

Findings of the IndicGEC and IndicWG Shared Task at BHASHA 2025

Pramit Bhattacharyya¹, Karthika N J², Hrishikesh Terdalkar³,
Manoj Balaji Jagadeeshan⁴, Shubham Kumar Nigam^{1,5},
Arvapalli Sai Susmitha¹, Arnab Bhattacharya¹

¹Dept. of Computer Science and Engineering, Indian Institute of Technology Kanpur, India

²Dept. of Computer Science and Engineering, Indian Institute of Technology Bombay, India

³Dept. of Computer Science and Information Systems, BITS Pilani, Hyderabad Campus, India

⁴Dept. of Computer Science and Engineering, Indian Institute of Technology Kharagpur, India

⁵School of Computer Science, University of Birmingham, Dubai, UAE

pramitb@cse.iitk.ac.in, karthika@cse.iitb.ac.in, hrishikesh.rt@hyderabad.bits-pilani.ac.in,

manojbalaji1@gmail.com, shubhamkumarnigam@gmail.com,

arvapallisaisusmitha@gmail.com, arnabb@cse.iitk.ac.in

Abstract

This overview paper presents the findings of the two shared tasks organized as part of the *1st Workshop on Benchmarks, Harmonization, Annotation, and Standardization for Human-Centric AI in Indian Languages (BHASHA)* co-located with IJCNLP-AACL 2025. The shared tasks are: (1) **Indic Grammar Error Correction (IndicGEC)** and (2) **Indic Word Grouping (IndicWG)**. For GEC, participants were tasked with producing grammatically correct sentences based on given input sentences in five Indian languages. For WG, participants were required to generate a word-grouped variant of a provided sentence in Hindi. The evaluation metric used for GEC was GLEU, while Exact Matching was employed for WG. A total of 14 teams participated in the final phase of the Shared Task 1; 2 teams participated in the final phase of Shared Task 2. The maximum GLEU scores obtained for Hindi, Bangla, Telugu, Tamil and Malayalam languages are respectively 85.69, 95.79, 88.17, 91.57 and 96.02 for the IndicGEC shared task. The highest exact matching score obtained for IndicWG shared task is 45.13%.

1 Introduction

India is the most populous country and Indian languages are among the most spoken languages on this planet with $\sim 1.4B+$ speakers. According to Ethnologue (Ethnologue, 2025), four of India’s 22 official languages—Hindi, Bangla, Urdu, Telugu—are within the top-20 most spoken languages on earth. Despite this, the state-of-the-art for NLP in Indian languages lags significantly behind high-resource languages like English and Mandarin as well as medium-resource languages like Arabic. The primary objective of the *1st Workshop on Benchmarks, Harmonization, Annotation,*

and Standardization for Human-Centric AI in Indian Languages (BHASHA) workshop co-located with IJCNLP-AACL 2025 was to bridge this gap for Indian languages.

Following the objective of the BHASHA workshop, we hosted two shared tasks: (1) **Grammatical Error Correction** for Indian languages (IndicGEC) and (2) **Word Grouping** for Indian languages (IndicWG). Both GEC and WG tasks are primarily unexplored territory for Indian languages from the computational point of view.

1.1 IndicGEC

Grammatical Error Correction (GEC) system focuses on detecting and correcting grammatical errors in a sentence automatically, regardless of the language. For instance, let us consider the following English sentence: “A ten year oldest boy go to school”. A good GEC system will identify errors in the use of the superlative degree and verb agreement, correcting it to “A ten-year-old boy goes to school.” For an Indian language, e.g., Bangla (Bengali), an effective GEC system will be able to detect the spelling error in the sentence “দেওয়ালের কোনে (direction) ধুলো জমেছে” (dēōyālēra kōnē dhulō jamēchē.) and correct it to “দেওয়ালের কোণে (corner) ধুলো জমেছে” (dēōyālēra kōṇē dhulō jamēchē). The CoNLL 2013 and 2014 shared tasks (Ng et al., 2013, 2014) significantly advanced GEC research for English. However, GEC for Indian languages is still in its early stages of development. Early works on GEC for Indic languages focused on rule-based and statistical models. Sonawane et al. (2020) categorized inflectional errors for Hindi GEC, and Rachel et al. (2023) proposed Vyakaranly, a toolkit for Hindi grammar correction. Alam et al. (2007) introduced

