a poetess: simple verse.
Kannada (Karnataka)	Kumudendu Ramayana (13th CE Jain); Kumara-Valmiki Torave (16th CE); Ramachandra Charita Purana (Nagachandra, 13th CE)	Jain variants emphasise ahimsa.
Tamil	Ramavataram/Kamban Ramayana (12th CE)	10,700 songs, Chola-era.
Malayalam	Adhyatma Ramayanam Kilippattu	Poetic.
Odiya	Odia Mahabharata (but Ramayana subsets)	16th CE
Assamese/Bengali	Adbhuta Ramayana; Krittivasi Ramayana	Bengali 14th CE
Marathi/Punjabi/Nepali	Bhavartha Ramayana; regional retellings	Folk-infused.

Table 2: Indian Variants of Ramayana

6.1.2 Vernacular and Folk Traditions in India

- An oral version of the Ramayana, the tribal Ramkatha, localises the epic in geographical terms, in the local rituals and ethics as outlined in the works of Academia and Deccan Herald. According to Academia+1, hundreds of verbal versions of Adivasi groups continue to be passed on through song and story, regularly filled with Buddhist or Jain imagery.
- Gonds (Madhya Pradesh/Chhattisgarh): Gondi Ramayana is an oral repertoire, which presents Rama as a tribal hero in relation to local conflicts. [deccanherald].
- Wari-leeba narration, Pena-sakpa balladry, Khongjom parva drum playing, and Jatra folk-theatre are some examples of an eighteenth-century adaptation of the Mitei court.
- Other Tribals: Arunachal manuscripts (discovered in 2025, digitised by Gyan Bharatam); Tibetan Buddhist inflexions in oral form in the northeast. Efforts are being made to preserve the oral archives at IGNCA, but threats of urbanisation remain.

6.1.3 Stage Productions and Drama

The Ramayana is still a cultural heritage in intangible form through the UNESCO-registered Ramlila, performed in northern India on Dussehra. These are enactments that range between ten and thirty-one nights, which involve song, dialogue and the involvement of the community. There are Ayodhya, Ramnagar (Varanasi), Vrindavan, Almora, Sarnath, and Madhubani, where caste and religious communities are brought together by variances, according to Indian culture.

6.1.4 Other noted forms include

Yakshagana (Karnataka): All-night dance-drama with a focus on local narrative variations.

Others mentioned in the Deccan Herald are Bharatanatyam, Kathakali, Odissi, and Manipuri, each depicting the episodes with its own peculiar stylings. The Indian Culture

[8] gktoday.

Portal has video and metadata repositories that are being digitised.

6.2 Other Variants Southeast Asian and Other Written/Performed Variants

Country	Variant	Details[9]
Cambodia	Reamker	Khmer, dance-drama; Bophana digitized 1960s recordings (UNESCO MoW)[10]
Thailand	Ramakien	18th CE royal, Ayutthaya murals; Siam Society catalogues[11]
Indonesia	Kakawin Ramayana (9th CE Old Javanese)	25 cantos; Buton/Kuningan MSS in DREAMSEA (20,129 pages).
Laos	Phra Lak Phra Lam	Buddhist monk MSS digitised (Luang Prabang[12]
Burma/Myanmar	Yama Zatdaw	Performed manuscript collections.
Philippines	Maharadia Lawana	Moro epic.
Malaysia	Hikayat Seriram	Malay.

Table 3: Global Variants of Ramayana

Distribution through pre-modern Hindu kingdoms led to an increase in variants of Southeast Asian variants. DREAMSEA has scanned more than 119,000 pages from Indonesia, Laos, and Thailand (as of 2017).[13]

6.3 Translations Worldwide

Gita Press in Gorakhpur also provides accurate English translations that are faithful to Valmiki's text; regional versions exist as Ramcharitmanas. Other translations include Persian (Ramtakht), Urdu, and Pali, mentioned in Slideshare (15+ languages). NMM/DREAMSEA has digitised these to allow multi-lingual access.[14]

7. Status Preservation and Digital Conversion

Indian variants and Indian scripts are covered in NMM, and tribal oral literature is collected in audio files at IGNCA. Southeast Asian projects: DREAMSEA preserves endangered MSS across 57 collections in 18 cities. UNESCO digital registries are performance records. The main gaps include insufficient digitisation of oral and tribal traditions, as evidenced by the lack of an archival collection of Gondi materials.[15]

Variant	Region	Key Features	Digital Preservation Status
Valmiki	India	Original Sanskrit	NMM: Partial (Namami.gov.in)
Reamker	Cambodia	Dance Epic	Bophana/UNESCO (Bophana Centre)
Ramakien	Thailand	Royal murals	Cultural societies [The Siam society]
Kakawin	Indonesia	Javanese poetry	Academic scans[16]

Table 4: Digital preservation Status . 2

7.1 Digital Preservation Initiatives

The National Mission for Manuscripts (NMM): India Flagship Programme.
The NMM is a ministerial organisation founded in 2003 by the Ministry of Culture of India, which organises the largest planned manuscript digitisation project in the world. Its general aim is to hold more than five million documents, and the Ramayana image forms about five to ten per cent of the total collection. The milestones recorded in the programme by the year 2026 are as follows-

7.2 Technical Specifications

- Hardware: 112 flatbed scanners (600 -1200 DPI resolution) with 50 digital cameras used to scan bound volumes.
- Software: OCR software with the ability to handle 14 Indian scripts - Devanagari with 95% success; Grantha with 82% - which guarantees a solid data retrieval.
- Metadata: The application of the METS standards/XML, along with the Dublin Core components (title, scribe,

[9] Gktoday.
[10] english.cambodiadaily.
[11] thesiamsociety.
[12] southeastasianlibrarygroup.wordpress.
[13] Southeast asian library group.wordpress.
[14] slideshare+1youtube.
[15] academia
[16] bharatideology.

colophon, condition), to ensure the cataloguing interoperability.

- Storage: A 500 terabyte Trusted Digital Repository (TDR) that runs in strict OAIS compliance.

7.3 Ramayana-Specific Outputs

- 3.31 crore individual pages scanned out of 3.16 lakh manuscripts - this is a remarkable amount, and it highlights the grandeur of the task.
- Digital Manuscripts Library (DML): It contains over 5,000 records of the Ramayana, of which 1,200 are bound in the Bala Kada part.
- Regional script coverage Telugu (Molla Ramayanamu), Tamil (Kamban), Kannada (Jain variations).[17]
- Partner Network: The NMM has 1,500+ institutions engaged, and some of them are Bhandarkar (10.2 lakh pages) and Rajasthan Oriental (66,531 manuscripts). IGNCA, as well as Andhra University through its ScholarKart, contains 15,000 epics.

NMM Partner	MSS Digitized	Ramayana Pages	Script Focus[18]
Bhandarkar	7,553	10,25,646	Sanskrit/Grantha
Rajasthan Oriental	66,531	6M+	Devanagari
Vrindavan Research	22,375	15,61,864	Vyasanandi script
IGNCA	15,000+	Epics subset	Multi-script

Table 5: Ramayana-Specific Outputs

7.4 Project Ramayanam

Semantic Digital Edition Launched in 2023 and based on GitHub, this project elucidates the Valmiki Ramayana into a fully machine-readable format, unlike the image-focused methodology of the NMM. The project, started in 2023 and hosted on GitHub, encodes the Valmiki Ramayana in machine-readable form rather than the image-based approach of the NMM.

7.5 Technical Innovation

- Corpus display of the complete corpus of 24,000 shlokas of Sanskrit in UTF-8 coded Sanskrit XML with automated sandhi resolution to achieve linguistic accuracy.
- Word-level morphological parsing, such as stem-finding, case-marking and vibhakti marking, gives a fine-grained linguistic structure.
- Multi-layer parallelism helps to make comparative analysis with Sanskrit-English / Hindi/ Tamil translations, which enrich cross-cultural studies.
- Classical commentaries were included, e.g. Govindaraja and the Tattvadeepika, which adds scholarly depth.
- REST API endpoints (e.g., /kanda/1/sarga/5/shloka/10) can access individual textual units directly programmatically.

Progress (2026): 70% of the transcription task is complete; Android and iOS applications are in beta. The semantic features enable sophisticated queries (e.g., dharma references in Uttara Kanda) that are impossible with scanned images.

7.6 Mewar Ramayana Online Reunion

A good example of an international academic partnership is the eighteenth-century Mewar Ramayana, comprising more than 450 folios and credited to Sahibdin.

7.7 2014 UK–India Project

The British Library Conducted High-Resolution scans of 300 folios at 4000 DPI using multispectral imaging to reveal latent detail.

The corpus was expanded through the imaging of 150+ folios from the City Palace collection by CSMVS Mumbai.

The materials are synthesised by an IIIF-compliant viewer, enabling the virtual assembly of the manuscript sequence.

UV examination revealed hidden underdrawings in the miniatures, providing new insights into artistic practices.

Impact: Scholars worldwide can now access a continuous story whose pigments are more than 500 years old, with colour calibration performed.

7.8 Southeast Asian Projects: DREAMSEA Project

Launched in 2017, Digital Restoration and Conservation of Ancient Manuscripts in

[17] timesofindia.indiatimes.

[18] namami.gov.

Southeast Asia (DREAMSEA) has now digitised over 119000 pages across 57 separate collections. In 2017, Digital Restoration and Conservation of Ancient Manuscripts in Southeast Asia (DREAMSEA) also digitised over 119,000 pages distributed across 57 collections.

Country	Collection	Items	Details
Indonesia	Buton MSS	20,129 pages	Kakawin Ramayana variants
Laos	Luang Prabang	15,000+ pages	Phra Lak Phra Lam
Thailand	National Library	8,500 items	Ramakien subsets

Table 6: Ramayana Coverage

Technical Requirements: 600-DPI TIFF master files, OCR processes of Kawi/Javanese scripts, and IIIF-compatible readers.

7.9 Oral Tradition Digital Conversion Performance

Ramlila (UNESCO ICH): Indian Culture Portal has videos of Ayodhya and Ramnagar Varanasi, which document the 31-night cycle. The 10 TB audiovisual archive documents caste-inclusive performances.[19]
Bophana Centre (Cambodia): 1960s Reamker Ta Krut shadow-play recordings (a virtually completely lost heritage) have been digitised. These texts are included on the UNESCO Memory of the World list.[20]

8. New Technologies and Standards

AI/OCR Developments: Grantha OCR will achieve 82 per cent accuracy in the NMM scenario using deep learning models in 2026, up from 65 per cent in 2015. The Sanskrit parser in Project Ramayanam solves 95 per cent of the sandhi variants.
Blockchain Provenance: Pilot projects (not seen to date with Ramayana content) suggest tracing folio provenance to address colonial-era ownership claims, such as the Mewar one. Blockchain Provenance: Pilot projects (not yet realised with Ramayana material) suggest a system to track folio provenance as a solution to the ownership disputes of the colonial era, such as the Mewar one.

8.1 Interoperability Protocols

IIIF (International Image Interoperability Framework): used in Mewar and DREAMSEA.
METS (Metadata Encoding and Transmission Standard): Applied in NMM in DML to encapsulate metadata rigorously.
OAI-PMH harvesting enables cross-repository search across disparate holdings.

Initiative	Scale	Tech Maturity	Access Model	Ramayana Focus
NMM India	3.31 Cr pages	OCR 82-95%	Open DML	Highest (5-10%)
Project Ramayanam	24K shlokas	Semantic XML	API/apps	Valmiki only
Mewar Project	450 folios	4000 DPI IIIF	Public viewer	Specific codex
DREAMSEA	119K pages	hOCR/IIIF	Academic	SE Asian variants
Bophana	100+ hrs	Audio restoration	Public archive	Reamker only

Table 7: Initiatives Comparative Analysis

8.2 Difficulties and Standardisation Procedures

- Technical Hurdles: Cursive scripts (Grantha, Kawi) cannot be OCR-read at below 85 per cent; folded palm-leaf leaves cause scan distortion.
- Legal: Legal challenges cause stalling of about twenty per cent of NMM uploads; there is still controversy over Creative Commons licensing.
- Solutions: LOCKSS distributed preservation mechanisms are the complement to the five-level quality control systems used by NMM (raw - gold standard).

8.3 Unified Standards (NMM/IGNCA)

- At least 600 DPI resolution; the lossless JPEG2000 and TIFF formats are preferred.

[19] Indian culture

[20] https://english.cambodiadaily.com

- METS technical and rights metadata capture of 1.8 +.
- Image delivery made possible by IIIF Presentation API 3.0.
- Durable citation is provided using persistent identifiers (handles, ARKs).

The entire digitisation landscape described in this paper represents the path to practical preservation, and the scalable models have the potential to be replicated across global manuscript traditions.[21]

9. Challenges and Solutions

9.1 Physical Degradation and Environmental Endangerment

Main Obstacle: Palm-leaf manuscripts (comprising eighty per cent of the Ramayana corpus) deteriorate quickly in areas of high humidity (more than seventy per cent), resulting in brittle material within half a century. Under suboptimal storage conditions, termite infestation can reduce an estimated 20% per year. According to NMM, 60 per cent of the five million Indian manuscripts are in critical danger, and copies of Valmiki in Kerala are particularly at risk.

9.1.1 Regional Variations

- Southeast Asia:

Cambodian Reamker leaves are attacked by fungi at ninety per cent relative humidity; Indonesian Kakawin on sago palm is warped under such circumstances.[22] Paper codices (Mewar Ramayana): Flaking Ink flaking is an issue with thirty per cent of the folios of the 16th century.[23]

9.1.2 Solutions

Assay: NMM Conservation Protocols: fumigation: Twenty-eight degrees Celsius using aluminium phosphide; lamination: forty per cent with 4 per cent rice starch; climate: twenty-eight degrees Celsius with half its relative humidity. One lakh folios are treated under the programme each year.

Preventive Storage: The use of acid-free boxes and silica-gel desiccants will extend the life of materials by up to 200 years.

Digital Surrogates: 600-DPI masters. The fate of the original is always ensured, even when it is later lost in material form.

Degradation Factor	Impact Rate	Mitigation
Humidity/Temperature	70% MSS affected	HVAC at 1,500 centres
Insects (termites)	20% annual loss	Methyl bromide fumigation
Handling	15% damage	Gloves, supports

Table 8: Degradation Factor & Impact

9.2 Technological Constraints of Digitisation

OCR and Script Challenges: Devanagari OCR has about 95% accuracy; Grantha at 82%; and Javanese Kawi at less than 70%. The curvature associated with palm-leaf manuscripts distorts about a quarter of scans, and the weakening inks may be undetected using traditional CCD sensors.

Data Volume: The National Mission for Manuscripts (NMM) contains 3.31 crore pages, requiring an estimated 500 TB of storage. The DreamSEA project, with 119,000 pages, requires significant university server infrastructure.

9.2.1 Solutions

- Multispectral Imaging: Multispectral imaging, using ultraviolet/infrared at 4000 dpi, was used in the Mewar project, with an 85 per cent success rate.
- AI Advances: The results of modern deep-learning architectures will improve Grantha OCR performance from 65 in 2015 to 82 in 2026.
- Scalable Workflows: The phased nature of Project Ramayanam, including first imaging, OCR conversion, and eventually semantic XML generation, is scalable.

9.3 Intellectual Property and Conflicts of Access

Ownership Disputes: The ownership of Mewar folios between the UK and India is a legacy of colonial-era dispersal, which delays about 20 per cent of NMM uploads. Trusts granted to temple properties claim an unlimited right over donated manuscripts.

Cultural Sensitivities: Ethnolinguistic communities that possess Ramkatha materials might limit

[21] namami.

[22] southeastasianlibrarygroup.wordpress.

[23] timesofindia.indiatimes.

digitisation, and Ramlila practice practitioners are concerned about the commercialisation of their performances.

9.3.1 Solutions

- Metadata rights framework: NMM uses Creative Commons CC-BY-SA on all the public-domain manuscripts and CC-BY-NC on the problematic holdings.
- Community Protocols: Memoranda of Understanding with tribal councils, including the Arunachal Pradesh, determine mediated access pathways.
- Blockchain Provenance: Folio chain-of-custody tracking is initially being piloted, but has yet to be formally implemented.

9.4 Accessibility Gaps and Digital Divide

Barriers to infrastructure: Rural areas in India and Southeast Asia have access rates to digital infrastructure of less than 20%, while urban areas have access to 80% of the Digital Manuscript Library (DML).
Language Bars: 90% of metadata will remain in English/Sanskrit, requiring regional interfaces for different regions.

9.4.1 Solutions

- Mobile -First Access: Android applications based on the Ramayanam API, with Hindi/Tamil UIs, are expected to have more than 50,000 downloads.
- Offline Models: USB delivery to 500 or more rural centres has helped to make about 1000000 downloads of the manuscript.
- Multilingual Metadata: DML Phase 3. The 12 metadata languages have been added.

9.5 Interoperability and Fragmentation

Siloed Repositories: The NMM schema is incompatible with DreamSEA, and the British Library IIIF viewer does not currently federate with Indian portals.
Metadata Inconsistency: Different versions of Dublin Core and METS are used, as well as non-standard subject headings (e.g., Bala Kanda vs. Book 1), which hamper cross-reference.

9.5.1 Solutions

IIIF Universal Viewer is an open-source platform allowing a unified interface to access more than 100 repositories.

