Transliterating Bangla to English: A Rule-Based Phonetic Approach and A Suggestion System

by

Prattoy Majumder
22266019

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
M.Sc. in Computer Science

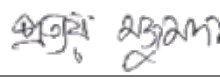
Department of Computer Science and Engineering
Brac University
April 2025

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:



Prattoy Majumder

22266019
Student ID

Approval

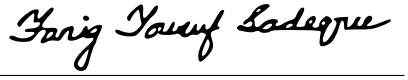
The thesis titled "Transliterating Bangla to English: A Rule-Based Phonetic Approach and A Suggestion System" submitted by

1. Prattoy Majumder (22266019)

Of Spring, 2025 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of M.Sc. in Computer Science on April 13, 2025.

Examining Committee:

Supervisor:
(Member)



Dr. Farig Yousuf Sadeque

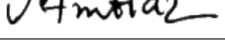
Associate Professor
Department of Computer Science and Engineering
BRAC University

Internal Examiner:
(Member)

Dr. Swakkhar Shatabda

Professor
Department of Computer Science and Engineering
Brac University

External Examiner:
(Member)



Dr. Mashrur Imtiaz

Assistant Professor
Department of Linguistics
University of Dhaka

M.Sc. Thesis Coordinator:

Dr. Md Sadek Ferdous

Professor
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Dr. Sadia Hamid Kazi

Chairperson and Associate Professor
Department of Computer Science and Engineering
Brac University

Ethics Statement

Title: Transliterating Bangla to English: A Rule-Based Phonetic Approach and A Suggestion System
Author: Prattoy Majumder
Supervisor: Dr. Farig Yousuf Sadeque
Department: Department of Computer Science and Engineering
University: BRAC University

1. Ethical Approval:

This thesis received approval from the aforementioned examining Committee to ensure adherence to ethical guidelines.

2. Data Management:

All data used in this project if created manually, in compliance with data protection regulations, stored securely, and retained according to institutional policies.

Abstract

Transliteration of Bangla to English is a major problem in Bangladesh. People use various spellings of the same word creating confusion for chat, social media posts, and other things. In this paper, I have developed a new way of transliterating Bangla words into English letters using a phonetic approach. At first, I gathered 110,000 Bangla words from various sources. Then, I converted Bangla words into the International Phonetic Alphabet (IPA) to capture their sound. Subsequently, I converted IPA symbols into English letters using an IPA chart for English dialect. Then, I made modifications to some words based on certain rules proposed by my linguists keeping user-friendliness in mind. I also took experts' advice on my transliterations to make sure these are correct and sound natural. Next, I developed a suggestion system where for one misspelled transliterated word, it will give n number of possible correct words. I wanted to have a proper balance of linguistical accuracy with user-friendliness. This data can help in language learning, computer programs that understand language, and working with digital text.

Keywords: Bangla Transliteration, Taklami Detection, Bangla Phonetics

Acknowledgement

All praise to the Great Creator for whom our thesis have been completed without any major interruption.
My supervisor Dr. Farig Yousuf Sadeque sir for his kind support and advice in our work. He helped us whenever we needed help.
My friend A.F.M. Mohimenul Joaa, for his support, motivation and contribution in the project and thesis.
The whole examining committee of the thesis.
And finally to my family without their throughout support it may not be possible.

Table of Contents

Declaration i

Approval ii

Ethics Statement iv

Abstract v

Acknowledgment vi

Table of Contents vii

List of Figures ix

List of Tables 1

1 Introduction 2
- 1.1 Preface . . . 2
- 1.2 Motivation & Significance of the Study . . . 2
- 1.3 Scope of the Study . . . 3
- 1.4 Research Design . . . 4
- 1.5 Research Questions . . . 5
- 1.6 Structure of the Thesis . . . 5