a rule-based statistical approach, but it failed to generalize beyond simple sentences. Islam et al. (2018) attempted to generate erroneous Bangla sentences via random word swaps, insertions, and deletions. However, to the best of our knowledge, no consolidated effort has been made for GEC for multiple Indian languages. Following the lines of the CoNLL-2013 shared task (Ng et al., 2013), we conducted a shared task on GEC for five Indian languages. To the best of our knowledge, this is the first attempt to organize a shared task on GEC for multiple Indian languages.

The IndicGEC shared task focused on the grammatical error correction task for five Indian languages: **Hindi, Bangla, Malayalam, Tamil, and Telugu**. The goal of the shared task is to create systems that can receive an input sentence in any of the five Indian languages and generate the corresponding correct sentence in the same language. Table 1 depicts the expected output from the GEC system for each of the 5 languages.

1.2 IndicWG

Indian languages exhibit rich morphology, flexible word order and a high degree of agglutination. These properties lead to equivalent parallel sentences written across different languages appearing structurally different. Most NLP models operate primarily on whitespace-separated words or tokens, which are more syntactic than semantic. This problem further leads to poor cross-lingual alignment and inconsistent representations. A semantically cohesive *word grouping* proposed by Karthika et al. (2025); Dangarikar et al. (2024) offers a solution to this problem by reorganizing the sentences to group semantically complete and meaningful units. The word groups are based on inflectional units, compounded verbs, named entities, compound words, and idioms. For instance, in the Bangla sentence “শচীন তেন্দুলকার ক্রিকেটে ১০০টি শতরান করেছেন।” (*śacīna tēndūlakāra krikēṭē 100ti śatarāna karēchēna*), the word grouping methodology will group NER (Person) “শচীন তেন্দুলকার” into one group and the resulting sentence “শচীন_তেন্দুলকার ক্রিকেটে ১০০টি_ শতরান করেছেন।¹” will be used for further processing. Dangarikar et al. (2024) has shown that when sentences are represented in this way, parallel sentences across languages become more structurally aligned, aiding easier mapping, and the

¹“_” denotes the boundaries of a predicted word group.

cross-lingual correspondence becomes more systematic and more predictable.

Building on this motivation, we organized a shared task on *Word Group Identification in Indian Languages (IndicWG)*. The goal of the shared task is to establish a benchmark for automatic word group identification in Indian languages, enabling the community to systematically study and model semantically cohesive units at scale. In this task, given a plain-text sentence, systems are required to output a sequence of word groups, where each group corresponds to a semantically meaningful unit. An example for Hindi word grouping is shown below.

- **Input:** कुक आइलैंड्स दक्षिण प्रशांत महासागर के बीच में पोलिनेशिया में स्थित एक द्वीप देश है, जिसका नूजीलैंड के साथ खुला सहयोग है।
- **Output:** कुक_आइलैंड्स_दक्षिण_प्रशांत_महासागर_के_बीच_में_पोलिनेशिया_में_स्थित_एक_द्वीप_देश है, जिसका_नूजीलैंड_के_साथ_खुला_सहयोग_है।

Prior works such as Bharati et al. (1991); Karthika et al. (2025); Dangarikar et al. (2024) showed that Hindi is most affected by (lack of) word grouping since even simple case markers such as के, etc. are written with added whitespaces in modern Hindi texts. Hence, the shared task is conducted in a fully supervised, single-language setting for Hindi only. We provide training and development data for Hindi, annotated with gold word groups. Participants develop systems (rule-based, statistical, or neural) to predict word groups on a held-out Hindi test set. Model performance is evaluated using exact match (EM), measuring both boundary accuracy and group-level correctness. To the best of our knowledge, this is the first attempt to conduct a shared task on such an important task from the Indian language perspective.