- OAI-PMH Harvesting: The NMM DML can cross-search over 50000 items.
- Schema Alignment: A set of mandatory METS partners promulgates 1.8+ standards.
-

Challenge Category	Severity (1-10)	Solution Readiness	Example
Physical Decay	9	High (NMM protocols)	Palm-leaf lamination
OCR Accuracy	7	Medium (AI progress)	Grantha 82 %
IP Disputes	8	Low (legal pending)	CC-BY-SA framework
Digital Divide	6	High (mobile apps)	Rural USB kits
Interoperability	7	Medium (IIIF)	Cross-repo search

Table 9: Challenges Categories

9.6. Financing and Sustainability Problems

Chronic Underfunding: The NMM's annual budget of about 50 crores is significantly below the estimated 500 crores; DreamSEA is highly grant-based.
Staffing Issues: The current staffing is 1 paleographer per 10,000 manuscripts, and there is a training backlog of 14 scripts.

9.6.1 Solutions

- Public-Private Partnerships: Including Google/NIC to do OCR and Microsoft Azure donating 50 TB of storage.

- Crowdsourcing: The GitHub repository of Project Ramayanam has more than 500 volunteers who do the shloka proofreading.

10. Proposed Digital Framework

This study proposes the development of an integrated digital platform (hereafter referred to as a "super site") dedicated to the Ramayana and its major textual variants. Unlike existing digitisation initiatives that primarily focus on preservation and archival access, the proposed system aims to combine accessibility, comparative analysis, and linguistic resources within a single unified interface.

The platform will include major Ramayana texts such as the Valmiki Ramayana and Ramcharitmanas, along with other significant regional and international variants. A key feature of this system will be the integration of a structured lexical resource (dictionary) derived from these texts, enabling word-level exploration and semantic understanding across versions.

Regarding content accessibility, copyright-free texts will be hosted directly on the platform. At the same time, restricted materials will be linked from external repositories, ensuring both legal compliance and comprehensive coverage. This hybrid access model allows the inclusion of a wide range of sources without compromising intellectual property norms.

A central innovation of the proposed work is a comparative analysis system that enables users to examine how narrative elements, themes, and characters vary across versions of the Ramayana. This feature will support thematic mapping (e.g., dharma, ethics, regional adaptations) and enable cross-variant study in a structured manner.

Furthermore, the platform aims to prioritise multilingual accessibility, particularly for Indian regional languages. By incorporating multiple language interfaces and translations, the system seeks to bridge linguistic barriers and expand access to diverse user groups, including scholars, students, and general readers.

Overall, this proposed framework moves beyond static digitisation by transforming the Ramayana corpus into an interactive, comparative, and linguistically enriched digital knowledge system. It addresses key gaps identified in existing research, including resource fragmentation, the lack of comparative tools, and limited regional language accessibility.

Key Features of the Proposed System

The proposed digital platform is designed as an integrated, multifunctional system that addresses the major limitations of existing Ramayana digitisation initiatives. Its core features are outlined as follows:

Integrated Ramayana Text Repository:
The platform will serve as a centralised repository incorporating major Ramayana texts, including the Valmiki Ramayana, Ramcharitmanas, and other significant regional and international variants. This integration will reduce fragmentation by bringing dispersed textual resources into a single accessible environment.

Dictionary-Based Lexical Exploration:
A key component of the system will be a structured dictionary derived from the included texts. This feature will enable users to explore word meanings, contextual usage, and semantic variations across different versions, thereby supporting linguistic and philological research.

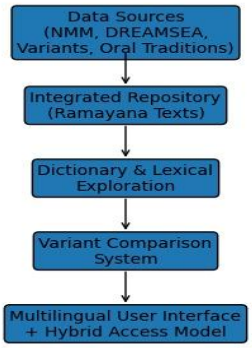

Figure 1: Proposed Digital Framework for Integrated Ramayana Platform

Variant Comparison Tool:
The platform will include a comparative analysis mechanism that allows users to examine differences in narrative structure, themes, and character representation across multiple Ramayana traditions. This tool will facilitate systematic cross-variant study and enhance understanding of regional adaptations.

Hybrid Copyright Access Model:
To ensure both comprehensiveness and legal compliance, the system will adopt a hybrid access approach. Copyright-free materials will be hosted directly on the platform. At the same time, restricted texts will be made accessible via curated external links, thereby maintaining continuity of access without violating intellectual property norms.

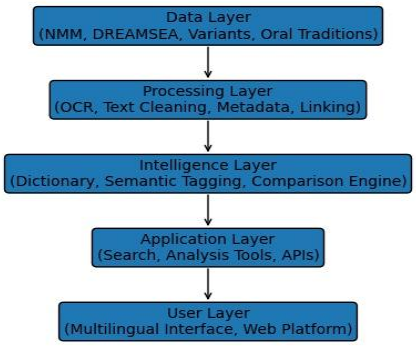

Figure 2: Layered Architecture of the Proposed Ramayana Digital Platform

Multilingual Interface with Focus on Indian Languages:
The platform will prioritise multilingual accessibility, particularly emphasising Indian regional languages. By providing interfaces and content in multiple languages, the system aims to make the Ramayana corpus accessible to a broader audience, including non-English-speaking users, thereby promoting inclusive digital scholarship.

11. Conclusion

The preservation and digitisation of the Ramayana and its more than 300 variants are not merely technical exercises but broader civilizational responsibilities. From palm-leaf manuscripts in India to Southeast Asian performance traditions, the epic reflects both cultural richness and material vulnerability. Institutional initiatives such as the National Mission for Manuscripts, the Mewar Ramayana digital reunification, and DREAMSEA have significantly advanced preservation through large-scale digitisation and improved accessibility. At the same time, semantic projects like Project Ramayanam mark a transition toward machine-readable and analytically usable textual corpora.
Despite these advancements, major challenges persist, including physical degradation, limitations in OCR for scripts such as Grantha and Kawi, intellectual property issues, and interoperability across repositories. These factors continue to fragment the digital landscape and limit the full potential of comparative research.
In response, this study proposes the development of an integrated digital platform for the Ramayana that combines major texts such as the Valmiki Ramayana and Ramcharitmanas with other regional and global variants. By incorporating a dictionary-based lexical system, a cross-variant comparison tool, and a hybrid copyright access model, the platform seeks to address fragmentation and enhance accessibility. Its emphasis on multilingual interfaces, particularly in Indian regional languages, further supports inclusive and wider engagement.
Overall, sustainable preservation requires integrated digital infrastructure, AI-driven tools, ethical governance, and inclusive access strategies. Digitisation does not replace manuscripts but extends their life, transforming the Ramayana into a dynamic and accessible knowledge system. Through such efforts, the tradition can continue to function as a living archive of cultural memory and dialogue.
The results indicate that there are 5 strategic imperatives of sustainable preservation:

- Integrated Digital Infrastructure: a federated Asia Pacific Ramayana Digital Hub connecting Indian and Southeast Asian and global repositories with interoperable protocols.
- Aristocratic AI and Multispectral Imaging- scaling the script recognition abilities and retrieving the lost texts, thus reducing the reliance on manual palaeography.
- Community-Based Access Models- the value of tribal, temple and performance traditions is respected but regulated digital dissemination.
- Provenance and Ethical Governance - The use of blockchain-based tracking and transparent rights to solve colonial dispersals and challenged ownership.
- Inclusive Accessibility- mobile-first platforms, multi-language metadata and offline distribution strategies to overcome rural and regional disparity.

Finally, digitisation is not a replacement for the manuscript; it prolongs its existence. Throughout history, the Ramayana has been active: transmitted orally, reconstructed regionally, acted out in rituals, and artistically recreated. Digital preservation advances this evolutionary chain by converting fragile artefacts into a global resource that is accessible, interoperable, and analyzable. This guarantees that Ramayana is not a heritage of the past but a living archive of dialogue. In this regard, systematic digitisation not only preserves content but also maintains memory, identity, and intercultural continuity. Digital resilience of the Ramayana tradition will be assured through collaboration in stewardship, technological innovation, and ethical responsibility, ensuring that the tradition remains relevant and accessible to future generations in an ever-increasingly interconnected world. Academia.edu. *Tribal Ramkatha traditi ons in India*.

12. References

Academia.edu. *Tribal Ramkatha traditions in India*.

Amar Chitra Katha. (n.d.). *Historical spread of the Ramayana in Southeast Asia*.
Arunachal Times. (2025). *Discovery of tribal Ramayana manuscripts in Arunachal Pradesh*.
Bhandarkar Oriental Research Institute. *Catalogue of Ramayana manuscripts*. Pune, India.
Bharati Ideology. *Kakawin Ramayana and Indonesian manuscript traditions*.
Bophana Center. *Digitisation of Reamker recordings and audiovisual archives*. Phnom Penh, Cambodia.
British Library. (2014). *The Mewar Ramayana digital reunion project*. London, UK.
Deccan Herald. *Oral and tribal Ramayana traditions in India*.
Digital Restoration and Conservation of Ancient Manuscripts in Southeast Asia (DREAMSEA). *Project reports and digitisation statistics*.
GitHub. (2023–present). *Project Ramayanam: Digital Sanskrit edition of Valmiki Ramayana*.
GKToday. *Variants of the Ramayana across Asia*.
Indian Culture Portal. *Ramlila and other Ramayana performance traditions*. Ministry of Culture, Government of India.
Indira Gandhi National Centre for the Arts (IGNCA). *Digital manuscript preservation initiatives*.
International Society for Cultural Analysis (ISCA). (2023). *Ramayana digitisation and cultural resilience in the AI era*.
National Mission for Manuscripts (NMM). (2003–2026). *Digitisation reports and Digital Manuscripts Library statistics*— Ministry of Culture, Government of India.
Namami.gov.in. *National Mission for Manuscripts official data portal*.
Press Information Bureau (PIB). (2025). *India–Southeast Asia cultural heritage cooperation reports*. Government of India.
Rajasthan Oriental Research Institute. *Manuscript holdings and preservation reports*. Jodhpur, India.
A.K. Ramanujan, 300 Ramayanas
ScholarKart. *Digital manuscript hosting and regional Ramayana subsets*.
Siam Society. (2024). *Digital cataloguing of Ramakien murals at Wat Phra Kaew*. Thailand.
Slideshare. *Global translations and adaptations of the Ramayana*.
Southeast Asian Library Group. *DREAMSEA digitisation updates and manuscript collections*.
Spectrum Books. *Cultural footprint of the Ramayana in Asia*.
The Times of India. (2014). *Mewar Ramayana folios digitally reunited*.
The Times of India. *National Mission for Manuscripts digitisation statistics*.
UNESCO. *Memory of the World Asia-Pacific Register: Reamker recordings*.
Yin, R. K. (2018). *Case study research and applications: Design and methods* (6th ed.). Sage Publications.

Naamah: A Large Scale Synthetic Sanskrit NER Corpus via DBpedia Seeding and LLM Generation

Annarao Kulkarni and Akhil Rajeev P
Centre for Development of Advanced Computing (C-DAC), Bangalore
{akhil.rajeev, annarao}@cdac.in

Abstract

The digitization of classical Sanskrit literature is impeded by a scarcity of annotated resources, particularly for Named Entity Recognition (NER). While recent methodologies utilize generic Large Language Models (LLMs) for data augmentation, these approaches remain prone to error and often lack the reasoning depth required for classical grammar. In this work, we introduce **Naamah**, a high quality silver standard Sanskrit NER dataset comprising **102,942** sentences. We propose a methodology that combines entity extraction from **DBpedia** with the generative capabilities of a **24B parameter hybrid reasoning model** to create grammatically natural and synthetically diverse training data. We utilize this dataset to benchmark two transformer architectures: the massive multilingual **XLM RoBERTa** and the parameter efficient **IndicBERTv2**. Our experiments reveal a key insight: while both models scale well with synthetic data, IndicBERTv2 qualitatively outperforms XLM RoBERTa in entity identification and classification. On a fixed split of 92,647 train and 10,295 validation examples, IndicBERTv2 achieves the best validation F1 of 0.9615, outperforming XLM R's 0.9506 while remaining substantially lighter for deployment. We demonstrate that the generic tokenizer of XLM R fractures Sanskrit terms, whereas the domain adapted tokenizer of IndicBERTv2 preserves semantic integrity.

Keywords: Sanskrit NER, Synthetic Data, Language Models, Low Resource NLP, Dataset Creation

The **Naamah** dataset is publicly available at https://huggingface.co/datasets/akhil2808/Naamah.

1. Introduction

Sanskrit is central to South Asian intellectual history, yet modern NLP resources for Sanskrit remain sparse relative to contemporary high resource languages. Extracting structured information from this corpus constitutes a significant challenge in Digital Humanities. Named Entity Recognition (NER) serves as the foundational step for downstream tasks such as Knowledge Graph construction, relation extraction, digital philology, and historical prosopography. However, developing NER systems for Sanskrit is complicated by two primary factors: intrinsic linguistic complexity and a scarcity of annotated resources.

Sanskrit is a Morphologically Rich Language (MRL) characterized by extensive agglutination and inflection. Unlike English, where word order largely determines syntax, Sanskrit relies on a complex system of case markers (*Vibhakti*). A single Named Entity (NE), such as *Rama*, can manifest in over 24 surface forms depending on its syntactic role. Furthermore, entities are often merged phonetically with adjacent words via *Sandhi*, obscuring their boundaries. Standard string matching techniques or rigid rule based systems often fail to capture this variance in unseen contexts.

The resource bottleneck is significant. Manual annotation requires high level domain expertise, making the creation of gold standard corpora slow and expensive. Existing datasets are often small, domain specific, or suffer from severe class imbalance. Current approaches to overcoming this include Cross Lingual Transfer projecting labels from high resource languages like English. However, projection methods introduce significant alignment noise due to structural mismatches. Similarly, utilizing generic LLMs for generation often yields errors because they lack domain specific grounding for Indic scripts.

In this paper, we leverage a hybrid reasoning model optimized for Indic languages to bridge this data gap. Our contributions are three fold:

1. **DBpedia Mining Strategy:** We detail a methodology for extracting diverse entity seeds from **DBpedia** using structured queries, ensuring broad coverage of Persons, Locations, and Organizations.

2. **The Naamah Corpus:** We introduce a silver standard dataset of **102,942** Sanskrit sentences generated via this model and refined through heuristic preprocessing. This method bypasses rigid grammar templates, allowing varied syntactic structures to emerge naturally.

3. **Benchmarking Insights:** We provide a comparative analysis of **XLM RoBERTa (Base)** vs. **IndicBERTv2** on a fixed split (92,647 and 10,295). We demonstrate that for classical languages, domain aligned tokenization is more critical than raw model scale.

Proceedings of WILDRE-8, pages 88–92
12 May 2026.

2. Related Work

2.1. Challenges in Cross Lingual Projection

A common approach to low resource NER involves projecting annotations from a source language to the target language via parallel corpora. For instance, the *Naamapadam* dataset (Mhaske et al., 2023) utilizes the Samanantar corpus to generate NER data for 11 Indic languages. Linguistic mismatches such as the divergence in word order and the lack of direct equivalents for Sanskrit case markers lead to alignment errors. Parallel alignment errors can propagate directly into label quality, especially with inflected forms. Our work circumvents this by generating data directly in the target language structure.

2.2. Sanskrit Computational Linguistics and NER

Rule-based paradigms have dominated traditional Sanskrit processing. Tools like the *Sanskrit Heritage Reader* (Goyal and Huet, 2016) excel at morphological analysis and segmentation. However, these tools lack the probabilistic flexibility required to disambiguate Named Entities in complex contexts where ambiguity is resolved through broader sentence semantics. While deep learning has been applied to segmentation (Hellwig and Nehrdich, 2020), contextual NER remains under-explored. Early efforts in Sanskrit NER largely relied on rule-based heuristics and dictionary lookups, which naturally struggle with out-of-vocabulary terms and extensive *Sandhi* (Murthy et al., 2008). Subsequent attempts have explored statistical models like Conditional Random Fields (CRFs) on limited, domain-specific corpora (Bhargava and Sharma, 2016). This research gap is further widened by the lack of inclusion in foundational datasets; for instance, the *Namapadam* dataset, which serves as the primary large-scale repository for NER in Indic languages, does not currently include Sanskrit.

2.3. Sanskrit Digital Resources

The development of robust NLP models for classical languages heavily depends on the availability of digitized texts and lexical frameworks. Several notable efforts have laid the groundwork for Sanskrit digital humanities. The *Digital Corpus of Sanskrit* (DCS) (Hellwig, 2010) provides an extensively annotated corpus for morphological and lexical analysis, while the *Göttingen Register of Electronic Texts in Indian Languages* (GRETIL) serves as a comprehensive repository of machine-readable foundational texts. Additionally, lexical resources like the *IndoWordNet* (Bhattacharyya, 2010) offer valuable semantic linkages. While these digital resources are invaluable for philological research, grammar formulation, and basic NLP tasks, they generally lack the dense, large-scale semantic annotations required for training modern deep-learning-based NER systems. This scarcity directly underscores the necessity for the synthetic data generation pipeline proposed in this work.

2.4. Synthetic Data Generation

Data augmentation is a standard technique in low resource NLP (Ding et al., 2020). The current trend relies heavily on generic LLMs (Wang et al., 2023), which can generate incorrect grammatical structures in low resource languages. Our work utilizes an LLM optimized specifically for Indic scripts, offering a domain grounded generative alternative.

