# Grammar Feedback for Non-Native Hindi Learners

**Uddip Yalamanchili**
Dept. of Computer Science
George Mason University
uyalaman@gmu.edu

**Aditya Shah**
Dept. of Computer Science
George Mason University
ashah49@gmu.edu

**Nikhil Chukka**
Dept. of Computer Science
George Mason University
nchukka@gmu.edu

## 1 Introduction

### 1.1 Task / Research Question Description

The task is to develop a tool that provides grammar feedback for non-native Hindi learners. The goal is to help learners identify their grammatical errors, facilitating their journey toward fluency.

### 1.2 Motivation and Limitations of existing work

Existing resources primarily focus on vocabulary and basic phrases, but they lack personalized grammar feedback. While others have addressed general Hindi language learning, this project aims to offer real-time, individualized feedback that enhances learners' understanding of grammar, which existing solutions do not sufficiently provide. Prior efforts have not been able to cater to personalized grammar refinement, especially for self-directed learners.

### 1.3 Proposed Approach

The proposed approach*(tentative approaches)* includes integrating rule-based and pre-trained NLP models to detect grammatical errors and provide suggestions. Initial ideas involve using existing public datasets, pre-trained models and machine translation tools. The development of an API that allows users to input sentences and receive corrections and explanations is also planned.

### 1.4 Likely challenges and mitigations

Challenges include accurately detecting complex grammatical errors, handling diverse sentence structures, and ensuring the tool's performance. To mitigate these, the project plans to use a combination of rule-based systems and machine-learning models. Additionally, there are contingency plans to involve human evaluations and real-world user testing to refine and improve the accuracy of the feedback tool.

## 2 Related Work

### 2.1 Vyakranly: Hindi Grammar & Spelling Errors Detection and Correction System:

Vyakranly is an automated tool for Hindi that detects and corrects both spelling and grammar errors using a combination of rule-based and statistical methods, including morphological analysis and part-of-speech tagging. It is designed for simple sentences but can handle some compound structures, distinguishing itself by integrating grammar checking, spell correction, and translation between Hindi and English (no actual implementation). However, the model struggles with complex sentence patterns and may produce less accurate results when dealing with intricate grammatical structures or idiomatic expressions.(S. et al., 2023)

### 2.2 Detection and Correction of Grammatical Errors in Hindi Language Using Hybrid Approach:

This article describes a Hindi grammar-checking system created with a hybrid approach that incorporates statistical and rule-based approaches. The approach efficiently corrects four common grammatical problems, including number and gender-related adjective-noun and noun-verb agreement difficulties. It uses a combination of morphological analysis, part-of-speech tagging, and pattern-based bigram and trigram models to detect and fix problems in short Hindi phrases. The system produced strong performance measures, including an accuracy of 0.83, recall of 0.91, and F-measure of 0.87. However, its architecture is largely focused on basic sentence constructions, restricting its ability to handle more complex grammatical

types. (Mittal et al., 2019)

### 2.3 Frequency-based Spell Checking and Rule-based Grammar Checking :

In this paper, authors developed a hybrid system combining frequency-based spell checking with rule-based grammar checking, focusing on implementing comprehensive tense rules using JSON for rule representation and demonstrating effective encoding of grammatical rules through POS tagging and parsing. (Singh et al., 2016)

### 2.4 Bangla Grammatical Error Detection Using T5 Transformer Model:

The Bangla Grammatical Error Detection paper (2023) showcased the application of T5 Transformer (60M parameters) fine-tuned on 9,385 sentences with error symbols, achieving a Levenshtein Distance of 1.0394 through extensive post-processing and specialized correction mechanisms for morphologically rich language features. (Shahgir and Sayeed, 2023b)

Our work differs by implementing a multi-model architecture specifically for Hindi that combines neural models with linguistic rules, moving beyond the purely rule-based approach of Singh et al. and the single-model architecture of the Bangla work, while incorporating Hindi-specific linguistic features and error patterns.

## 3 Methodology

To develop a tool that provides personalized grammar feedback for non-native Hindi learners, we propose a hybrid approach that combines the power of transformer-based language models with rule-based grammar correction mechanisms. Our solution is designed to provide real-time and personalized feedback to facilitate the learners' journey toward fluency in Hindi. The approach is as follows, data preparation, model selection and training, error detection, error correction, and personalization.