2 Data Curation

In this section, we discuss the methodologies employed to curate datasets for the shared tasks IndicGEC and IndicWG, respectively.

2.1 IndicGEC

Following Bhattacharyya and Bhattacharya (2025) we categorized the grammatical errors in each of these 5 languages into 4 broad categories. These categories can be further subdivided into finer categories; however, for the IndicGEC shared task, we

Language	Wrong sentence	Expected correct sentence
Hindi	ये केवल ज्ञान अर्जन तक ही सिमित नहीं है। yē kēvala jñāna arjana taka hī simita nahīṁ hai.	ये केवल ज्ञान अर्जन तक ही सीमित नहीं है। yē kēvala jñāna arjana taka hī simita nahīṁ hai.
Bangla	দেওয়ালের কোনে ধুলো জমেছে। dēoyālēra kōne dhulō jamēchē.	দেওয়ালের কোনে ধুলো জমেছে। dēoyālēra kōnē dhulō jamēchē.
Malayalam	പിലർക്ക് ജീവൻ നാഷ്ടപ്പെട്ടു. cilarkk jivan naṣṭappat̄tu .	പിലർക്ക് ജീവൻ നാഷ്ടപ്പെട്ടു . cilarkk jivan naṣṭappet̄tu .
Telugu	దినికి రెండు రకాల పరిశాయలు ఉంచాయి dīniki reṁdu rakālā parināmālu um̄tāyi	దినికి రెండు రకాల పరిశాయలు ఉంచాయి. dīniki reṁdu rakālā parināmālu um̄tāyi.
Tamil	தொழிற்சாலை இயந்த்துன் சத்தம் tolircālai iyanattan cattam	தொழிற்சாலை இயந்தரத்துன் சத்தம் tolircālai iyanatarattan cattam

Table 1: Examples of wrong and corresponding corrected sentences for GEC across languages

focused only on these four categories, analogous to the CoNLL-2013 shared task (Ng et al., 2013). The categories are as follows:

- **Spelling Errors:** Spelling errors include both the errors about the presence of non-dictionary words and Homonym errors.
- **Word Errors:** Word errors encompass all types of errors at the word level other than spelling errors. It encompasses tense errors, person errors, case errors, gender, and number errors for Indian languages.
- **Punctuation Errors:** Punctuation errors comprise all the errors that occur due to the omission of punctuation markers and the use of wrong punctuation markers.
- **Multiple Errors:** Sentences containing multiple errors of the same kind or of different kinds fall under this category.

We then further employed the methodology applied by Bhattacharyya and Bhattacharya (2025) in their Bangla GEC work to collect data for the shared task. We organized a survey asking participants to write an essay on a particular topic. Each participant was asked to write an essay within 20 minutes comprising at least 15 sentences and 150 words, on a topic chosen by them from a set of choices. The survey took place in a proctored environment to generate an exam-like situation, enabling us to gather real-world data (with errors) on all five languages. Table 2 shows the data statistics for each of the five languages on which the shared task is conducted.

This dataset has been used to conduct the IndicGEC task.

Language	Train	Dev	Test
Hindi	599	107	236
Bangla	659	102	330
Malayalam	312	50	102
Tamil	91	16	65
Telugu	603	100	315

Table 2: Dataset statistics (number of sentences) for IndicGEC task for different languages

2.2 IndicWG

The dataset for IndicWG has been created following the methodology described by Karthika et al. (2025). We employed a rule-based methodology to generate word-grouped sentences. Some of the basic rules followed to identify the word groups in the sentences are (i) *Named Entities* such as name of a person (grouping salutation, first name, middle name, last name), names of places, institutions etc. (ii) *Inflections*, where a noun followed by postpositions/ case-markers are grouped to form a single semantic unit, (iii) *Derivations*: verbs grouped along with auxiliary verbs in a sentence, resulting in a word group together representing a single action, and (iv) *Numbers* followed by measurement units are grouped together.