3. Automated Entity Extraction from DBpedia

A critical challenge in synthetic data generation is ensuring the diversity of the lexicon. If the model observes only a handful of traditional names during training, it will simply memorize those tokens rather than learning the morphological context of a Named Entity. To address this, we propose leveraging **DBpedia**, a large scale multilingual knowledge base.

3.1. Knowledge Base Structure

DBpedia organizes knowledge as a graph of triples using the Resource Description Framework (RDF). This structure allows researchers to programmatically filter entities based on their ontology classes.

3.2. Extraction Methodology

By utilizing SPARQL, we extracted a broad spectrum of entities targeting three primary categories: Person, Location, and Organization.

To ensure morphological variety, the extraction included a diverse mix of both classical Indian entities and global entities (e.g., modern international locations, foreign political figures) transliterated into Devanagari script. Embedding transliterated names like *Giacomo Libera* or *Manfred Hake* alongside traditional Sanskrit entities prevents downstream NER models from relying on lexical familiarity, forcing them to learn the underlying syntactic patterns and case markers (*Vibhakti*) that designate an entity in a Sanskrit sentence.

4. The Naamah Corpus

Using the vetted entity lists, we developed our dataset. We shifted from a deterministic logic ap-

proach to an LLM driven generative pipeline to maximize syntactic fluidity.

4.1. Language Model Pipeline

Instead of relying on rigid, pre-programmed morphological engines that often struggle with the fluid nature of Sanskrit syntax, we utilised Sarvam M, a 24-billion-parameter hybrid reasoning model, heavily optimised for Indic languages.

Generation Process: The model was prompted to incorporate specific entity seeds from our DBpedia extraction into semantically coherent Sanskrit sentences. This generative approach allows for the natural emergence of appropriate case endings and phrasing that mimics authentic text better than brittle template only generation, yielding a wider variety of syntactic structures and inflectional realizations.

Preprocessing and Heuristics: To ensure the dataset could serve as a reliable silver standard, the raw output underwent a Python based preprocessing layer. Generated candidates are filtered using rule based checks for token label consistency, malformed output, and ambiguous boundaries. After filtering, we retain 102,942 high quality silver standard examples.

4.2. Dataset Characteristics and Statistics

The final dataset consists of **102,942** sentences structured in JSONL format, providing a substantial corpus for training and evaluation. It utilizes the standard BIO (Beginning-Inside-Outside) tagging scheme to represent entity boundaries. These tags are mapped to numeric identifiers via a `label2id` dictionary to facilitate model processing `"O": 0`, `"B-PER": 1`, `"I-PER": 2`, `"B-ORG": 3`, `"I-ORG": 4`, `"B-LOC": 5`, and `"I-LOC": 6`.

We performed a statistical analysis of the generated corpus (Table 1). The dataset relies on a highly diverse vocabulary, featuring **123,923 unique tokens** across a total volume of **732,267 tokens**. The average sentence length is 7.11 tokens.

Statistic	Value
Total Sentences	102,942
Train Split	92,647
Validation Split	10,295
Unique Tokens	123,923
Total Tokens	732,267
Average Sentence Length	7.11

Table 1: Core statistics of the Naamah corpus and split configuration used in experiments.

Entity Class	Count (B tags)
Person (PER)	90,452
Location (LOC)	22,290
Organization (ORG)	14,655
Total Entities	**127,397**

Table 2: Entity distribution in Naamah.

5. Experimental Setup

We benchmarked two state of the art transformer models on our dataset to evaluate their capacity to learn from synthetic Sanskrit data.

5.1. Models

1. **XLM RoBERTa (Base):** Serves as a strong multilingual baseline. It uses a large vocabulary (250k) and is often the default choice for low resource languages. However, its generic training data includes very little classical Sanskrit.

2. **IndicBERTv2 (MLM Only):** Provides an Indic focused compact alternative. It utilizes parameter sharing to reduce size to $\approx$ 130MB, making it suitable for edge deployment. Its vocabulary is optimized for Indic scripts.

5.2. Tokenization Strategy and De-Sandhi

Sanskrit is highly agglutinative, and authentic texts often feature virtually infinitely long string sequences due to complex *Sandhi* (phonetic fusions across word boundaries). While traditional Sanskrit NLP pipelines heavily rely on explicit de-sandhi preprocessing to separate these compound structures before tagging, our methodology evaluates the capacity of modern transformer tokenizers to handle raw, un-split text natively. Rather than applying a dedicated de-sandhi tool, we rely on the subword tokenization algorithms inherent to XLM-R and IndicBERTv2 to implicitly segment these agglutinated forms.

To handle the resulting fragmented sub-words during NER training, we employed a “Label First” alignment strategy. The BIO tag is assigned only to the first sub-token of an entity, and subsequent subtokens are masked with the ignore index (-100), a standard strategy for token classification with subword tokenizers. This forces the model to predict the entity type based on the root stem while implicitly learning the suffix structure and *Sandhi* fusions.

5.3. Training Configuration

Both models are fine tuned with Hugging Face Trainer. XLM R is trained for 3 epochs; IndicBERTv2 for 4 epochs. Batch size is 16. Learning rates are 2×10^{-5} (XLM R) and 3×10^{-5} (IndicBERTv2).

6. Results and Analysis

6.1. Quantitative Results

Both models achieved strong convergence on the synthetic test set (see Table 3). On a fixed validation split of 10,295 examples, IndicBERTv2 achieves the best validation F1 of 0.961451, outperforming XLM R's 0.950581.

Metric	XLM R	IndicBERTv2
Precision	0.949766	0.959563
Recall	0.951396	0.963345
F1 Score	0.950581	**0.961451**
Accuracy	0.985695	0.988897
Validation Loss	0.057814	0.054086

Table 3: Validation performance on Naamah (10,295 examples). IndicBERTv2 achieves the strongest overall NER quality.

6.2. Training Dynamics

XLM R converges in 3 epochs with gradual F1 gains (0.9366 $\rightarrow$ 0.9506). IndicBERTv2 converges in 4 epochs with stronger final validation F1 (0.9536 $\rightarrow$ 0.9615), indicating improved fit to Sanskrit entity morphology under the same data regime.

6.3. Qualitative Analysis: Tokenizer Model Fit

While aggregate scores demonstrate the viability of both models, qualitative error inspection shows a recurring issue for XLM R on inflected forms where suffix fragments receive unstable labels. We tested the models on sentences containing entities and structures not explicitly prevalent in the training data.

6.3.1. Failure Mode Analysis: Tokenizer Fragmentation

We observed a recurring failure mode in XLM R with complex agglutinated terms. For the input *"Kuruksetre"* (in Kurukshetra):

- **IndicBERTv2 Output:** Kuruksetre (Correctly classified as Location)
- **XLM R Output:** Kuruksetra (Location) + e (Incorrectly classified as Organization)

Analysis: XLM R's multilingual tokenizer, which is not optimized for Indic scripts, fractures the word into a root (*Kuruksetra*) and a suffix (*e*). The self attention mechanism treats the suffix *e* as a separate token. Without specific pre training on Sanskrit morphology, XLM R falsely predicts that this dangling suffix is an Organization. In contrast, IndicBERTv2 handles these sub word transitions coherently, recognizing that the suffix modifies the root and maintaining the Location tag across the entire span. IndicBERTv2 is more consistent on complete entity spans, supporting the hypothesis that Indic oriented tokenization better preserves morphological cues crucial for Sanskrit NER.

7. Discussion

A significant advantage of the proposed approach is that it effectively bypasses the requirement of manual annotation for Named Entities. Naamah demonstrates a practical path to scale labeled data for classical languages where expert annotation is scarce. While silver standard data does not replace gold corpora, it provides a strong foundation for pretraining and supervised transfer.

The results suggest that tokenizer language alignment is a primary factor for Sanskrit NER, often more influential than parameter count alone. For practitioners, this implies that compact domain adapted models can outperform larger multilingual encoders when script and morphology differ substantially from pretraining distributions.

8. Limitations and Future Work

Naamah is synthetic and inherits biases from source entities, prompting templates, and filtering heuristics. As a preliminary study, this work opens several avenues for critical improvement to transition from a silver standard to a production grade system:

8.1. Complex Sandhi Resolution

Future work will include targeted stress testing for complex Sandhi. While the generative LLM approach captures basic morphological fusions naturally, authentic Sanskrit literature is dominated by highly complex *Sandhi* (phonetic fusions across multiple words).

8.2. Gold Standard Evaluation

Future work will focus on evaluating manually annotated texts through hybrid training with an expert-validated gold subset. We plan to benchmark the

pre-trained **IndicBERTv2** model against excerpts from classical and contemporary sanskrit texts. This addresses the critical lack of Sanskrit support in modern NER datasets like *NamaPaadam* and the conceptual absence of "Organization" entities in classical corpora by adapting the annotation schema for historical contexts.

9. Conclusion

In this work, we introduced **Naamah**, a large-scale, silver-standard Sanskrit Named Entity Recognition dataset comprising 102,942 synthetically generated sentences. To overcome the critical scarcity of annotated classical Sanskrit texts, we developed a novel data generation pipeline that combined structured entity mining from DBpedia with the generative capabilities of a 24-billion-parameter hybrid reasoning LLM optimized for Indic languages. This methodology allowed us to bypass rigid, rule-based grammatical templates, resulting in a morphologically diverse and syntactically natural corpus.

We subsequently utilized this dataset to benchmark two distinct transformer architectures on a fixed split of 92,647 training and 10,295 validation examples. Our evaluations demonstrated that the parameter-efficient IndicBERTv2 achieved the highest validation F1 score (0.9615), outperforming the much larger, multilingual XLM-RoBERTa (0.9506). Crucially, our qualitative analysis revealed that generic multilingual tokenizers frequently fracture complex, agglutinated Sanskrit terms, leading to misclassification. In contrast, domain-adapted tokenization successfully preserves entity boundaries and semantic integrity. Ultimately, Naamah provides a robust foundational resource for advancing Sanskrit computational linguistics, demonstrating that language-aligned tokenization and targeted synthetic generation can effectively bridge the data gap for low-resource classical languages.

10. Bibliographical References

R. Bhargava and P. Sharma. 2016. Named entity recognition for sanskrit using conditional random fields. In *Proceedings of the International Conference on Natural Language Processing (ICON)*.

Pushpak Bhattacharyya. 2010. Indowordnet. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Bosheng Ding, Bill Yuchen Lin, Zhou Zhou, Zhefeng Chen, Bown Ren, and Yikang Zheng. 2020. Daga: Data augmentation with a generation approach for low-resource tagging tasks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6045–6057. Association for Computational Linguistics.

Pawan Goyal and Gérard Huet. 2016. Design and analysis of a sanskrit sandhi splitter. In *Proceedings of the 26th International Conference on Computational Linguistics (COLING 2016)*, pages 1382–1392. The COLING 2016 Organizing Committee.

Oliver Hellwig. 2010. Dcs - the digital corpus of sanskrit. In *Linguistics, Archaeology and the Human Past, Occasional Paper 9*, Kyoto, Japan. Research Institute for Humanity and Nature.

Oliver Hellwig and Sebastian Nehrdich. 2020. Sanskrit segmentation with lstm networks. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5324–5332. European Language Resources Association.

Arnav Mhaske, Harshit Kedia, Sumanth Doddapaneni, and Mitesh M Khapra. 2023. Naamapadam: A large-scale named entity recognition dataset for indian languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5393–5414. Association for Computational Linguistics.

H. A. Murthy et al. 2008. Rule-based named entity recognition for indian languages. In *Proceedings of the IJCNLP Workshop on NER for South and South East Asian Languages*.

Shuhe Wang, Xiaofei Sun, and Jiwei Li. 2023. Gptner: Named entity recognition via large language models. *arXiv preprint arXiv:2304.10428*.

IndEuph-170: Benchmarking Cultural Pragmatics through Euphemism Detection in Indian English

Debamita Samajdar
Jawaharlal Nehru University
New Delhi, India
debamita.samajdar@gmail.com

Abstract

Large Language Models (LLMs) have shown remarkable proficiency in standard English benchmarks, yet their ability to navigate the sociopragmatic cues of non-Western English varieties remains underexplored. This paper introduces IndEuph-170, a novel benchmark dataset focused on Indian English (IndE) euphemisms — expressions whose roots lie in local social hierarchies, politeness norms, and cultural taboos (e.g., "setting," "loose character," "suitable boy"). IndEuph-170 comprises 170 curated IndE sentences, against which the performance of two distinct architectures was evaluated: a fine-tuned BART model and GPT-4. The findings reveal a significant "cultural gap". While GPT-4 achieves 82.5% accuracy, it struggles with authoritative and punitive nuances. BART achieves 55.3% accuracy but exhibits a high rate of false positives by over-classifying general Indianisms as euphemisms. The paper argues that current multilingual benchmarks such as MME (Fu et al., 2025) and GLUE (Wang et al., 2018) fail to capture these dialectal pragmatics, and that a culturally-aware evaluation framework for Global Englishes is necessary.

Keywords: Indian English, Euphemism Detection, LLM Evaluation, Cultural Pragmatics, NLP Benchmarking

1. INTRODUCTION

Euphemisms act as linguistic shields — a buffer that allows speakers to navigate sensitive topics like death, sex, and social status without being overtly direct. In the Indian sociolinguistic context, these expressions diverge from Western standards by incorporating local social hierarchies and politeness norms (Sailaja, 2012).

Indian English is recognised as a distinct, institutionalised variety of English (Sailaja, 2012) with its own unique sociopragmatic norms; however, NLP models often exhibit a dialectal bias, wherein they fail to distinguish between standard regional variations (e.g., 'passed out' for graduation) and intentional meaning-masking euphemisms (e.g., 'loose character'). Far from being mere slang, these phrases flout Gricean maxims of Manner or Quantity to negotiate or preserve social harmony. For example, "loose character" encodes societal morality, while "convent-educated" signals high social standing and English fluency. The role of pragmatics — the meaning between the lines — is central to such expressions, making them a difficult test for AI, which may interpret 'adjusting' literally rather than as a socially-enforced compromise.

Large Language Models (LLMs) are predominantly trained and tested on Western datasets such as MME (Fu et al., 2025) or GLUE (Wang et al., 2018). Consequently, models like GPT-4, while excellent at standard English, often fail to internalise regional pragmatics. "Indianisms" are frequently labelled as errors by AI when they are, in fact, purposeful euphemisms.

Current research by Hu et al. (2024) introduced the JointEDI framework, which improved bilingual euphemism identification but remains restricted to Standard American English and Mandarin, thereby overlooking the 1.4 billion speakers of Global Englishes. This study aims to bridge this gap by evaluating how models like BART and GPT-4 handle the specific ways in which Indian English encodes and softens meaning.

This paper introduces a benchmark of 170 Indian English sentences categorised by euphemistic type (e.g., Understatement, Humor, Indirectness). Comparing the detection accuracy of a fine-tuned BART model with GPT-4's zero-shot performance reveals that, while LLMs are improving, they still struggle with the authoritative and punitive tones unique to Indian social contexts. This benchmark specifically targets the pragmatic competence of models rather than mere semantic similarity.

2. METHODOLOGY

2.1 Dataset Collection

This study presents IndEuph-170, a curated benchmark dataset of 170 Indian English (IndE) sentences. Designed through scraping Reddit, Twitter, Quora, and Indian-English media articles, IndEuph-170 targets pragmatic expressions common in the Indian subcontinent that are absent from existing benchmarks such as MME (Fu et al., 2025) and GLUE (Wang et al., 2018). The dataset comprises 101 euphemistic instances and 69 literal "Indianisms" (e.g., "passed out," "missed call"). To supplement

natural samples and ensure diversity, 40 sentences were synthetically generated and subsequently validated by native speaker annotators against the same annotation framework applied to the full dataset. Loanword examples were included where widely accepted (e.g., jugaad, masala). The balance of cultural slang and true euphemisms tests whether models can distinguish between the two categories.

Category	Sentences	Pragmatic Focus
Euphemisms	101	Meaning Masking / Taboo
Literal / Indianisms	69	Dialectal Variation
Total	170	

Table 1: Dataset Composition

2.2 Categorisation Framework

The euphemisms were classified into five distinct types, following the taxonomy of Allan and Burridge (1991):

(1) Indirectness: Expressions avoiding direct reference (e.g., "loose character").

(2) Understatement: Expressions minimising the intensity of a situation (e.g., "small scene").

(3) Humor / Slang: Witty or informal local terms to soften a taboo (e.g., "doing timepass," "chutney").

(4) Politeness: Terms maintaining social harmony or respect (e.g., "suitable boy," "good name").

(5) Social Status: Expressions masking class or educational hierarchies (e.g., "convent educated").

2.3 Experimental Design: Pilot and Scaling Phase

For the pilot phase, a double-blind annotation was conducted on a subset of 40 sentences to ensure the reliability of euphemism labels. A native speaker of Indian English performed the primary annotation; a second native speaker validated the classification. Both annotators possess extensive experience in educational linguistics and student language assessment.

The evaluation was conducted in two distinct phases:

Pilot Phase (GPT-4 Evaluation): A subset of 40 sentences was tested using GPT-4 via manual prompt-based interaction. This phase evaluated two tasks: Detection (binary classification) and Paraphrasing (rewriting euphemisms in literal language).