**Data Preparation**: We planned to create a high-quality dataset of 20,000 Hindi sentences, both correct and incorrect, simulating common errors made by non-native speakers. This helps the model effectively learn error patterns for accurate corrections.

**Model Selection and Training**: Planning to fine-tune a small HindiT5 model (60M parameters) for grammatical error detection. The model was trained for 100 epochs using 80% of the data, optimizing with AdamW.

**Error Detection**: The model uses an encoder-decoder structure to identify and correct common grammatical errors. It detects issues like incorrect gender or number agreement and verb conjugations.

**Post-Processing and Error Correction**: Post-processing involves character-level corrections, dictionary lookups, and regular expressions to refine the output. This ensures grammatical correctness while retaining sentence meaning.

**Evaluation Metrics**: We evaluate the model using Levenshtein Distance, Precision, Recall, and F1 score.

**Reproducibility**: We document the dataset, model architecture, and hyperparameters to ensure reproducibility. Regular expression rules and model scripts will also be provided for further research.

**Evaluation**: We plan for both automated and human evaluation for comprehensive analysis. This approach ensures the model's effectiveness across various user proficiencies.

**Automated Evaluation**: We used metrics like Levenshtein Distance and F1 score to objectively assess corrections. This provides consistency in performance evaluation across grammatical error types.

**Human Evaluation**: The human evaluation addresses nuances that automated metrics might miss.

**Hybrid Approach**: A hybrid of automated metrics and human evaluation will be used to balance efficiency and qualitative insights. This ensures a comprehensive understanding of model performance.

## 4 Experiments

### 4.1 Datasets

We found a dataset that has both grammatically correct and incorrect examples. The incorrect sentences were generated by simulating common mistakes made by non-native speakers, such as errors in adjective-noun agreement, verb conjugations, and incorrect postpositions. The added errors were bracketed using a special character to mark the erroneous parts of the sentences, similar to the approach outlined by (Shahgir and Sayeed, 2023a) for Bangla grammatical error detection. This an-

notation makes it easier for the model to understand and identify error patterns during training.

### 4.2 Implementation

We were initially unable to find a dataset that suited our needs, most of the datasets that we found only had grammatically correct sentences. So, we had to figure out a way to implement or incorporate grammatically incorrect data into the dataset. Once, we had done that, we started to implement a simple rule-based system that could check if the given sentence is grammatically correct or not, we are still yet to complete that and then we are going to implement a neural-based approach.

### 4.3 Results

We were able to incorporate incorrect grammar into the data, that is our result at the moment and we plan on working further on the implementation part.

### 4.4 Discussion

We are yet to remain with the results and implementation part.

### 4.5 Resources

Large development time is needed for data collecting and model fine-tuning, as well as large computational resources, such as GPUs for model training, in the suggested method. Data annotation, appraisal, and language validation all require human labor, therefore effective development and quality control call for a small team of specialists.

### 4.6 Error Analysis

There is no analysis as we do not have any results yet.

## 5 Conclusion

In conclusion, our research created a Hindi grammar-checking system that uses statistical and rule-based approaches to detect common errors. The system is now effective for short phrases, but ongoing work intends to increase its capacity to handle complicated structures, increasing its general usefulness.

## References


M. Mittal, S. K. Sharma, and A. Sethi. 2019. Detection and Correction of Grammatical Errors in Hindi Language Using Hybrid Approach. 7(5):421–426.

Rachel S., Vasudha S., Shriya T., Rhutuja K., and Lakshmi Gadhikar. 2023. Vyakranly : Hindi Grammar & Spelling Errors Detection and Correction System. In *2023 5th Biennial International Conference on Nascent Technologies in Engineering (ICNTE)*, pages 1–6.

H. A. Z. Sameen Shahgir and Khondker Salman Sayeed. 2023a. Bangla Grammatical Error Detection Using T5 Transformer Model. ArXiv:2303.10612.

HAZ Shahgir and Khondker Salman Sayeed. 2023b. Bangla grammatical error detection using t5 transformer model. *arXiv preprint arXiv:2303.10612*.

Shashi Pal Singh, Ajai Kumar, Lenali Singh, Mahesh Bhargava, Kritika Goyal, and Bhanu Sharma. 2016. Frequency based spell checking and rule based grammar checking. In *2016 international conference on electrical, electronics, and optimization techniques (iceeot)*, pages 4435–4439. IEEE.