We created the dataset by using sentence sourced from the IN22 corpus (Gala et al., 2023). We automatically generated word-grouped sentences using Gemini-2.5 Pro (Comanici et al., 2025), which was guided by a linguistically informed prompt (refer to Figure 1) that helped perform the grouping of only semantically cohesive units such as named entities, compound nouns, complex verbs, number-unit, inflectional unity, and derivational unity expressions using underscores, while avoiding standard adjective–noun or

noun–verb combinations. All automatically generated groupings were subsequently proofread and corrected by two language experts to ensure high-quality annotations. We have thus created a dataset of 876 Hindi sentences, along with their word-grouped variants. These sentences were provided for the shared task, comprising 550 sentences as a training set, 100 sentences for validation, and 226 sentences for testing. The total number of sentences was less than 1000, and thus the task was conducted in a low-resource setting, in accordance with the low-resource nature of Indian languages.

3 System Submission

In this section, we summarize the methodologies adopted by the participating teams for both the shared tasks.

3.1 IndicGEC

A total of 14 teams participated in the final phase of IndicGEC (32 teams participated in the dev phase) across 5 languages. In the final phase, 11 teams participated in GEC task for Hindi, while 8 teams for Bangla, 8 teams for Telugu, 9 teams for Tamil, and 8 teams for Malayalam participated respectively. Of these 14 teams, 9 submitted system papers describing their methodologies. Almost all the teams that participated in this task attempted to create generic systems that can be applied to multiple Indian languages. In this section, we summarise the findings of these papers corresponding to the IndicGEC task. Table 3, Table 4, Table 5, Table 6, and Table 7 depict the leaderboard for final phase of GEC for Hindi, Bangla, Telugu, Tamil, and Malayalam respectively.

Rank	Team	GLEU
1	AiMNLP	85.69
2	OneNRC (Vajjala, 2025)	84.31
3	akhilrajeevp	81.44
4	devdot	80.75
4	priyam_saha17	80.75
6	HindiLlama	80.72
7	Horizon	80.44
7	villa_vallabh	80.44
9	Niyamika	79.47
10	A3-108	79.45
11	Dynamic Trio	72.67

Table 3: Leaderboard for final phase of Hindi GEC

Rank	Team Name	GLEU
1	priyam_saha17	95.79
1	Dynamic Trio	95.79
3	devdot	93.20
4	AiMNLP	92.86
5	A3-108	92.44
6	Horizon	82.69
7	Niyamika	81.83
8	hmd_123	57.75

Table 4: Leaderboard for final phase of Bangla GEC

Rank	Team Name	GLEU
1	priyam_saha17	88.17
2	AiMNLP	85.22
3	Niyamika	85.03
4	OneNRC (Vajjala, 2025)	83.78
5	A3-108	81.90
6	Horizon	72.00
6	villa_vallabh	72.00
8	ramanirudh	69.56

Table 5: Leaderboard for final phase of Telugu GEC

Team Niyamika have tried to simulate realistic grammatical errors by applying both character-level and word-level augmentations. For this purpose, the authors created an additional corpus of 10,000 sentences from the IndicCorp v2 dataset (Doddapaneni et al., 2023). They performed random insertion, deletion, and swapping of either characters or words in these sentences to generate erroneous sentences. The authors augmented 7,000 sentences for each language and added them to the provided dataset. Using IndicTrans, the authors then performed transliteration to change the non-native script words to their canonical form. On that dataset, the authors fine-tuned mT5 (Xue et al., 2021) and achieved GLEU scores of 79.47, 81.43, 89.77, 84.48, and 85.03 for Hindi, Bangla, Malayalam, Tamil and Telugu languages, respectively. The authors observed various inconsistencies with the data, specifically with spaces around punctuation marks or inside quotations and acronyms. They have also observed a few transliteration errors.