Scaling Phase (BART Evaluation): The full 170-sentence dataset was processed using the facebook/bart-large-mnli model via Hugging Face in a Google Colab environment. This phase focused exclusively on binary detection to measure model robustness at scale.

It is acknowledged that this two-phase design does not permit a direct, head-to-head comparison between GPT-4 and BART under a unified protocol. The asymmetry was a deliberate methodological choice. It was driven by two constraints: (1) the cost and rate-limiting of GPT 4's API made full 170-sentence evaluation impractical within this study's scope, and (2) the two phases serve distinct research purposes. The pilot phase assesses qualitative pragmatic competence (detection and paraphrasing). And the scaling phase stress-tests binary detection robustness at volume. Table 2 metrics are indicative of each model's characteristic failure modes, and not a direct performance race. Future work will evaluate both architectures under a unified experimental protocol on an expanded dataset in order to enable fair comparison.

2.4 Annotation and Validation

To ensure label reliability, a double-blind annotation was conducted. A primary native speaker of IndE and a secondary native speaker, both with expertise in educational linguistics, annotated the pilot subset. A Cohen's kappa of 0.79 was achieved, indicating "substantial agreement."

Annotators identified: (a) whether a euphemism was present, (b) the target taboo or sensitive word, and (c) a literal paraphrase of the sentence. Each entry was thus labelled with a binary label, the target word, and a standard English translation of the intended meaning. Discrepancies were resolved through consensus based discussion to ensure that the final "Gold Standard" labels reflect a shared cultural understanding of IndE.

3. RESULTS AND DISCUSSION

3.1 Pilot Phase Results (GPT-4)

In the pilot phase, GPT-4 demonstrated high proficiency, achieving 82.5% accuracy (7 mismatches) in the detection task. In the paraphrasing task, it achieved 90% accuracy, with

only 4 semantic mismatches where the model failed to capture the specific Indian social gravity of an expression.

Model	Acc.	Prec.	Rec.	F1
BART Large	55.3%	58.6%	84.1%	69.0%
GPT-4	82.5%	80.5%	84.6%	82.5%

Table 2: Performance Metrics

3.2 Scaling Phase Results (BART) The scaling phase revealed a significant performance drop. As detailed in Table 2, BART Large yielded an accuracy of only 55.3%. While the model showed high Recall (84.1%), its Precision was low (58.6%) due to excessive false positives.

3.3 Error Analysis: The "Indianism" Bias

The most striking result from the BART evaluation is the high rate of False Positives (FPs). Out of 69 literal control sentences, BART incorrectly flagged 60 as euphemisms, revealing a "dialectal bias": the model treats any non-standard English construction as a euphemism, despite such constructions being standard usage in the Indian subcontinent.

Phrases such as "Give me a missed call," "I passed out of college," and "What is your good name?" were classified as meaning-masking expressions. BART's training on Standard American/British English evidently causes it to treat regional pragmatic variation as inherently "suspicious" or indirect.

Three compounding factors appear to drive this bias. First, BART-large-mnli was pre-trained predominantly on Western English corpora (BookCorpus, CC-News, OpenWebText), meaning its entailment priors have no representation of IndE as a grammatically coherent variety. Second, class imbalance in the IndEuph-170 dataset — with euphemistic instances (101) outnumbering literal controls (69) — may have reinforced a positive-classification tendency. Third, BART's zero-shot NLI framing (hypothesis: "This sentence is a euphemism") is an imprecise instrument for the pragmatic detection task. It requires sensitivity to context and speaker intent rather than surface lexical cues. Future work offers a promising avenue to address this bias through targeted fine-tuning on Indic English corpora, or by substituting Indic specific architectures such as IndicBERT (Kakwani et al., 2020) or MuRIL (Khanuja et al., 2021).

3.4 Qualitative Gaps in GPT-4: Pragmatic Blind-Spots

Despite its higher accuracy, GPT-4 failed on culturally specific expressions involving Indian authority and social dynamics. It missed the punitive tone in "The police launched a campaign to crack down…" and the corruption-related nuance in "He has a setting with the manager."

These results highlight a Western-centric bias and an absence of deep sociopragmatic mapping of the Indian context, even in the most advanced LLMs.

A closer breakdown of GPT-4's seven detection errors reveals a consistent pattern: the model struggles specifically with Social Status and authority-coded expressions. Three of the seven errors involved Social Status euphemisms (e.g., "convent-educated," "decent family"), where GPT 4 identified the literal meaning correctly but failed to recognise the hierarchical subtext. Two errors involved Indirectness expressions with punitive or institutional authority registers (e.g., "the officer asked him to cooperate"). GPT-4 parsed these as literal requests rather than coercive softening. One error involved a Humour/Slang category item ("jugaad fix") where the loanword was treated as untranslatable rather than euphemistic. The final error involved a code-mixed token where the English frame around a Hindi noun misled the classifier. These patterns suggest that GPT-4's failures are not random but cluster around power asymmetric social registers that are culturally specific to the Indian context. Few-shot prompting with culturally annotated exemplars may improve performance on these categories in future evaluations. Techniques such as explicitly encoding the euphemism taxonomy as part of the system prompt could also prove helpful.

3.5 Type-Specific Performance

A category-wise breakdown of the 101 euphemisms reveals that Humour and Slang (e.g., "chutney," "timepass") achieved the highest detection rates across both models. Understatement (e.g., "small scene," "adjusting") proved most difficult to paraphrase correctly, frequently resulting in literal translations that strip the expression of its social gravity.

4. CONCLUSION

This paper demonstrates that Indian English euphemisms pose a significant challenge for current NLP models. Through the curation of IndEuph-170, a 170-sentence benchmark, it has been shown that state-of-the-art models consistently either over-classify dialectal variations as euphemisms (BART) or miss

culturally specific taboos related to authority and social negotiation (GPT-4).

The interpretative failure is twofold: (1) smaller models over-classify dialectal variations as euphemisms, and (2) larger models miss the social power dynamics embedded in Indian expressions. It is not sufficient for an AI to be "multilingual"; it must also be trained to be "multicultural."

Future work will expand IndEuph-170 to over 1,000 samples, incorporating Bengali-English and Tamil-English code-mixed euphemisms to determine whether the bias persists in mixed language settings. To ensure annotation reliability and cultural authenticity at scale, expansion will employ a structured protocol: each regional variety will be annotated by a minimum of two native speakers with demonstrated competency in the target variety, with Cohen's kappa ≥ 0.75 required for inclusion. Taxonomic consistency will be maintained by anchoring annotations to the five-category Allan and Burridge (1991) framework used in the current dataset, with an additional code-mixing category to capture intra sentential switching. Additionally, training Indic specific models such as IndicBERT (Kakwani et al., 2020) or MuRIL (Khanuja et al., 2021) on the expanded dataset, and evaluating both models under a unified protocol, represents the primary next step toward fair cross-architecture comparison.

Once this gap is bridged, it becomes possible to move toward NLP systems that better address the complex indirectness that defines human communication in the Global South.

ACKNOWLEDGEMENTS

The authors thank the native speaker annotators whose cultural expertise was indispensable to the dataset validation process. We also thank the reviewers for their constructive feedback, which has substantially improved this work. The author specially thanks Dr. Ashwini Vaidya (IIT Delhi) for the intellectual grounding and guidance provided through her course on 'Computational Models of Meaning', which inspired the framing of this work.

BIBLIOGRAPHICAL REFERENCES

Allan, K. and Burridge, K. (1991). Euphemism and Dysphemism: Language Used as Shield and Weapon. Oxford University Press.

Fu, C., Chen, P., Shen, Y., Qin, Y., Zhang, M., Lin, X., et al. (2025). MME: A comprehensive evaluation benchmark for multimodal large language models. In Proceedings of the Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track.

Grice, H. P. (1975). Logic and conversation. In Syntax and Semantics, Vol. 3: Speech Acts (pp. 41–58). Academic Press.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D. and Steinhardt, J. (2021). Measuring massive multitask language understanding. In Proceedings of the International Conference on Learning Representations (ICLR).

Hu, Y., Li, J., Wu, M., Huang, Z., Chen, G. and Sha, Y. (2024a). A unified generative framework for bilingual euphemism detection and identification. In Findings of the Association for Computational Linguistics: ACL 2024 (pp. 403–417). Association for Computational Linguistics.

Hu, Y., Li, J., Wu, M., Huang, Z., Chen, G. and Sha, Y. (2024b). Uncovering and mitigating the hidden chasm: A study on the text-text domain gap in euphemism identification. In Proceedings of the AAAI Conference on Artificial Intelligence, 38(16), 18270–18278. https://doi.org/10.1609/aaai.v38i16.29786

Kakwani, D., Kunchukuttan, A., Golla, S., Gokul, N. C., Iyer, A., Khapra, M. M. and Kumar, P. (2020). IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages. In Findings of the Association for Computational Linguistics: EMNLP 2020 (pp. 4948–4961).

Khanuja, S., Bansal, D., Mehtani, S., Khosla, S., Dey, A., Gopalan, B., Margam, D. K., Aggarwal, P., Nagipogu, R. T., Dave, S., Gupta, S., Gali, S. C., Subramanian, V. and Talukdar, P. (2021). MuRIL: Multilingual representations for Indian languages. arXiv preprint arXiv:2103.10730.

Sailaja, P. (2012). Indian English. Edinburgh University Press.

Srivastava, V. and Singh, M. (2021). Challenges and considerations with code-mixed NLP for multilingual societies. arXiv preprint arXiv:2106.07823.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O. and Bowman, S. R. (2018). GLUE: A multi task benchmark and analysis platform for natural language understanding. In Proceedings of the 2018 EMNLP Workshop BlackboxNLP (pp. 353–355). Association for Computational Linguistics.

Zeng, L. (2024). Leveraging large language models for code-mixed data augmentation in sentiment analysis. In Proceedings of the 2024 Symposium on Social Influence and Conversational AI (pp. 1–17).

Zhang, R. and Eickhoff, C. (2024). CroCoSum: A benchmark dataset for cross-lingual code switched summarization. In Proceedings of LREC-COLING 2024 (pp. 367–382). European Language Resources Association.

Integrating Syntactic and Discourse Signals through Multi-Encoder Fusion in NMT for Low-Resource Indian Language Pairs

Sobha Lalitha Devi, Vijay Sundar Ram R, Pattabhi RK Rao

AU-KBC Research Centre
MIT Campus of Anna University,Chennai, India
sobha@au-kbc.org

Abstract

Neural Machine Translation (NMT) for low-resource Indian language pairs such as Hindi–Tamil and Tamil–Malayalam remains challenging due to morphological richness, syntactic divergence, and limited availability of high-quality parallel corpora. While Transformer-based architectures achieve strong performance in high-resource settings, they often struggle to model syntactic structure and discourse-level dependencies in low-resource scenarios, resulting in errors in agreement, word order, and pronoun translation. In this work, we propose a linguistically informed multi-encoder fusion framework that explicitly incorporates syntactic and discourse signals into NMT. Experiments conducted on Hindi–Tamil and Tamil–Malayalam parallel corpora demonstrate consistent improvements over strong Transformer baselines in BLEU and ChrF scores, along with gains in pronoun translation accuracy and agreement consistency. The results highlight the effectiveness of explicit linguistic integration for improving NMT in low-resource Indian language settings.

Keywords: Multi-Encoder Fusion, Linguistic Features, Neural Machine Translation, Low Resource Languages, Hindi, Tamil, Malayalam

1. Introduction

Neural Machine Translation (NMT) has significantly improved translation quality with the introduction of encoder–decoder architectures, especially with the Transformer model introduced in Attention Is All You Need (Vaswani, 2017). Despite these advances, NMT systems especially for Indian languages still struggle with:

- Syntactic ambiguities
- Long-distance dependencies
- Pronoun resolution and discourse consistency

Two major linguistic signals that can address these issues are:

- Part-of-Speech (POS) information – captures syntactic structure
- Anaphora resolution – resolves pronouns and coreference relations

Encoder fusion integrates these linguistic features directly into the NMT encoder to enhance contextual representation and improve translation quality.

Modern NMT systems typically use, Encoder and Decoder. An Encoder converts source sentence into contextual embedding. A Decoder generates target sentence based on encoded representation. The Transformer model in general uses:

- Multi-head self-attention
- Positional encoding
- Feed-forward layers

However, pure data-driven learning does not effectively capture explicit syntactic and discourse-level information.

POS information will help in disambiguation of homographs, improved word reordering and better syntactic alignment. Anaphora resolution identifies the antecedent for anaphor which it refers to. And in translation, especially for languages such as Hindi, Tamil and Malayalam, incorrect pronoun resolution leads to grammatical errors. Anaphora resolution helps in clarity and cohesion in discourse by resolving references to previously mentioned entities, which can effect gender and number agreement.

This paper is further organised as follows: Section 2 describes the related works in this area of research. Section 3 describes the data and its preparation. Section 4 describes the methodology. Section 5 describes experiments and results. Section 6 concludes the paper.

2. Related Works

Encoder fusion has emerged as an important architectural strategy in deep learning, particularly in sequence-to-sequence and multimodal models, where information from multiple encoder representations is combined to improve downstream performance. Gao et al (Gao, 2020) categorized fusion strategies into early, intermediate, and late fusion, depending on the stage at which representations are combined. Encoder fusion typically falls under intermediate fusion, where latent feature representations from one or more encoders are integrated to form a richer joint representation. Liu et al (2021) has presented work on improving translation output using encoder fusion technique for sequence-to-sequence model. They proposed a simple fusion method by fusing only the encoder embedding layers for the softmax layer. Their experiment revealed that this methodology learns more expressive bilingual word embedding by

Proceedings of WILDRE-8, pages 98–103

building between relevant source and target embeddings.

Das et al (2022) has used encoder fusion in their Personalized response selection system, where persona, emotion, and entailment information are fusion. They used the fusion strategies and concept-flow encoding to train a BERT-based model which outperforms the previous methods by 2.3% on original personas.

Recent survey work by Jiao et al (2024) emphasized that intermediate fusion at the encoder level allows models to preserve modality- or feature-specific information while still enabling effective interaction between representations.

Huang et al (2025) has used encoder fusion architectures to improve information retrieval tasks, where they have fused the text and image features.

Encoder fusion is utilized in many multi-model learning experiments, where separate encoders process different modalities such as text, images, or speech and fused together. Li and Tang (Li et al, 2024) provided a recent survey on multimodal alignment and fusion, highlighting attention-based and representation-level fusion methods that combine outputs from multiple encoders.

Building on these insights, this work adopts an encoder fusion framework to integrate linguistic features into neural machine translation. Unlike prior approaches that rely solely on final-layer representations or treat linguistic annotations as simple input embeddings, we explicitly fuse encoder representations to jointly model lexical and syntactic information. This approach is particularly relevant for morphologically rich and syntactically divergent language pairs, such as Hindi – Tamil and Tamil-Malayam, where explicit modeling of syntactic structure can help alleviate data sparsity and re-ordering challenges.

3. Data

There is a need of large number of parallel sentences to build a robust Neural Machine Translation (NMT) system. Thus data is very crucial in the development of NMT systems. Publicly available data are English centric. One of huge parallel corpus which is available is NLLB (Costa-Jussà, et. al., 2022). This is a large-scale multilingual bitext (parallel text) corpus created by Meta AI as part of their effort to support translation across many languages including low-resource ones. The dataset covers 148 English-centric language pairs (i.e., English ↔ X) and 1,465 non-English-centric pairs (i.e., X ↔ Y) using metadata mined bitext. (Costa-Jussà, et. al., 2022)

Samanantar is one of the largest publicly-available parallel corpora for Indic languages. It contains ~49.7 million sentence-pairs between English and 11 Indic languages (Assamese, Bengali, Gujarati, Hindi, Kannada, Malayalam, Marathi, Odia, Punjabi, Tamil, Telugu). (Ramesh et. al., 2022)

Bharat Parallel Corpus Collection (BPCC) is another comprehensive and publicly available parallel corpus that includes both existing and new data for all 22 scheduled Indic languages. It is comprised of two parts: BPCC-Mined and BPCC-Human, totaling approximately 230 million bitext pairs. BPCC-Mined contains about 228 million pairs. (Jay et.al, 2023)

Most of the parallel sentences available in these corpora are between language X and English. Opus is an Open Parallel Corpora, which aggregates the publicly available parallel corpus and helps to build the parallel corpus for the available parallel corpora.

From the above corpora, we prepared parallel sentences for Hindi-Tamil and Tamil-Malayalam. We have collected 2987320 parallel sentences for Tamil to Hindi pairs. These parallel sentences are collected from Tamil-English and Hindi-English parallel sentences available in the above mentioned parallel data. Hindi to Tamil parallel sentences are collected by considering the Tamil and Hindi sentences which have common English sentence. Similarly we collected 1492138 Tamil-Malayalam parallel sentences. These parallel sentences were filtered using Comet and LaBSE scores. LaBSE (Language-agnostic BERT Sentence Embedding) is a multilingual BERT to produce language-agnostic sentence embeddings for 109 languages. This model combines masked language model (MLM) and translation language model (TLM) pretraining with a translation ranking task using bi-directional dual encoders. (https://github.com/bojone/labse). COMET (Crosslingual Optimized Metric for Evaluation of Translation) uses a pre-trained multilingual transformer encoder such as XLM-RoBERTa. It uses deep multilingual semantic representations. Each input sentence is encoded into contextual embeddings.