Team Horizon categorized the errors specifically for Hindi into 12 categories following the categorization of Bhattacharyya and Bhattacharya (2025) for Bangla. Like Niyamika, they also cu-

You are an expert in Hindi linguistics. Your task is to process Hindi sentences by applying the following specific formatting rules for grouping.

Grouping Rule () *Semantically cohesive word groups are to be combined and marked using underscore ().*

This grouping applies only to words that form a single, inseparable semantic unit.

Group these:

- Multi-word proper nouns and named entities: e.g., फर्नांडो_अलोन्सो, हंगेरियन_ग्रैंड_प्री, यू_एस_सेना
- Noun with postposition or case-marker: e.g., घर_से, लड़के_ने, खाने_के_लिए
- Specific technical terms or compound nouns: e.g., पिट_स्टॉप, सुरक्षा_कार
- Compound verbs: e.g., गूँज_उठा, बात_करना, करना_चाहिए, जा_रहा_था
- Number and unit: e.g., 10_ग्राम, 6_लाख, 400_मीटर

Do NOT group:

- Standard adjective + noun phrases
- Standard noun + verb phrases

Instruction: Whenever the user provides one or more Hindi sentences, apply the above grouping rules to those sentences and return the processed version. Grouping should be done only if the words are semantically cohesive units as explained above. Output each processed sentence on a new line. Note that each sentence is independent of each other.

Figure 1: Prompt used for generating word-grouped sentences using Gemini

Rank	Team Name	GLEU
1	AiMNLP	91.57
2	jharishr	86.52
3	ashwinarumugam	86.30
4	priyam_saha17	86.29
5	Horizon	86.03
5	villa_vallabh	86.03
7	A3-108	85.52
8	DLRG	85.34
9	Niyamika	84.48

Table 6: Leaderboard for final phase of Tamil GEC

rated a corpus out of IndicCorp v2 and generated erroneous sentences by injecting noise following grammatical rules. Team Horizon used 42 grammatical rules to generate erroneous sentences. On the curated dataset, they finetuned mT5 and IndicBART (Dabre et al., 2022). mT5 performs significantly better than IndicBART and achieves scores of 82.69, 80.44, 86.03, 72.00, 84.36 for Bangla, Hindi, Tamil, Telugu and Malayalam respectively.

The **Dynamic Trio** team also fine-tuned IndicBART (Dabre et al., 2022) for Hindi and Bangla

Rank	Team Name	GLEU
1	devdot	96.02
2	DLRG	95.06
3	priyam_saha17	94.42
4	A3-108	94.16
5	AiMNLP	92.97
6	akhilrajeevp	92.41
7	Niyamika	89.77
8	Horizon	84.36

Table 7: Leaderboard for final phase of Malayalam GEC

grammatical error correction using the IndicGEC 2025 datasets, applying minimal pre-processing to retain natural error patterns. Their system achieved GLEU score of 72.67 for Hindi, standing 11th in the leaderboard, while scoring a GLEU of 95.79 for Bangla, ranking first alongside team priyam_saha17, demonstrating the effectiveness of multilingual pretraining for low-resource Indic GEC. The team shows that strong generative performance is possible even without synthetic augmentation when leveraging an Indic-pretrained model. Their system focused on correcting syntactic ordering, morphological agreement, and punctuation.

tuation errors without over-editing.

Team **Akhilrajeevp** developed an augmentation-free GEC system for Hindi and Malayalam by applying Instruction Fine-Tuning (IFT) to Gemma-3 12B (Team et al., 2025) using LoRA adapters, combined with a deterministic, classifier-guided prompting strategy. Their minimal-edit decoding achieved 81.44 GLEU in Hindi, standing 3rd in the leaderboard and 92.41 GLEU in Malayalam (6th place), showing that careful prompt/decoding design can be competitive even under sub-1000-example supervision.