The encoder captures Cross-lingual semantic alignment, (https://unbabel.github.io/COMET/)

We selected the parallel sentences with scores greater than 0.85. We got 9,26,962 Tamil-Hindi parallel sentences and 3,73,165 Tamil-Malayalam parallel sentences.

274	நாட்	##களில்	இந்த	விலங்கு	##களின்	பால்	உற்பத்தி	280	கிலோ	##கி	##ரா	##ம்	ஆகும்	.
B-QT_QTC	B-N_NN	I-N_NN	B-DM_DMR	B-N_NN	I-N_NN	B-N_NN	B-N_NN	B-QT_QTC	B-N_NN	I-N_NN	I-N_NN	I-N_NN	B-V_VM_VF	B-RD_PUNC

Figure 1: Example 1 sentence after sub-word processing with each token aligned with the POS information

As mentioned earlier we intend to train the models with the linguistic information, namely, Part-of-Speech (POS) tags and Anaphor information. So we processed the source sentences with POS tagger and Anaphora resolution engine.

For any neural machine learning, tokenization process is very crucial step. In the following paragraphs we explain about the tokenization that is used in this work.

3.1 Tokenization

In Neural Machine translation, tokenization is beyond separation of words based on white space. Here sub-word tokenization is done using Statistical measures or linguistic features. This reduces the vocabulary size and increases the frequency of the tokens and improves the translation by handling rare words and unknown words. The sub-word tokenization is very beneficial to morphologically rich languages as vocabulary size is large compared to other languages due to productive inflectional and derivational suffixations. Sennrich et. al. (2016) presented the statistical word segmentation techniques which is based on simple character n-gram model and segmentation based on the byte pair encoding (BPE) comparison algorithm. BPE sub-word algorithm is one of the widely used sub-word tokenization algorithm.

The other sub-word tokenization algorithms include, WordPiece, SentencePiece, and language specific algorithms such as Mecab (a morphological analysis based Japanese tokenizer), Stanford Word Segmentation (a Chinese word segmenter based on Conditional Random Fields). Ram and Sobha (2023) has presented a comparative study on effectiveness of morphological based and BPE sub-word segmentation.

In this work, we have tokenized the data using Indic-tokenizer[1]. The major disadvantage with the BPE sub-word algorithm for Indian languages is that, it segments word into tokens which do not form valid letters of the language's alphabet set. Consider the following example 1.

Example 1:
Tamil Sentence:
"274 நாட்களில் இந்த விலங்குகளின் பால் உற்பத்தி 280 கிலோகிராம் ஆகும்"

[1] https://github.com/sudarsun/indic-tokenizer

After sub-word tokenization using BPE:

274 நாட@@ ்களில் இந்த விலங்க@@ ுகளின் பால் உற்பத்தி 280 க@@ ிலo@@ ோ@@ கி@@ ராம் ஆகும் .

Here in the above example, the glyphs (mathras) such as ' ு, ், ி' should occur with the previous token to form a proper letter.

Example 2 shows the tokenized output for same sentence as in example 1, when it is tokenized using the Indic-tokenizer.

Example 2:

"274 நாட் ##களில் இந்த விலங்கு ##களின் பால் உற்பத்தி 280 கிலோ ##கி ##ரா ##ம் ஆகும் ."

Here the tokens have the valid letters.

Thus in the work presented here Indic-tokenizer is used in the NMT system development.

In the sub-word tokenization in both the methodologies the words are divided into sub-words. Now we need to align the POS and Anaphora information to the sub-words, which are originally for the wordforms before tokenization. We aligned the POS and Anaphora information with the subword tokenized sentences, by distributing the POS and Anaphora information assigned to word to its sub-words also. For this purpose the BIO format is used as in BIO format.
Consider the following example:

விலங்குகளின் (vilangkukaLin)
when processed with sub-word tokenization it is split into 'விலங்கு ##களின்'. The POS for this word is 'N_NN'. Here POS for the sub words are assigned as follows:

விலங்கு/B-N_NN ##களின்/I-N_NN

Similarly the same is followed for the Anaphora information also. The anaphora information is assigned to the sub-words as follows:

விலங்கு/B-Ante-3 ##களின்/I-Ante-3

Figure 1 shows the sentence in example 1 after sub-word processing in which each token is aligned with the POS information.

4. Methodology

Encoder fusion is a family of architectures where multiple encoders are combined to produce a single representation for downstream tasks. This shows up a lot in multilingual NLP, multimodal models, domain adaptation, and parameter-efficient fine-tuning.
The fusion mechanism operates as an additive multi-stream encoder where:

- Primary stream: Standard token embedding's capture semantic and lexical information
- Auxiliary streams: Separate embedding spaces for POS tags and chunk labels capture syntactic structure
- Fusion point: Early fusion at the embedding layer, before the self-attention stack

This differs from late fusion (combining features after encoding) or feature concatenation approaches by maintaining the original embedding dimensionality through projection.
In this work late fusion approach is used, having gating mechanism. In encoder fusion, multiple representations are combined:
Word embedding's, POS embedding's, Coreference/anaphora embedding's. These are fused within the encoder layers. The gating architecture mechanism works as follows:

$$H = \alpha\ H\{word\} + \beta\ H\{pos\} + \gamma\ H\{coref\}$$

Where weights α, β and γ are learned dynamically.

POS Embedding Layer:

Each POS tag is mapped to a dense vector:

$$E\_\{pos\} = Embedding(POS\_tag)$$

then, combined with word embedding:

$$E\_\{input\} = E\_\{word\} + E\_\{pos\}$$

Coreference Embedding Layer:

Coreference signals are encoded as:

- Binary features (is pronoun, is antecedent)
- Entity cluster embeddings
- Distance-based embeddings

These are fused into encoder representations.

The architecture used in this work adopts a multi-encoder design consisting of dedicated encoders for lexical tokens, Part-of-Speech (POS) sequences, and coreference annotations. Each encoder captures complementary aspects of linguistic structure—semantic content, syntactic function, and discourse-level referential relations—before their representations are integrated through attention-driven and learnable gating fusion mechanisms. By explicitly modeling token-level semantics, syntactic structure, and discourse-level coreference, this multi-encoder fusion framework produces richer contextual embedding's, facilitating discourse-coherent and syntactically accurate translations in low-resource Indian language pairs.

5. Experiments and Results

We evaluated the translations of the NMT models for both Hindi to Tamil and Tamil to Malayalam, using BLEU score (Papineni et al., 2002). We used Sacrebleu python library to calculate the BLEU scores. The results are presented in Table 1. The BLEU scores show that the model with POS and Anaphora/Coreference features integrated has improved by 3% the translation in both Hindi-Tamil and Tamil-Malayalam. We have developed two models viz.,

a) **Sys-1** – NMT trained with just the Parallel data using the indic tokenization (as explained in section 3.1)
b) **Sys-2 –** NMT trained with encoder fusion using indic tokenization.

SNo	**Details**	**Hindi to Tamil**		**Tamil to Malayalam**	
		BLEU	**chrF**	**BLEU**	**chrF**
1	Sys-1	26.23	52.33	31.46	61.87
2	Sys-2	29.88	57.97	35.66	66.42

Table 1: BLEU Score for Hindi - Tamil and Tamil - Malayalam

On analysis of the translation output from different experiments in both Hindi to Tamil and Tamil to Malayalam, our observations are as follows,

Sys-1: Translated sentences were complete but most of these translations were not the exact translation. Translations convey a different sense due to the choice of the verb generation.

There were also words omitted in the translation. Technical words and rare words were handled, but there were errors in it.

Sys-2: Clausal sentences were translated better than the systems. Verb phrase generation was exact, though there were errors.

Overall this output was observed to be more coherent and closer to human translation.

We have explained the translation output with examples in the further part of this section.

Ex 1.(HI to Tamil):
Hindi-Input:
म्यूटेशन आनुवंशकि में मलि सकते हैं.
(Mutations can be found in genetics.)
Tamil Translations:

Sys-1:பிறழ்வுகள் மரபணு மரபணுவில் இருக்கலாம்.
(Mutations can occur in the genetics genetics.)

Sys-2: பிறழ்வுகள் மரபணுவில் கிடைக்லாம்.
(Mutations can be found in genetics.)

In this example we observe that both Sys-1 and Sys-2 outputs are proper sentences but Sys-2 translation has better sense translation.

Ex2: Clausal Sentence
Hindi-Input:
इतहिास उस दौर से शुरू होता है जब लोग लखिने की कला जानते थे.
(History begins from the time when people knew the art of writing.)

Tamil Translations:
Sys-1: மக்கள் எழுதும் கலத்திலிருந்து தொடங்கும்போது வரலாறு தொடங்குகிறது.
(History begins from when people started writing.)

Sys-2: மக்கள் எழுதும் கலையை மக்கள் அறிந்த போது வரலாறு தொடங்குகிறது.
(History begins from the time when people knew the art of writing.)

In Ex 2, the Hindi sentence has a relative participle clause. The clause construction was correctly translated by the **Sys-2** system. It has generated the relative participle verb, 'அறிந்த' (aRintha).

Ex 3:
Hindi-Input:
ग्राउंड रखरखाव उपकरण, जैसे लॉन मोवर, रोलर्स, लाइम पाउडर मशीन, मार्कगि मशीन, घास काटने वाली तलवारें, दरांती, श्रब मास्टर, कटर .
(Ground Maintenance Equipments like Lawn Mower, Rollers, Lime Powder Machine, Marking Machine, Mower, Sickle, Shrub Master, Cutter etc.)
Tamil Translations:
Sys-1: தரை பராமரிப்பு சாதனங்கள், ஒவ்வாமைகள், லீக் தூள் இயந்திரங்கள், புல்வெளி இயந்திரங்கள், புல்வெளிகள், ஆப்பிரிக்கா, ஆப்பிரிக்கா, கேரட் மற்றும் பலர் போன்ற தரை பராமரிப்பு சாதனங்கள்.
(Ground care equipment like Ground care equipment, Allergies, Leek powder machines, Lawn machines, Lawns, Africa, Africa, Carrot and many others.)

Sys-2: புல் வெட்டும் இயந்திரம், ரோலர்கள், சுண்ணாம்பு பொடி இயந்திரங்கள், நின்மஞ்சள் அழைப்பான், குறியீட்டு இயந்திரம், புல் வெட்டும் வாள்கள், நறுமணங்கள், சிரோமாஸ்டர், கடர் போன்ற தரை பராமரிப்பு உபகரணங்கள் அகும்.
(Ground care equipment like such as Lawn care equipment such as lawnmowers, lawnmowers, trucks, lime powders, machines, machines, machines, machines, mowers etc.)

In example 3, the Hindi sentence has series of noun phrases. **Sys-1** has generated output with many words which are not in the input sentences such as 'Africa', 'Carrot' etc. **Sys-2**, most of the noun phrases is translated and also the structure of the Tamil sentence is generated properly. This shows that Sys-2 has better performance than Sys-1. It requires little more data in the training.

We have also performed human evaluation using three human evaluators. The evaluation metrics used was Fluency **(F)** and Comprehensibility **(C).** Human evaluators had scored in a scale of 1-10 (1 indicates the lowest and 10 the highest score). In the table 2 we present the average human evaluation scores for both **Sys1** and **Sys2.**

SNo	Details	Hindi to Tamil		Tamil to Malayalam	
		F	C	F	C
1	Sys -1	62.66	69.88	67.65	71.34
2	Sys-2	71.27	79.45	78.85	81.35

Table 2: Human Evaluation Scores for the Sys1 and Sys2

6. Conclusion

We have presented our work in building Neural Machine Translation system in which we incorporated the syntactic and semantic features into the model development through encoder fusion gated mechanism. We have compared our Encoder Fusion NMT with NMT without Encoder Fusion. Hindi is an Indo-Aryan language. Malayalam and Tamil are Dravidian languages. All the three languages are morphologically rich language. And Tamil and Malayalam are highly agglutinative. The languages have different semantic and syntactic features such as the pronominals usage, PNG agreements etc. In our experiments we have observed that the encoder fusion has significant improvement. We have obtained 3% improvement. In future we plan to conduct ablation studies and also compare with LLMs.

7. Acknowledgments

This work is part of the research project titled "Discourse Integrated Dravidian Language to Dravidian Language Macine Translation (DL-DiscoMT)" funded under National Language Translation Mission (NLTM), Bhashini by Ministry of Electronics and Information Technology (MeitY), Government of India.

8. Bibliographical References

Castor, A. and Pollux, L. E. (1992). The use of user modelling to guide inference and learning. *Applied Intelligence*, 2(1):37–53.

Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan AK, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Mahalakshmi J, Divyanshu Kakwani, Navneet Kumar, Aswin Pradeep, Srihari Nagaraj, Kumar Deepak, Vivek Raghavan, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh Shantadevi Khapra. (2022). Samanantar: The largest publicly available parallel corpora collection for 11 Indic languages. Transactions of the Association for Computational Linguistics, 10:145–162.

Costa-Jussà, M. R., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., Heffernan, K., Kalbassi, E., Lam, J., Licht, D., Maillard, J., Sun, A., Wang, S., Wenzek, G., Youngblood, A., Akula, B., Barrault, L., Mejía González, G., Hansanti, P., Hoffman, J., Jarrett, S., Sadagopan, K. R., Rowe, D., Spruit, S., Tran, C., Andrews, P., Ayan, N. F., Bhosale, S., Edunov, S., Fan, A., Gao, C., Goswami, V., Guzmán, F., Koehn, P., Mourachko, A., Ropers, C., Saleem, S., Schwenk, H., amp; Wang, J. (2022). "No Language Left Behind (NLLB): Scaling Human-Centered Machine Translation". arXiv preprint arXiv:2207.04672 https://doi.org/10.48550/arXiv.2207.04672

Souvik Das, Sougata Saha, and Rohini K. Srihari. (2022). "Using Multi-Encoder Fusion Strategies to Improve Personalized Response Selection". In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 532–541, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Gala, Jay, Pranjal A. Chitale, A. K. Raghavan, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar M., Janki Atul Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, Pratyush Kumar, Mitesh M. Khapra, Raj Dabre, and Anoop Kunchukuttan. (2023). "IndicTrans2: Towards High-Quality and Accessible Machine Translation Models for All 22 Scheduled Indian Languages." Transactions on Machine Learning Research.

Wang, Yang. (2020). Deep multi-modal data analytics: Collaboration, rivalry and fusion. arXiv preprint arXiv:2006.08159.

Li, Songtao and Hao Tang. (2024). Multimodal alignment and fusion: A survey. arXiv preprint arXiv:2411.17040.

Liu, Xuebo and Wang, Longyue and Wong, Derek F. and Ding, Liang and Chao, Lidia S. and Tu, Zhaopeng. (2020). "Understanding and Improving Encoder Layer Fusion in Sequence-to-Sequence Learning.", Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing(EMNLP) https://arxiv.org/abs/2012.14768

Vijay Sundar Ram and Sobha Lalitha Devi. (2023). "Hindi to Dravidian Language Neural Machine Translation System". In: Proceedings of Recent Trends in Natural Language Processing (RANLP), 2023.

Gao, Jing, Peng Li, Zhikui Chen, and Jianing Zhang. (2020). A survey on deep learning for multimodal data fusion. Neural Computation 32(5):829–864.

Jiao, Tianzhe, Chaopeng Guo, Xiaoyue Feng, Yuming Chen, and Jie Song. (2024). A comprehensive survey on deep learning multi-modal fusion: Methods, technologies and applications. Computers, Materials & Continua 80(1):1–35.

NE-LID: A Fast and Accurate Language Identification System for Northeast Indian Languages

Badal Nyalang
MWire Labs, Shillong, Meghalaya, India
nyalang@mwirelabs.com

Abstract

Language identification (LID) is crucial for natural language processing systems, yet Northeast Indian languages remain severely underserved by existing multilingual LID models. We present NE-LID, a fast and accurate language identification system specifically designed for eleven languages of Northeast India. Built using character n-gram features with fastText, NE-LID achieves 99.09% accuracy on a balanced test set, significantly outperforming existing multilingual systems including GlotLID (73.12%), OpenLID (42.03%), IndicLID (39.30%), and LangDetect (24.33%). Our model processes predictions in 0.084 milliseconds on average, enabling real-time applications. We demonstrate that character-level modeling outperforms transformer-based approaches for script-diverse, low-resource languages.

Keywords: language identification, low-resource languages, Northeast India, fastText, multilingual NLP

1. Introduction

Northeast India is home to rich linguistic diversity with over 200 languages from multiple language families including Tibeto-Burman, Austroasiatic, and Indo-Aryan. Despite this diversity, these languages remain critically underserved by modern natural language processing tools, including language identification systems.

Language identification is the foundational task of automatically determining which language a given text is written in. While significant progress has been made in multilingual LID systems covering hundreds of languages (5), we demonstrate through empirical evaluation that existing systems perform poorly on Northeast Indian languages, with many systems completely failing to detect several languages in the region.