Team **AiMNL**P explored prompt-driven approaches to perform the IndicGEC task, leveraging three large instruction-tuned models *viz.*, GPT 4.1 Mini, Gemini-2.5-Flash(Comanici et al., 2025), and Llama-4-Maverick-17B-128E-Instruct to perform the task at inference time, with zero-shot and few-shot prompting. Additionally, they also experimented with a LoRA-based fine-tuned Sarvam-M 24B baseline. The team achieved strong multilingual results—ranking 1st in Tamil (GLEU 91.57) and Hindi (85.69), 2nd in Telugu (85.22), 4th in Bangla (92.86), and 5th in Malayalam (92.97), demonstrating the effectiveness of careful prompting of LLMs, while emphasising the importance of language-specific fine-tuning for achieving robust and culturally consistent error correction in low-resource Indic contexts.

Team **A3-108** tackled the Grammatical Error Correction (GEC) for five low-resource Indic languages (Bangla, Hindi, Malayalam, Tamil, and Telugu) by framing the task as a monolingual machine translation problem. The proposed approach utilized a two-stage pipeline that first generates large-scale synthetic noisy-to-clean training data using Statistical Machine Translation (SMT) on monolingual corpora, followed by training Transformer-based models. To optimize performance, the models(Ott et al., 2019) employ an Asymmetric Byte Pair Encoding (BPE) strategy, utilizing different vocabulary sizes for the source (erroneous) and target (corrected) text to better capture language-specific error patterns. The team achieved competitive results, securing 4th for Malayalam(GLEU 94.16), 5th for Bangla(GLEU 92.44) and Telugu(GLEU 92.44), 7th for tamil(GLEU 85.52) and 10th for Hindi(GLEU 79.45).

Team **DLRG** presented a hybrid neurosymbolic architecture for Grammatical Error Correction (GEC) in Tamil and Malayalam, strategically

combining neural generalization with precise symbolic rule-based pattern matching. Pre-trained mT5 models(Xue et al., 2021) i.e. mT5-base for Tamil and mT5-small for Malayalam, were finetuned using Parameter-Efficient LoRA adaptation on aggressively augmented datasets to address data scarcity and morphological complexity. Ensemble mechanism was employed to select the best output from exact matches, neural predictions, or rule-based corrections, utilizing strict safety thresholds to prevent catastrophic deletions or over-corrections. The approach achieved impressive results on the IndicGEC blind test sets, securing 2nd position for Malayalam(GLEU 95.06), and 8th position for Tamil(GLEU 85.34), thus demonstrating that the hybrid neurosymbolic architecture, offers a robust and effective solution for Grammatical Error Correction in extremely low-resource Indic languages.

3.2 IndicWG

Only 2 teams made the submissions in the test phase of the IndicWG task. One of them, team name **Melba247** has achieved an exact marching score of 44%. The team has not submitted a system description paper outlining their methodology and, therefore, we are not summarising their methodology for this task. Team **Horizon**, on the other hand, employed a model to model the word-grouping task as a sequence classification problem and finetuned MuRIL (Khanuja et al., 2021) to achieve an EM score of 58.18%. We discuss the method applied by team **Horizon** in this section.

- **Data Augmentation:** Team Horizon augmented the given dataset with a publicly available Hindi dataset consisting of 5,000 annotated sentences (Mishra et al., 2024). The augmentation is based on a rule-based local word group finder² that uses chunk labels and POS tags to form noun and verb groups. Augmenting the given data achieves an EM score of 30.58% using MuRIL.
- **Weighted Loss:** The authors probed into the dataset and found out that word-grouping datasets typically have many tokens aligned to the ‘O’ label (delimiters), producing an ‘all-O’ bias. To address this, they compute the simple inverse frequency of class weights from the training labels and use a custom

²<https://github.com/Pruthwik/Rule-Based-LWG>

“weighted” loss wrapper around the standard cross-entropy to slightly upweight B and I labels during training. Application of class weighting improves the exact matching score by 1-2% against the baseline score of 45.13%.