In this work, we present NE-LID, a language identification system specifically designed for eleven languages of Northeast India: Assamese, Bodo, English, Garo, Hindi, Khasi, Kokborok, Meitei, Mizo, Nagamese, and Nyishi. Our contributions are:

- A curated dataset of 22,000 sentences across 11 Northeast Indian languages for training and evaluation
- NE-LID, achieving 99.09% accuracy with 0.084ms inference time, significantly outperforming existing multilingual systems
- Comprehensive benchmark of four existing LID systems on Northeast Indian languages, revealing critical gaps in current multilingual models
- Empirical evidence that character n-gram models outperform transformer-based approaches for script-diverse, low-resource language identification

2. Related Work

2.1. Language Identification Systems

Modern language identification has evolved from early statistical methods to sophisticated neural approaches. GlotLID (5) covers over 2000 languages using fastText, while OpenLID (1) supports 201 languages. IndicLID (6) focuses specifically on 22 Indic languages including both native-script and romanized text. LangDetect (7), based on character n-gram profiles, supports 55 languages. However, coverage does not equal accuracy for low-resource languages: a system may list a language as supported while failing to correctly identify it in practice, particularly when training data for that language is scarce or absent.

2.2. Low-Resource Language Processing

Earlier attempts at language identification for Northeast Indian languages were primarily acoustic and prosodic in nature (2), leaving text-based LID severely underexplored. Northeast Indian languages face unique challenges including limited digital corpora, script diversity, and lack of standardized orthography. Recent efforts have begun addressing these gaps: ILID (3) created a dataset for 22 official Indian languages but excludes most Northeast languages (Khasi, Garo, Kokborok, Nyishi, Nagamese), and Tonja et al. (10) developed the first parallel corpora for 13 Northeast Indian languages. Terhiija et al. (9) surveyed spoken language technologies for Northeast Indian languages and highlighted the near-zero-resource status of most Tibeto-Burman varieties, underscoring the prerequisite role of accurate text-based language identification. The NE-BERT project (8) trained multilingual encoder models for nine Northeast Indian languages, providing the foundational corpus from

Proceedings of WILDRE-8, pages 104–108

which NE-LID draws its training data. However, language identification specifically for Northeast India has received limited attention despite being a prerequisite for other NLP tasks.

3. Languages and Data

3.1. Target Languages

We focus on eleven languages spanning four language families (Table 1). Hindi and English are included as anchor languages given their widespread use across Northeast India in administrative, educational, and digital contexts.

Family	Languages
Austroasiatic	Khasi
Tibeto-Burman	Garo, Bodo, Kokborok, Meitei, Mizo, Nagamese, Nyishi
Indo-Aryan	Assamese, Hindi
Germanic	English

Table 1: Target languages by language family

These languages exhibit significant orthographic diversity, using Latin script (Khasi, Garo, Mizo, Kokborok, Nyishi, Nagamese), Bengali-Assamese script (Assamese, Meitei when written in Bengali script), and Devanagari script (Bodo, Hindi). This script diversity poses challenges for LID systems that rely primarily on character-level features.

3.2. Dataset Construction

We constructed our dataset by sampling from the NE-BERT corpus (8), which contains web-scraped and curated text from various publicly available sources including news articles, social media, and digitized documents. We extracted 2,000 sentences per language, totaling 22,000 sentences. The dataset comprises approximately 90% publicly sourced content, with the remaining 10% drawn from curated and digitized materials.

We used a stratified split of 70% training (15,400 samples), 15% development (3,300 samples), and 15% test (3,300 samples). The larger development and test sets (15% each, yielding 300 samples per language) were chosen to ensure statistically reliable evaluation, which is especially important in low-resource settings where per-language sample sizes are small. Sentences range from short phrases (2-10 tokens) to longer passages (50+ tokens), ensuring diversity in text length.

4. Methodology

4.1. Model Architecture

We employ fastText (4) for supervised text classification. FastText learns vector representations of character n-grams and word n-grams, making it particularly suitable for morphologically rich and low-resource languages. Character n-grams allow the model to capture script-specific orthographic patterns without requiring large amounts of training data, making it well-suited for the script-diverse languages of Northeast India.

Our model configuration uses character n-grams of length 2-5 to capture subword patterns, word unigrams only, learning rate of 0.5, 25 training epochs, softmax loss function, and 8 training threads.

4.2. Training

Training was conducted on the 15,400-sample training set using the fastText supervised learning algorithm. The model converged after 25 epochs. FastText models are inherently compact and efficient, operating entirely on CPU without GPU requirements.

5. Experiments

5.1. Evaluation Setup

We evaluate on our held-out test set of 3,300 sentences (300 per language). We report overall accuracy and per-language accuracy. We additionally benchmark four existing multilingual LID systems: GlotLID, OpenLID, IndicLID, and LangDetect.

5.2. Main Results

Table 2 shows overall accuracy comparison across all five systems. NE-LID achieves 99.09% accuracy, outperforming the best competitor (GlotLID) by 25.97 percentage points, representing a 2.7$\times$ improvement. Development set accuracy is 99.00%, confirming that the model generalises well and is not overfitting to the test set.

Model	Accuracy
NE-LID (Ours)	**99.09%**
GlotLID	73.12%
OpenLID	42.03%
IndicLID	39.30%
LangDetect	24.33%

Table 2: Overall accuracy comparison

5.3. Per-Language Analysis

Table 3 shows per-language accuracy for all five systems. NE-LID achieves near-perfect accuracy ($>$95%) on all eleven languages, while competitor systems show significant gaps.

Language	NE-LID	GlotLID	OpenLID	IndicLID	LangDetect
Assamese	100.00	100.00	100.00	100.00	100.00
Bodo	98.67	89.33	0.00	96.67	0.00
English	96.00	79.00	96.00	83.00	94.33
Garo	99.67	0.00	0.00	0.00	0.00
Hindi	96.33	86.00	79.67	54.67	73.33
Khasi	99.67	95.33	0.00	0.00	0.00
Kokborok	99.33	99.33	0.00	0.00	0.00
Meitei	99.67	97.00	95.33	98.00	0.00
Mizo	99.00	92.67	91.33	0.00	0.00
Nagamese	100.00	0.00	0.00	0.00	0.00
Nyishi	99.33	65.67	0.00	0.00	0.00

Table 3: Per-language accuracy (%) for all systems

5.4. Inference Speed

We measured inference speed on CPU. NE-LID processes predictions extremely fast: short sentences ($<$50 chars) in 0.028ms, medium sentences (50-150 chars) in 0.065ms, and long sentences ($>$150 chars) in 0.160ms, with an overall average of 0.084ms ($\sim$12,000 predictions/second).

This speed enables real-time language identification for interactive applications and high-throughput batch processing. FastText operates entirely on CPU, making it accessible for deployment without specialized hardware.

6. Analysis

6.1. Coverage Gaps in Existing Systems

Our benchmark reveals critical gaps in existing multilingual LID systems:

- GlotLID fails completely on Garo and Nagamese (0% accuracy), despite claiming to support 2000+ languages
- OpenLID (Meta) only detects 5 of 11 Northeast Indian languages, completely missing Khasi, Garo, Bodo, Kokborok, Nagamese, and Nyishi
- IndicLID, despite focusing on Indic languages, covers only 4 of 11 languages. Notably, it achieves only 54.67% accuracy on Hindi, likely due to confusion with Maithili and Marathi which share the Devanagari script
- LangDetect performs worst overall (24.33%), essentially unusable for Northeast Indian languages

6.2. Why Character N-Grams Work

We initially attempted transformer-based approaches (NE-BERT, XLM-R) but found they performed poorly (9-37% accuracy) even when trained on our dataset. Character n-grams succeed because:

- **Script awareness**: Character n-grams directly capture script-specific patterns (Devanagari vs. Latin vs. Bengali-Assamese)
- **Orthographic distinctiveness**: Many Northeast languages have unique character combinations and diacritics
- **Low-resource robustness**: Character n-grams require less training data than transformer models to learn discriminative patterns
- **Efficiency**: FastText models are orders of magnitude faster than transformer inference

6.3. Error Analysis

The model produces 30 misclassifications (0.91% error rate) across the full test set. Table 4 shows development and test accuracy, confirming consistent performance across both splits.

Split	Accuracy
Development	99.00%
Test	99.09%

Table 4: Development vs. test accuracy

Table 5 shows accuracy broken down by sentence length. Short sentences ($<$50 characters) are the most challenging, consistent with the intuition that very short inputs provide fewer character n-gram features for discrimination.

Length	Count	Accuracy
Short ($<$50 chars)	1,107	98.37%
Medium (50–150 chars)	1,370	99.49%
Long ($>$150 chars)	823	99.39%

Table 5: Accuracy by sentence length on test set

Table 6 shows mean confidence scores per language on correctly classified samples. Languages with script-distinctive orthographies (Meitei, Assamese, Kokborok, Nyishi) yield the highest confidence, while Hindi and English show the lowest

confidence, consistent with their script overlap with other languages in the dataset.

Language	Mean Confidence
Hindi	0.9839
English	0.9941
Mizo	0.9949
Bodo	0.9949
Khasi	0.9956
Naga	0.9968
Garo	0.9975
Kokborok	0.9980
Nyishi	0.9985
Assamese	0.9985
Meitei	0.9986

Table 6: Mean prediction confidence per language (correct predictions only)

Manual inspection of the 30 misclassified samples reveals that several cases involve label noise in the source corpus, for example fully Hindi sentences labeled as Bodo, or fully English sentences labeled as Mizo, suggesting the true model error rate may be lower than 0.91%. The remaining errors follow interpretable patterns: Bodo-Hindi confusion due to shared Devanagari script; Khasi-Garo confusion as both use Latin script with similar phonological patterns; and Nyishi misclassifications on extremely short inputs with ambiguous character patterns.

7. Limitations

NE-LID has several limitations: (1) The model is designed for monolingual sentences and may struggle with code-mixed text; (2) While overall accuracy is high, performance may degrade on extremely short inputs ($\leq$2 tokens); (3) The model relies heavily on script patterns, which may fail when languages are transliterated to non-standard scripts; (4) Many other Northeast Indian languages (e.g., Ao, Tangkhul, Dimasa) are not covered.

8. Conclusion

We present NE-LID, a fast and accurate language identification system for eleven Northeast Indian languages. With 99.09% accuracy and 0.084ms inference time, NE-LID significantly outperforms existing multilingual systems and addresses critical gaps in language technology for Northeast India.

Our comprehensive benchmark reveals that "multilingual" LID models often fail to adequately support low-resource languages, with many systems completely missing 6-7 Northeast Indian languages. We demonstrate that character n-gram models are more effective than transformer-based approaches for script-diverse, low-resource language identification. The model and dataset are publicly available at https://huggingface.co/MWirelabs/ne-lid.

9. Future Work

Future work includes extending coverage to additional Northeast Indian languages, handling code-mixed text, improving robustness on very short inputs, and integrating NE-LID into downstream NLP pipelines for the region.

10. Bibliographical References

[1] Laurie Burchell, Alexandra Birch, Nikolay Bogoychev, and Kenneth Heafield. 2023. An open dataset and model for language identification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 865–879, Toronto, Canada. Association for Computational Linguistics.

[2] Chuya China Bhanja, Mohammad Azharuddin Laskar, and Rabul Hussain Laskar. 2019. A pre-classification-based language identification for northeast indian languages using prosody and spectral features. *Circuits Syst. Signal Process.*, 38(5):2266–2296.

[3] Yash Ingle and Pruthwik Mishra. 2025. ILID: Native script language identification for Indian languages. *arXiv preprint arXiv:2507.11832*.

[4] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. Association for Computational Linguistics.

[5] Amir Hossein Kargaran, Ayyoob Imani, François Yvon, and Hinrich Schuetze. 2023. GlotLID: Language identification for low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12028–12051, Singapore. Association for Computational Linguistics.

[6] Yash Madhani, Mitesh M. Khapra, and Anoop Kunchukuttan. 2023. Bhasa-Abhijnaanam:

Native-script and romanized language identification for 22 Indic languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 816–826, Toronto, Canada. Association for Computational Linguistics.

[7] Shuyo Nakatani. 2010. Language detection library for java.

[8] Badal Nyalang. 2026. NE-BERT: A multilingual language model for nine Northeast Indian languages. In *Proceedings of the Second Workshop on Language Models for Low-Resource Languages (LoResLM 2026)*, pages 1–12, Rabat, Morocco. Association for Computational Linguistics.

[9] Viyazonuo Terhiija, Samudra Vijaya, and Priyankoo Sarmah. 2019. Spoken language technology for north-east indian languages. In *Proceedings of the 1st International Conference on Language Technologies for All*, pages 182–185, Paris, France. European Language Resources Association (ELRA).

[10] Atnafu Lambebo Tonja, Hellina Hailu Nigatu, Olga Kolesnikova, Grigori Sidorov, and Alexander Gelbukh. 2023. First attempt at building parallel corpora for machine translation of Northeast India's very low-resource languages. *arXiv preprint arXiv:2312.04764*.

11. Language Resource References

[1] MWire Labs. (2025). NE-LID: Language identification model for Northeast Indian languages. HuggingFace. https://huggingface.co/MWirelabs/ne-lid

[2] MWire Labs. (2025). NE Multilingual Corpus: Web-scraped text for Northeast Indian languages. HuggingFace. https://huggingface.co/datasets/Badnyal/ne-multilingual-corpus

Integrating Cultural Wisdom and Digital Technologies for Children's Moral and Emotional Development

Garima[1], Girish Nath Jha[2]

School of Sanskrit and Indic Studies, Jawaharlal, Nehru University
Rankawatgarima03@gmail.com, girishjha@jnu.ac.in

Abstract

The influence of technology on children's education is increasing rapidly in the digital age, but with it comes the challenge of how to develop children's cultural and moral values in a balanced manner in the age of AI. In traditional societies, moral and cultural teachings have often been imparted through religious and philosophical texts, memorisation, interpretation, and oral traditions. This research presents an AI-based value-oriented learning framework that aims to make cultural and ethical teachings more structured, simple, and technologically accessible to children. The study includes a brief analysis of memory-based teaching systems prevalent in various religious traditions and incorporates insights from existing approaches to propose an integrative model based on selected verses from the Bhagavad Gītā. The proposed system includes data generation and processing, simplified interpretation, semantic understanding, pronunciation analysis, and interactive learning features based on selected cultural content. While the framework is conceptual in nature and has not yet undergone pilot implementation, it is designed as a foundation for future empirical validation. The study suggests that, through AI and modern technologies, traditional cultural knowledge can be delivered to children in a more effective and engaging manner, thereby opening new possibilities for strengthening their moral and cultural development

Keywords: Bhagvad Gītā**,** AI integration, ASR, TTS, Value-Based pedagogy, AI-integrated learning

1. Introduction

In the present global society, children are growing in an environment where technology, social media, digital media and rapid information flow have become a natural part of their daily experience. While it has led to benefits such as social cooperation, global reach, development of self-expression and access to communities, its long-term use by children and adolescents poses serious risks to mental health (Zsila and Reyes, 2023). Today, while the education systems seem to focus primarily on cognitive competencies and academic achievement, the development of values, self-control, compassion, and discernment of right and wrong among children has been relatively neglected.

Today, the question of student ethics and accountability has become one of the biggest challenges in the field of education (Ramadhani at el., 2024). In such a situation, value education can play an important role in building a generation that is not only intellectually intelligent but also possesses moral awareness, social responsibility, and high morality.(Ruhiyat at el., 2026). (Lamb and Brooks, 2026) believe that character education should not be limited to indirect moral formulation but should be adopted in well-planned strategies that develop both a cognitive understanding of virtues and their practical application.

Some efforts are being made in all countries regarding this subject. Historically, all religions have also used their own cultural narratives, life-philosophies, narratives, symbols and dialogue traditions for moral and personality development through different texts, civilisations, and cultures. These pedagogical methods were not just a means of communicating knowledge, but were also a means to nurturing a relationship with society, life and self and for establishing a balance.

According to (Banerjee, 1999), the uniqueness of religion and theology lies in the fact that they determine the duties for every individual in society and provide a moral and ethical framework for proper conduct. (Dubey,2012) has highlighted their importance and presented religion as a righteous way of ethical and moral life. However, due to modern lifestyles, nuclear family structures, and the progressive use of digital gadgets, this cultural and behavioural dialogue is gradually becoming disconnected from the children's world. Moreover, the complexity of cultural texts, which can neither be understood quickly nor easily linked to the present situations, children have challenges in developing an interest in these subjects.

In this context, the research begins with the fundamental question of whether modern technology, rather than being a barrier to the moral, cultural and emotional development of children, can be used as a bridge. That is, can culture and moral values be effectively communicated to children through technology as well. (Park at el., 2024) in their Scoping Review have shown that research based on digital media is still insufficient to address spiritual and mental health among Generation Z. Although the effectiveness and popularity of digital methods for addressing mental health problems are growing, more studies are needed on digital platforms that address spirituality and mental health together. The successful design of value-based learning models for children in technology-rich

environments demands a balance between technological innovation and established developmental and ethical principles. The goal is not just the enrichment of knowledge, but also the creation of a strong moral and character base in the coming generation. (Alfusanah at el., 2024).

This study is based on the assumption that the combination of culture + Moral Values + Technology can offer a strong educational model for the holistic development of children. If moral and cultural teachings are redesigned to suit children's psychological level, interests, and contemporary learning habits, they can become meaningful and experiential for children rather than being cumbersome or didactic.