- **Decoding and Reconstruction:** In this stage, the authors converted the predicted label IDs to BIO tags and then reconstructed grouped sentences by concatenating words labelled as the same group. Exact-match computation compared the reconstructed grouped sentence with the given gold standard sentences.

The application of this approach yielded an EM score of 58.18%, using MuRIL as the pre-trained model. The authors observed that for the wrongly predicted sentences, the model either over-merges (54.8% of all wrong sentences) or over-splits (31.5% of all wrong sentences). The authors also observed that the EM score is 63.27% for sentences with <20 words, 45.99% for sentences with 21 to 40 words, and only 20% for sentences with >40 words, highlighting the sensitivity of the task with respect to sentence length. Team Horizon also showed that models that preserve casing and have better Indic vocabularies, such as MuRIL, produce fewer tokenization errors and thus perform better than other models, IndicBert v2 (Doddapaneni et al., 2023).

4 Conclusions and Future Work

We have organized two shared tasks, IndicGEC and IndicWG, at the BHASHA workshop co-located with IJCNLP-AACL, 2025. A total of 37 teams participated in the development phase of the task, and 14 teams participated in the final phase. For IndicGEC, the highest GLEU scores obtained are 85.69 for Hindi, 88.17 for Telugu, 95.79 for Bangla, 91.57 for Tamil and 96.02 for Malayalam. For IndicWG 45.13% is the maximum exact-matching score, which has been enhanced to 58.18% by applying a weighted loss and decoding reconstruction method. From the EM score, it is evident that the Indic word grouping is a challenging task. On the other hand, teams have scored quite highly on the IndicGEC task. It may be due to a lack of data, which fails to capture the lexical diversity of a language. We hope that these shared tasks will provide impetus to grammatical error correction and word grouping for Indian languages. In future, we will collect more handwritten

data for IndicGEC and may use a literary corpus (Bhattacharyya et al., 2023) to capture the lexical diversity of the languages.

5 Limitations

The data provided for the tasks was insufficient to fine-tune pre-trained transformer models, as noted by all participating teams. However, handwritten data is not readily available. Despite conducting a multi-week survey effort, we were able to gather fewer than 1,000 handwritten sentences. In addition, the handwritten data may not be lexically diverse. Literary data may help in the curation of a lexically diverse dataset. However, large corpora of literary data are not readily available for languages other than Bangla.

6 Ethics Statement

We have made efforts to ensure that the curated corpus is devoid of any objectionable statements. We have also conducted a manual essay writing survey to gather real-world errors. The participants have kindly allowed us to use their essays for research purposes; hence there is no copyright infringement in curating the dataset.

References

- Md. Jahangir Alam, Naushad Uzzaman, and Mumit Khan. 2007. [N-gram based statistical grammar checker for bangla and english](#).
- Akshar Bharati, Vineet Chaitanya, and Rajeev Sangal. 1991. Local word grouping and its relevance to indian languages. *Frontiers in Knowledge Based Computing (KBCS90)*, VP Bhatkar and KM Rege (eds.), Narosa Publishing House, New Delhi, pages 277–296.
- Pramit Bhattacharyya and Arnab Bhattacharya. 2025. [Leveraging LLMs for Bangla grammar error correction: Error categorization, synthetic data, and model evaluation](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 8220–8239, Vienna, Austria. Association for Computational Linguistics.
- Pramit Bhattacharyya, Joydeep Mondal, Subhadip Maji, and Arnab Bhattacharya. 2023. [VACASPATI: A diverse corpus of Bangla literature](#). In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1118–1130, Nusa Dua, Bali. Association for Computational Linguistics.