Contemporary child psychology indicates that children learn more effectively through stories, visual experiences, dialogues, music, and creative expressions than through abstract theories. For this reason, value-based teachings, when presented in the form of narrative, visual presentation, interactive activities, and creative mediums, can have a profound impact on children's behaviour, thinking, and emotional responses.

Modern technology can play the role of a facilitator in this process. VR technology, in particular, has shown potential in improving children's learning abilities through highly interactive and engaging simulated environments (Andryani at el., 2024). Modern technology can play the role of a facilitator in this process. VR technology, in particular, has shown potential in improving children's learning abilities through highly interactive and engaging simulated environments (Andryani at el., 2024).

The present research seeks to impart practical, cultural, and social knowledge to children by applying education, child psychology, culture and technology. The research also predicts that technology can play a positive and creative role in the moral, mental and emotional development of children if used sensitively and purposefully. In the present society, while there have been efforts to make AI safe, it is equally very important to awaken the power of the inner conscience as remarked by the great Indian poet Tulsidas in Rāmacaritamānasa – the good man grasps goodness, while the base man clings to baseness. Nectar is praised for its immortality, while poison is known for bringing death.

No matter how safe technology is made, there is a possibility of its misuse, which is more likely to cause distraction and problems for children. The findings of the research suggest that digital technology is neither completely harmful nor inherently value-creating; rather, its effect depends on how it is being used and under what guidance (Nursiti at el., 2023).

(Kristjánsson, 2025) provides an overview of the recent developments and research trends of morality education, underscoring the growing interest in digital technologies, global contexts, and diverse aspects of ethical education. It mentions that there are special issues and interdisciplinary efforts emerging in the current research on combining moral education with AI and digital technology, providing the necessary direction for the moral development of children in the digital age.

2. Related Work

Various researchers have pointed out that while integrating traditional knowledge and ethical teachings with modern technology, careful consideration of aspects of intellectual property rights, ethical challenges, and cultural preservation is necessary. At the same time, virtues and value-based education have been recognised as important for the mental and moral development of adolescents, and an understanding of developmental psychology and human-centred educational technology has also been suggested to be essential for developing effective learning models (Wang & Xu, 2024; Shane McLoughlin & Kristján Kristjánsson, 2025; Gupta, 2025; Nurhabibah, 2025).

A study found that with the help of Technology-Enhanced Learning (TEL), Islamic religious education in early childhood can be made more interesting and value-added. In the research, digital content (such as animated stories and interactive applications) increased children's learning interest, social interaction, and positive character traits (Sulastri & Ismail, 2025;Mesurado at el., 2025) developed a web-based intervention called "Little Hero", which aimed to promote moral values in children. Their research found that this program increased children's co-morbid behaviours, empathy, and positive emotional responses (Mesurado & Resett, 2025).

(Betawi, 2023) found that telling moral stories to preschool children significantly improved their integrity values, such as honesty, empathy, respect, and courage, making it clear that culturally grounded narrative content can have a positive impact on children's moral and emotional development (Betawi, 2023). Christian education has also been redefined as a transformative teaching method that enables religious communities to meaningfully assimilate the gospel in a digital, cultural, and ecclesiastical environment (Clair, 2025).

The findings suggest that the integration of educational technology positively contributes to the development of religious and moral values in early childhood, offering numerous benefits for promoting these essential qualities (Warmansyah at el., 2023).

Modern psychological studies have also shown that school-based mindfulness and

self-compassion programs develop self-regulation and compassion in children, which are foundational qualities for ethical conduct (Razza at el., 2025). Contemporary scholarship emphasises that cultural heritage and traditional knowledge are not merely relics of the past but dynamic resources that contribute to social cohesion, identity formation, and ethical awareness. Integrating cultural heritage into education strengthens learners' connection with community values and fosters responsible citizenship (Heritage Science, 2021;Trček, 2022). Although recent research on Artificial General Intelligence (AGI) highlights its transformative potential in education, it primarily focuses on personalisation and cognitive outcomes, with limited attention to culturally grounded moral and emotional development (Latif at el., 2023).

Globally, institutions such as the Jubilee Centre for Character and Virtues, Character Lab, and Centre for Curriculum Redesign emphasise virtue development, resilience, and ethical reflection, while UNESCO, OECD, Stanford Centre for Ethics and the Partnership on AI are contributing to the development of responsible, human-centred, and inclusive technology. In the Indian context, the Indira Gandhi National Centre for the Arts (IGNCA) has digitised cultural texts, monuments, and spaces. EdTech platforms such as DIKSHA and Jio Shiksha have increased the reach of digital education. The Laboratory for Computational Cultural Dynamics and various spiritual organisations are promoting value-based learning with AI chatbots, mobile applications, and immersive technologies. Multi-pronged efforts are being made through initiatives such as the Ministry of Culture, NEP 2020, NCERT, e-Pathshala, Scheme for Safeguarding the Intangible Cultural Heritage, Guru Shishya Parampara Scheme, Ek Bharat Shreshtha Bharat and PM SHRI Schools for cultural preservation and transfer of knowledge to children. Further in the Indian context, a number of initiatives are being taken to inculcate culture, moral values and traditional knowledge among children in different states. For example, the Happiness Curriculum provides education based on meditation, self-reflection, and emotional balance. This work is also being done by various universities at the academic level, so that the texts can be easily understood. In this, many efforts have been made, such as the digitisation of Purāṇas in Indian culture and online searches of databases of technical terms of Sāṅkhya-Yoga. Online search and indexing systems have also been developed for texts like the Mahābhārata, the Nirukta, the Medinīkośa, the Maṅkha-kośa, etc.

Bamberg's Cultural Informatics Research Group is using digital technology to preserve and make cultural data and knowledge (history, art, tradition) interactive, with geogames or digital archives, AI and computational models, etc., to explain cultural knowledge, but no such model has been developed with a focus on children.

Although there has been substantial work in previous research on the digitisation of cultural heritage, the use of technology in moral education, and the computational analysis of cultural contexts, there has been limited attention to the development of AI-based educational models for systematic communication of cultural and moral values to children. Most studies focus either on the preservation and digitisation of cultural data or on analysing the general effects of moral education through technology. Similarly, AI research has often been limited to the analysis and prediction of cultural behaviour. The development of age-appropriate, culturally sensitive and pedagogically structured AI-based value-based learning frameworks for children is still a relatively unexplored area.

The objective of the presented research is to use AI and technology not just to provide information, but to build rational, ethical and culturally aware citizens in children. The technology, if used in coordination with cultural values, moral education, and overall personality development, can significantly reduce the potential repercussions posed by AI. The objective of this research is to make an equally serious effort towards the initiative to make AI safe for the creation of a future-oriented society, as well as to awaken the recognition of right and wrong, conscience and cultural consciousness in children, because ultimately the impact of technology depends more on the consciousness of its user than on its structure. The purpose of this research is to try to equip children with a moral conscience and critical thinking through cultural understanding.

In this context, technology can be used as a positive tool. The use of AI and digital media is to impart cultural knowledge to children, develop their thinking process, and present values by connecting them to contemporary contexts. Through technology, complex concepts can be presented in simple, intuitive, engaging, and visualised forms, making the learning process more effective.

3. Cultural Traditions of Knowledge Transmission and Memorisation

All the religious traditions of the world are rich in spiritual texts such as Hinduism, Buddhism, Jainism, Sikhism, Islam and Christianity that present not only faith-based legislation but also a structured tradition of well-organised life-philosophy, moral discipline and spiritual self-realisation. In these texts, the concepts of purpose, duty, self-restraint, social responsibility and ultimate liberation of human

life are philosophically systematic. Classical literature such as the Vedas, Upaniṣad, Tripiṭaka, Āgama, Gurubānī, Quran and Bible have been functioning as knowledge systems in their respective cultural contexts. Their core tone has been associated with the spirit of self-development, moral balance, and collective well-being. Historically, religious education was not only informative but also transformative; that is, its purpose was to build the character, consciousness and social conduct of the individual.

4. Education and memorised tradition

Memorised tradition was a central feature of ancient religious education systems. The memorisation of religious texts was not only a means of preserving the text, but also a cognitive, ethical, and spiritual practice.

For example

- In the Vedic tradition, the correct pronunciation and memorisation of the Vedas were the basis for the preservation of knowledge.
- In the Buddhist Sangha, oral memorisation of sutras was the medium of community discipline.
- The memorised practice of Agamic texts by Jain monks was associated with self-restraint.
- In the Islamic tradition, hifz (memorisation of the Quran) was considered a spiritual discipline.
- In Christian monasteries, the memorisation of hymns and scriptural passages was related to the process of meditation.

This tradition was not just for the preservation of knowledge, but was based on the belief that internalised knowledge becomes part of the individual's consciousness. When texts are not merely read, but are established in memory, they create the possibility of discovering new meanings according to the circumstances. In the present global context, the in-depth philosophical study of religious texts seems to be limited to the religious leadership class—such as priests, monastics, clerics, or clergy—respectively. The general society is attached to these texts on a symbolic, cultural or ritualistic level, but can be observed a decline in their interpretive and philosophical engagement.

The growing temporal nature of formal education, the utilitarian tendency of knowledge, and the rapid information culture of digital media have marginalised the tradition of deep learning and memory. Numerous sociological and psychological studies indicate that the current generation appears to be experiencing a crisis of moral reasoning, self-regulation, and existential clarity. Long-term verbal memory training showed changes related to plasticity in the brain, which are related to the development of areas associated with memory and attention. These results show that memory-based practice can not only be a means of knowledge preservation but also a form of cognitive training (Kumar, Singh & Paddakanya, 2021).

In some studies, religious activities, such as the recitation of sacred texts or the practice of memorisation, were also found to be positively correlated with cognitive development and mental health(Abdullah, 2025; Ganguly at el.,2021). In such a scenario, the memorised tradition of religious texts can be rethought not just as a religious practice, but as a cognitive-ethical training method. Listening, reading, or memorising the Quran can be a useful remedy for improving physical and mental health (Rozali, 2022). When a text is not just read, but is established in memory, it becomes part of the person's inner dialogue. This internalised knowledge can activate moral discernment according to the circumstances.

Religious narratives and classical themes can be helpful in developing children's imagination, language learning, and moral wisdom. In this context, "memorisation" can be considered as an educational tool for the promotion of moral-cultural literacy rather than a means of constructing a narrow religious identity(senthilkumar and shubhlakshmi,2024). Memorisation here is not merely memorisation, but the process of understanding the meaning, context, and message of the original text.

The methods of remembrance developed in the Vedic oral tradition were not merely mechanical rote memorisation, but the development of memorisation, concentration, hearing, and intellectual discipline. When students listen to and repeat an original text over and over again, the text is permanently established in their minds. Later, in different situations of life, the same memory becomes the basis of contemplation and understanding for them (Besra, 2025).

Through the chanting of the Torah, the reader does not merely read the scripture but establishes a deeply personal and community connection with it. Remembrance, music, and group interaction play a crucial role in this process, making the sacred text part of the individual's spiritual experience(Stephen, 2024). Modern digital platforms such as Tarteel, Huffaz, Mu'alim (Qur'an) and Scripture Stack (Bible) are encouraging the memorisation tradition of sacred texts using artificial intelligence, voice-recognition, and progress-tracking techniques. These tools facilitate sequential rendering of verses/verses, partial concealment for recall testing, responding to pronunciation errors, and repetition-based practice, effectively adapting the traditional memorisation method to the digital environment.

5. The Bhagavadgītā as a Value-Based Learning Framework for Children

cultures can communicate their moral and cultural knowledge to children more effectively through modern technological means. The Śrīmad Bhagavadgītā, a distinct cultural tradition, has been used as an illustrative reference in this study to illustrate how value-based education can be linked to the contemporary lives of children.

yadā yadā hi dharmasya glānir bhavati bhārata abhyutthānam adharmasya tadātmānaṁ sṛijāmyaham -Bhagavadgita, 4.7

5.1 AI-Based Shloka Recitation (TTS Integration)

In the first step of this process, the verse is pronounced correctly and clearly with the help of TTS technology.

5.2 Learner Repetition and Voice Interaction

The student will then attempt to repeat the verse spoken by AI. This process enhances active participation.

5.3 Pronunciation Analysis and Feedback

The AI-based voice-recognition system will analyse the student's pronunciation. If a word is mispronounced, such as dharmasya or glānirbhavati, the system will suggest the correct pronunciation, highlighting the mistake. In this way, the student gets a quick and personalised response.

5.4 Iterative Practice and Mastery

The student will continue to practice until he speaks the verse with precision and confidence. This process develops memory power, concentration, and linguistic correctness.

5.5 Word-by-Word Meaning Exploration

When the student is able to speak the verse with the correct pronunciation, the next step is to explain the meaning of each word of the verse in simple language.

yadā yadā – Whenever

hi – definitely / indeed

dharmasya – dharma or morality

Perform actions while being established in yoga, abandoning attachment, O Dhananjaya. Being equal in success and failure, equanimity is called yoga.

That is, by keeping the mind in a stable and balanced state

yoga-sthaḥ – established in a state of mental balance

In this research, no attempt has been made to evaluate any one religion, scripture or tradition by keeping it at the centre. It seeks to identify a universal educational approach whereby different

glāniḥ – fall or decline, weakness

That is, the loss or weakening of dharma

abhyutthānam – rise, increase

adharmasya – of unrighteousness or evil

That is, the growth or spread of unrighteousness

tadā ātmānam sṛijāmi aham – then I manifest myself

This process clarifies both language comprehension and philosophical meaning.

The Meaning of Loss of *dharma*

- The Fall of Truth: When People Begin to Lie
- Lack of justice when injustice begins to happen
- Lack of compassion: When people become unkind to each other
- Immorality: When the wrong seems to be considered right

Growth of Unrighteousness:

- When evil, violence, greed, and selfishness begin to spread in society, then *dharma* becomes weak

5.6 Contextual Storytelling and Cultural Connection

Krishna tells Arjuna that whenever there is a loss of dharma and an increase in unrighteousness on earth, then I myself incarnate in the ages to protect the virtuous, destroy the wicked and restore dharma.

Has this ever happened in the past? The incarnation of Maryādā Puruṣottama Rāma, the life of Kṛṣṇa, and the Narasimha Avatāra to protect the devotee Prahlāda are some examples. The stories of these avatāras can be made interesting and engaging through animation and storytelling. In this way, children can not only understand moral values in a better manner but also connect with them practically.

yoga-sthaḥ kuru karmāṇi saṅgaṁ tyaktvā dhanañjaya,siddhy-asiddhyoḥ samo bhūtvā samatvaṁ yoga uchyate - Bhagavad-gita 2.48

kuru– perform

karmāṇi– actions

saṅgam– attachment

tyaktvā – abandoning

Dhanañjaya – Arjuna

siddhi – success

asiddhi – failure

siddhy-asiddhyoḥ– in both success and failure

samaḥ– equal

bhūtvā – becoming

samatvam – equanimity (state of balance)

yogaḥ– yoga (state of inner balance)

ucyate – is called

This verse explains the basic principle of Karma Yoga.

Simple Meaning (Meaning) - O Dhananjaya (Arjuna), you should do your duty by being established in Karma Yoga. While performing actions, let go of attachment to the fruit. Have the same feeling in both success and failure, because this equality and balance of the mind is called yoga.

Philosophical Meaning (In-depth Understanding)

This verse explains the basic principle of Karma Yoga.

- Performing actions in a yogic manner - working by keeping the mind in peace, alertness and spiritual balance. That is, the mind should not be affected by anxiety, fear or greed while working.
- Giving up attachment to the fruit – Man should do karma, but should not be overly attached or worried about the result.
- Equal Feeling in Success and Failure – If there is success, there should be no arrogance, and if there is failure, there should be no disappointment.
- This is yoga – when the mind is balanced in every situation,

So that equanimity is the real sum.

Its meaning in life – this verse teaches that

- Do your duty with full devotion
- Don't worry too much about the result
- Keep the mind steady in success and failure

Only then does a person become calm, intelligent, and spiritually strong.

6. Core Idea

- Memorise the verse so that the verses gradually become a permanent part of his memory and contemplation
- Explain the semantics in simple words (words familiar to children) - so that they not only repeat but also understand its meaning.
- Explain the meaning of the verse - Objective- To clarify the philosophical and moral meaning of the verse, so that children develop a perspective.
- Animation, presentation by reels or short stories - Objective - To present the idealistic, theoretical, and philosophical knowledge of the scriptures in a practical form
- Establishing a Connection to the Present Life - The Scriptures Knowledge should be presented by connecting it with the daily life, problems, interests, lifestyle, etc. of the children, so that they do not consider it as mere theoretical knowledge, but also as a guide to adopt in life.
- Giving small practical tasks - The practice of equanimity, the practice of doing good deeds without any result, the practice of helping others selflessly, etc. Purpose - Our nature is formed by our small actions. To build a good nature from childhood, give small tasks to children.

7. Technique of Data Creation and Processing

This method will generate systematic training data using sequential steps, which will be used to train AI based tools and present cultural and moral teachings in digital form.

Selected verses, narratives, and teachings from various religious texts will be compiled to prepare research data rich in cultural and moral knowledge. An integrated dataset will be created by collecting different types of information for each verse.

7.1 Resource Identification: First of all, authentic religious texts and sources will be identified. For this, various religious texts, digital repositories, academic databases, official websites of universities, research journals and authentic books will be used. Texts related to moral values and life guidance will be selected from these sources.

Ex. Śrīmad Bhagavadgītā ,Rāmāyaṇa, Mahābhārata, Manusmṛti, Brahma Vaivarta Purāṇa etc.

7.2 Scriptural Text, Story and Cultural Context Collection - Verses from selected texts will be collected, and each verse will be accompanied by relevant cultural references, historical examples, and explanatory material. This dataset will include the original Sanskrit verse, IAST transliteration, word-by-word meaning, brief interpretation, moral or cultural message, related narrative or cultural examples, and elements for each verse.

This information will be compiled on the basis of authentic books, research articles, classical commentaries and university sources.

Semantic Meaning and Moral Scenario Development - Semantic analysis of the meaning of each verse will be done to

Understand its meaning. In this, the semantic

concepts related to the verse will be identified.

In addition, various social and behavioural situations related to the shlokas will also be prepared so that these teachings can be linked to practical life contexts. Techniques such as natural language processing and concept mapping can be used for semantic analysis.

The two ślokas given above are examples of this.

7.3 Authentic Translation and Content Validation - The translation and interpretation of the verses will be verified based on authentic sources to ensure the authenticity of the dataset. Research journals, authentic books, official publications of universities, digital libraries and academic database sources will be used in this process. Based on these sources, a comparative study of the material will be done to find a correct and authentic interpretation in the dataset.

7.4 Audio and Speech Dataset Creation for AI Training - An audio dataset related to the pronunciation of verses will be generated to train the AI-based tool. This dataset will include the pronunciation of pure Sanskrit, slow speed and normal speed. Additionally, speech samples will also be compiled, taking into account different pronunciation patterns and possible errors.

The ASR system will also be trained using diverse voice datasets, including speech data from children across different age groups, as well as from both male and female voices. This approach will enable the system to accurately recognize and process children's speech.

7.5 Practical Exercises and Interactive Data Generation –

Various exercises and activities will be devised based on the classical teachings, which will aim to present the moral messages of the verses in practical terms. These activities may include the following types of content –

Situation-based questions-What should you do if someone tells a lie?

Examples related to moral judgment-Examples will be provided to help children understand right and wrong.

Interactive Exercises- You will not get angry for the whole day' or 'You will help others' can be included. Additionally, animation and multimedia content can also be produced for visual presentation. Various digital tools and animation software can be used for this.

8. AI-Supported Cultural Learning Framework

An AI-supported cultural learning framework will be developed based on the dataset prepared earlier. In this framework, the AI system will be trained by integrating different types of data (verse, meaning, narrative, moral concept, circumstance, and audio). Content verification, annotation and training process will be followed by experts at each stage to ensure the reliability of the system. The key steps of the framework will be as follows.

8.1 Knowledge database creation - The compiled dataset will be organised into a consolidated knowledge database, with each verse accompanied by the original verse, IAST transliteration, semantics, brief interpretation, ethical/cultural message, related narrative, Situational examples and audio pronunciations will be attached. This structure will facilitate easy acquisition and analysis of data for the AI system.

8.2 Meaning Generator Development – The AI model will be built based on the meaning and interpretations of the verses. The dataset will be annotated with semantic tags and verified by language experts. The model will be trained on the annotated dataset, and the accuracy of the meanings generated will be improved through an expert.

8.3 Story Generator Development – An AI-based Story Generator will be developed using narratives and cultural references. Each narrative will be linked to the concept tags of the verse and the moral concept. Through expert validation and iterative training, the model will ensure the quality and relevance of the narratives.

8.4 Scenario Interpreter Development - AI models will be trained by establishing connections between ethical concepts and behavioural situations. Verses and social\ behavioural situations will be added through concept mapping. After verification by the specialist, the model will be able to recognise the appropriate verses and moral messages according to the circumstances.

8.5 Audio Processing and Feedback Module Development - The audio module will be developed to practice and improve the pronunciation of verses. The steps will include audio recording, annotation, sound analysis by Praat, and the use of TTS and ASR technology. Results obtained from ASR will be provided with immediate feedback on errors by comparing them to the reference audio.

ASR, TTS tools, digitisation of diverse cultural data, indexing, and digital video creation of stories like the Panchatantra, e-learning1 have already been developed by school of Sanskrit and Indic studies, Jawaharlal Nehru University, Similar individual efforts have already been made, such as BharatGen's[1] Sarvam AI, Vachan, Suktam, Bhashini's[2] ASR and TTS models in various languages, and the Bible learning platform[3] If all these technologies are integrated together, a comprehensive and fruitful AI-based environment can be developed for children's cultural and moral learning.

E-learning animated stories[4]

9. Features of the Proposed AI System

The proposed AI-based cultural learning system aims to present traditional cultural and moral knowledge in a more systematic, accessible and participatory manner through modern technological means. The system will include a variety of structural and functional features through which classical teachings can be presented not just as textual content but as a holistic learning experience. The salient features of the system are as follows:

9.1 Integration of Traditional Knowledge and Artificial Intelligence - This system seeks to integrate the knowledge contained in ancient religious and cultural texts with modern AI techniques. Through this integration, traditional knowledge can be systematised, preserved and presented more effectively through technological means.

9.2 Multi-dimensional knowledge structure – A key feature of the system is the multi-dimensional knowledge structure of verses. Each verse will be accompanied by a variety of supporting information, original verse, IAST transliteration, semantic and concise explanation, moral message, related narrative or cultural context, situational examples, and audio pronunciation of the verse in an integrated form. The system will thus provide the user with an opportunity for deep and structured learning while providing different levels of information related to the verse on a single platform.

9.3 AI-based Meaning and Interpretation Feature - The proposed system will have the ability to present the meaning and concise interpretation of verses through AI-based analysis. This system will be able to present the key moral and philosophical messages of the verses in a clear and organised form based on datasets and semantic annotations.

9.4 Narrative-Based Presentation - The system will present stories and cultural references related to the shlokas to make the classical teachings more intuitive and interesting. This approach can help to understand abstract philosophical ideas in narrative terms and make the learning process more meaningful.

9.5 Pronunciation Analysis and Feedback System - The system may include a special module for the practice and correction of the pronunciation of verses.

TTS technique will be used to recite the correct pronunciation of the verses.

ASR technology will be used to recognise the user's pronunciation Tools like PRAAT can be used for sound analysis, with the help of which sound properties such as frequency, rhythm and pauses can be studied. Based on this, the system will be able to provide appropriate feedback to the user about possible pronunciation errors.

9.6 Interactive and multimedia-based learning experiences -

The system can include a variety of interactive and multimedia content, such as

- Situational questions
- Examples of ethical judgment
- Audio-visual presentation
- Animation-based interpretation

The use of these media can make the learning process more engaging, participatory and effective.

10. Expected Outcomes

The proposed AI-based cultural learning system is expected to yield the following important results.

10.1 Child-Centric Pedagogy -

This research will present a new children-centric pedagogy, which will attempt to explain traditional classical knowledge through modern technology. Through this approach, the moral and cultural teachings

[1] https://bharatgen.com/text-models/
[2] https://bhashini.gov.in
[3] https://www.scripturestack.com/
[4] https://sanskrit.jnu.ac.in/download/elearning_animated_stories.zip

contained in the scriptures will be presented in a simpler, structured and engaging manner.

10.2 Direction of safe use of digital mediums -

This AI-based system can encourage the positive and educational use of technology. Thus, an effort can be made towards using digital devices as a medium of knowledge and value-based learning instead of just a means of entertainment.

10.3 Development of Moral Values and Rationality in Children -

Through verses, stories, and moral messages, children can help in the development of discriminative understanding and willpower to distinguish between right and wrong. Thus, this system can contribute to the development of moral values from an early age.

10.4 Development of interest in classical texts -

Often, the language and style of spiritual and scriptural texts seem complicated for the new generation. This AI-based system can present these texts in a simpler and more engaging form through the meaning, narrative, and context-based presentation of verses, which can develop an interest in children and young people.

10.5 Context-Based Understanding of Scriptural Texts -

Through AI-based analysis, an attempt will be made to connect the ideas contained in the verses to different life situations. This will enable the user to understand how classical teachings can be relevant in contemporary social and personal contexts.

10.6 Development of Moral Thinking and Reasoning Ability -

Through AI-based analysis, an attempt will be made to connect the ideas contained in the verses to different life situations. This will enable the user to understand how classical teachings can be relevant in contemporary social and personal contexts.

10.6 Technological representation of traditional oral traditions -

The study of scriptures in the Indian knowledge tradition has been done through oral tradition for a long time. The proposed system can provide a new medium to preserve and understand these oral traditions through digital technologies.

10.7 Possibility of Reducing Cultural and Spiritual Distance -

Keeping in mind the growing gap between the new generation and traditional cultural-spiritual knowledge in contemporary society, the system strives to present classical knowledge through digital and AI-supported means. Thus, it can be helpful in establishing a dialogue between the current generation and cultural traditions.

11. Future Scope

11.1 Age-Adaptive Interpretation System -

In the future, this research can be expanded in the direction of how to explain the teachings of cultural and spiritual texts to children of different age groups at different levels. Through AI, the complexity, examples, and interpretation of content can be customised according to age.

11.2 Multilingual Presentation and Game-Based Learning -

This model can be further developed in different languages, enabling children from different cultural backgrounds to understand these teachings in their mother tongue. Also, cultural learning can be made more interesting and experiential through game-based learning and interactive activities.

11.3 Culturally Safe AI Systems for Children -

In the future, AI models may be developed that provide safe digital environments for children and present content that addresses cultural and ethical boundaries. This will enable children to get proper cultural direction along with the use of technology.

11.4 Use of Advanced AI and Natural Language Processing -

Further advanced AI and NLP technologies can be used to develop a deeper interpretation of verses and statements, automatic narrative generation, and better semantic understanding, so that the teaching process can be made more effective.

11.5 Mobile and Web-Based Learning Platforms -

Based on this framework, mobile and web-based learning platforms can be developed, through which users will be able to access cultural and moral teachings from any location.

11.6 Advanced Pronunciation Analysis System -

In the future, voice analysis techniques can be further developed to more accurately analyse the subtle aspects of pronunciation, such as rhythm, intonation, and sound structure, making it easier to learn the correct pronunciation of a verse or traditional text.

11.7 Use of this model in the study of various texts -

This approach is not limited to moral education but can also be useful for the study and presentation of various cultural and spiritual texts in the future, so that traditional knowledge can be presented in a more accessible and systematic form through modern technology.

11.8 Continuous Development of Feedback-Based Cultural Data -

Cultural data and learning materials can be continuously updated based on feedback from users and changing socio-cultural challenges, making the system more relevant and effective over time. This model can be further developed in different languages, enabling children from different cultural backgrounds to understand these teachings in their mother tongue. Also, cultural learning can be made more interesting and experiential through game-based learning and interactive activities.

12. References

Abdullah. (2025). Education in the Vedic era: A historical and philosophical study. *International Journal of Novel Research and Development*, 10(9)

Alfusanah, F., Ramada, E., Mukarohmah, A. H., Fathurrohman, A., Anwar, C., & Anwar, S. (2024). The urgency of value education in forming students' character in the era of Society 5.0. TOFEDU: *The Future of Education Journal,* 3(5):1957–1963

Andryani, R., Gernowo, R., and Negara, E. S. (2024). The potential of virtual reality technology in children's learning success. *Indonesian Research Journal in Education* (IRJE), 8(1),374–387.

Banerji, S. C. (1999). A Brief History of Dharmaśāstra. *Abhinav Publications*.

Besra, S. (2025). Memory techniques in the Vedic oral tradition and their application in education. *International Journal for Multidisciplinary Research*, 7(6).

Che Wan Mohd Rozali, W. N. A., Ishak, I., Mat Ludin, A. F., Ibrahim, F. W., Abd Warif, N. M., & Che Roos, N. A. (2022). The impact of listening to, reciting, or memorizing the Quran on physical and mental health of Muslims: Evidence from systematic review. *International Journal of Public Health*, 67, 1604998.

Clair, M. (2025). Missional contextual theology for Christian ethics and education in the age of disruption: A framework for scholars, practitioners, and faith-based educators. *ResearchGate*.

Dubey, V. K. (2012). Vedang Shiksha Sahitya me Vyasshhhiksha ek Parisheelan. Doctoral thesis, Dr. Rammanohar Lohia Avadh University, Faizabad, India.

Farisia, H. (2020). Nurturing religious and moral values at early childhood education. Didaktika Religia, 8(1).

Frananda, A., Niva, M., and Maharjan, K. (2024). The positive impact of memorizing the Qur'an on the cognitive intelligence of primary school children. *World Psychology*, 3(1):128–144.

Ganguly, M., Mohanty, S., Mishra, S., & Patra, S. (2021). Impact of Sanskrit prosody on anxiety, mindfulness, and self-concept in young adolescents: A four-armed control trial. *Yoga Mimamsa,* 53(2), 85–92.

Girish, S. and Jairam, R. (2025). Impact of Bhagavad Gita teachings on cognitive abilities of adolescents. *International Journal of Innovative Research in Technology*, 12(8).

Gupta, P. (2025). Integrating tradition and technology: Relevance of the Indian knowledge system today. *Archives.9*(14).

Kumar, U., Singh, A., & Paddakanya, P. (2021). Extensive long-term verbal memory training is associated with brain plasticity. *Scientific Reports*, 11(1), 9712.

Kok, C. L., Koh, Y. Y., Ho, C. K., Teo, T. H., and Lee, C. (2024). Enhancing learning: Gamification and immersive experiences with AI. In: TENCON 2024 – 2024 IEEE Region 10 Conference, pp. 1853–1856.

Kristjánsson, K. (2025). Recent developments in the field of moral education—and some prompts for authors, old and new. *Journal of Moral Education*, 54(4):519–525.

Lamb, M. and Brooks, E. (2026). Teaching virtue literacy: A strategy for character education in the university. *Journal of Moral Education*.

Latif, E., Mai, G., Nyaaba, M., Wu, X., Liu, N., Lu, G., Li, S., Liu, T., and Zhai, X. (2023). AGI: Artificial General Intelligence for Education. *arXiv.*

McLoughlin, S. and Kristjánsson, K. (2025). Virtues as protective factors for adolescent mental health. *Journal of Research on Adolescence*, 35(1):e13004.

Mesurado, B. and Resett, S. (2025). 'Little Hero', a web-based intervention to promote moral values among children: A study of its development and effectiveness. *Journal of Moral Education*, pp. 1–29.

Miswanto, M., Lestari, D., and Nurhayati, D. (2024). The role of digital in early childhood Islamic education. *International Journal of Early Childhood Education and Development*, 3(1):207–212.

Nurhabibah, S. (2025). The relevance of Jean Piaget's theory of moral development in addressing the challenges of the Society 5.0 era. Proceedings of International Conference on Education (*ICE Proceedings),*3(1):233–238.

Nursiti Khodijah D., S., Saptiani, S., Santi, N. E., and Utama, M. M. A. (2023). Investigation of religious and moral values in children in the digital era. *Jurnal Asy-Syukriyyah*, 24(2):212–227.

Park, S. Y., Do, B., Yourell, J., Hermer, J., and Huberty, J. (2024). Digital methods for the spiritual and mental health of Generation Z: Scoping review. *Interactive Journal of Medical Research*, 13:e48929.

Ramadhani, T., Widiyanta, D., Sumayana, Y., Santoso, R. Y., Agustin, P. D., and Al-Amin. (2024). The role of character education in forming ethical and responsible students. *International Journal of Graduate of Islamic Education* (IJGIE), 5(2)110–124.

Razza, R. A., Liu, Q., Feng, R., Hao, X., Kirkman, K. A., and Merrin, G. J. (2025). Cultivating adolescents' self-compassion through mindfulness: The role of self-regulation at both the individual- and classroom-level. Contemporary School Psychology, 29:843–854.

Ruhiyat, E., Sugiyanto, S., Sukmana, E., Mariati, M., Hayati, S., and Muliawati, K. I. (2026). The capitalization training for MSMEs actors in Waringin Jaya Village, Bojonggede Subdistrict, Bogor Regency. *TOFEDU: The Future of Education Journal,* 5(1).

Sulastri and Ismail. (2025). The application of technology-enhanced learning in the development of Islamic religious education in early childhood. Absorbent Mind: *Journal of Psychology and Child Development,* 5(1):64–75.

Senthilkumar, R., & Subhalakshmi, R. T. (2024). How religious texts influence cognitive development in children. *Journal of Science Technology and Research*, 705–714.

Trček, D. (2022). Cultural heritage preservation by using blockchain technologies. *Heritage Science*, 10(6).

University of Bamberg. (n.d.). Recent studies in cultural informatics emphasize digital heritage mapping.

Wang, W. and Xu, X. (2024). Transformation and development of intangible cultural heritage through technology. *Journal of Library & Information Science in Agriculture,* 36(1).

Warmansyah, J., Zalzabila, Z., Yuningsih, R., Sari, M., Helawati, V., and Sari, E. N. (2023). Educational technology applications for enhancing religious and moral values in early childhood development: A bibliometric analysis. *Tarbiyah al-Mustamirrah: Jurnal Pendidikan Islam,* 4(2):154–168.

Zsila, Á. and Reyes, M. G. (2023). Pros & cons: impacts of social media on mental health. *BMC Psychology,* 11(1):201.

Author Index