- Gheorghe Comanici, Eric Bieber, Mike Schaeckermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.
- Raj Dabre, Himani Shrotriya, Anoop Kunchukuttan, Ratish Puduppully, Mitesh Khapra, and Pratyush Kumar. 2022. Indicbart: A pre-trained model for indic natural language generation. In *Findings of the Association for Computational Linguistics: ACL 2022*. Association for Computational Linguistics.
- Chaitali Dangarikar, Arnab Bhattacharya, Karthika N J, Hrishikesh Terdalkar, Pramit Bhattacharyya, Annarao Kulkarni, Chaitanya S Lakkundi, Ganesh Ramakrishnan, and Shivani V. 2024. *Samanvaya: An Interlingua for Unity of Indian Languages*. Central Sanskrit University, India.
- Sumanth Doddapaneni, Rahul Aralikatte, Gowtham Ramesh, Shreya Goyal, Mitesh M. Khapra, Anoop Kunchukuttan, and Pratyush Kumar. 2023. Towards leaving no Indic language behind: Building monolingual corpora, benchmark and models for Indic languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12402–12426, Toronto, Canada. Association for Computational Linguistics.
- Ethnologue. 2025. Ethnologue: Languages of the world.
- Jay Gala, Pranjal A Chitale, A K Raghavan, Varun Gumma, Sumanth Doddapaneni, Awanth 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*.
- Sadidul Islam, Mst. Farhana Sarkar, Towhid Hussain, Md. Mehedi Hasan, Dewan Md Farid, and Swakkhar Shatabda. 2018. Bangla sentence correction using deep neural network based sequence to sequence learning. In *2018 21st International Conference of Computer and Information Technology (ICCIT)*, pages 1–6.
- N J Karthika, Adyasha Patra, Nagasai Saketh Naidu, Arnab Bhattacharya, Ganesh Ramakrishnan, and Chaitali Dangarikar. 2025. Semantically cohesive word grouping in indian languages. *Preprint, arXiv:2501.03988*.
- Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, Shruti Gupta, Subhash Chandra Bose Gali, Vish Subramanian, and Partha Talukdar. 2021. Muril: Multilingual representations for indian languages. *Preprint, arXiv:2103.10730*.
- Pruthwik Mishra, Vandana Mujadia, and Dipti Misra Sharma. 2024. Multi task learning based shallow parsing for indian languages. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 23(9).
- Hwee Tou Ng, Siew Mei Wu, Ted Briscoe, Christian Hadiwinoto, Raymond Hendy Susanto, and Christopher Bryant. 2014. The CoNLL-2014 shared task on grammatical error correction. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–14, Baltimore, Maryland. Association for Computational Linguistics.
- Hwee Tou Ng, Siew Mei Wu, Yuanbin Wu, Christian Hadiwinoto, and Joel Tetreault. 2013. The CoNLL-2013 shared task on grammatical error correction. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–12, Sofia, Bulgaria. Association for Computational Linguistics.
- Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. *arXiv preprint arXiv:1904.01038*.
- S. Rachel, S. Vasudha, T. Shriya, K. Rhutuja, and Lakshmi Gadherkar. 2023. Vyakaranly: Hindi grammar & spelling errors detection and correction system. In *2023 5th Biennial International Conference on Nascent Technologies in Engineering (ICNTE)*, pages 1–6.
- Ankur Sonawane, Sujeet Kumar Vishwakarma, Bhavana Srivastava, and Anil Kumar Singh. 2020. Generating inflectional errors for grammatical error correction in Hindi. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 165–171, Suzhou, China. Association for Computational Linguistics.
- Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.
- Sowmya Vajjala. 2025. Indicgec: Powerful models, or a measurement mirage? *Preprint, arXiv:2511.15260*.
- Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. *Preprint, arXiv:2010.11934*.