2 Background 6
- 2.1 Introduction to Transliteration . . . 6
- 2.2 Challenges in Bangla-to-English Transliteration . . . 6
- 2.3 Computational Techniques for Transliteration and Error Correction . . . 7
  - 2.3.1 In Algorithm 1 . . . 7
  - 2.3.2 In Algorithm 2 . . . 8
- 2.4 Summary . . . 9

3 Literature Review 10
- 3.1 Introduction . . . 10
- 3.2 Historical Perspective on Bangla Romanization . . . 10
- 3.3 Transliteration in Machine Translation . . . 11
- 3.4 Convergence of Digital Bangla Spellings . . . 12
- 3.5 Comparative Analysis of Transliteration Models . . . 12
- 3.6 Autocorrection in Transliteration Systems . . . 13
- 3.7 Transliteration in Other Indo-European Languages . . . 14

3.8 Conclusion . . . 15

4 Methodology **16**
4.1 Data . . . 16
4.1.1 Initial Data Collection . . . 16
4.1.2 Modern Word Collection . . . 16
4.1.3 Data Merge . . . 17
4.1.4 Ethical Considerations . . . 17
4.2 Approach Evolution . . . 18
4.2.1 Alphabet-by-alphabet Transliteration Approach . . . 18
4.2.2 Phonetic Approach . . . 18
4.2.3 Word Conversion . . . 22
4.3 Algorithms for Generating Correct Transliteration Suggestions . . . 22
4.3.1 Initial Approach: Edit Distance and Phonetic Similarity (Limited Scalability) . . . 22
4.3.2 Improved Approach: Custom Phonetic Representation (Scalable Solution) . . . 24
4.3.3 Misspelled Word Generation . . . 26

5 Experimental Results **27**
5.1 Rules . . . 27
5.1.1 Ambiguities in Transliteration . . . 31
5.2 Algorithm Results . . . 32
5.2.1 Approach 1: Small Dictionary-Based Matching . . . 32
5.2.2 Approach 2: Phonetic-Based Matching with Large Dictionary 32
5.3 Comparison and Observations . . . 33
5.4 User Interface Demonstration of Suggestions . . . 33

6 Conclusion **35**
6.1 Challenges and Limitations . . . 35
6.2 Future Direction . . . 35
6.3 Summary . . . 36

Bibliography **38**

List of Figures

4.1 Collected Word Distribution 17
4.2 Flowchart of the Transliteration Process 19
4.3 Algorithm 1 Flowchart 23
4.4 Custom Closeness of Certain Letters 24
4.5 Algorithm 2 Flowchart 25

5.1 Suggestion as Transliterated Word 33
5.2 Suggestion as Bangla Word 34

6.1 Context-Based Current Word Prediction 36
6.2 Full Sentence Correction 36
6.3 Next Word & Full Sentence Prediction 36

List of Tables

5.1 Transliteration Rules 27

Chapter 1

Introduction

1.1 Preface

In Bangladesh, it is common for people to write Bangla using English alphabets, often resulting in varied and inconsistent transliterations. This informal practice, known as "Taklami", can lead to confusion about the intended meaning. Transliterating Bangla into English is a common practice in digital communication, social media, and search queries. But, there are no universally accepted standard rules for transliteration which results in multiple variations of the same word, and it leads to inconsistencies. For example, the Bangla word "ভালো" may be transliterated as "bhalo", "valo", or even "bhallo", depending on individual preferences and phonetic interpretations. Different variations like these create challenges in natural language processing (NLP), search engines, and text-based applications, making it difficult for machines and users to interpret transliterated text correctly.

Existing transliteration systems mainly rely on statistical models, machine learning approaches, or direct character mapping. However, these methods often fail to preserve phonetic accuracy or provide meaningful suggestions for incorrect transliterations. To address this issue, I propose a standardized approach for transliterating Bangla words using English alphabets.

My approach is based on phonetics, where each word's International Phonetic Alphabet (IPA) representation will be used to guide the transliteration. However, since not all words can be accurately represented through IPA alone, I will supplement the approach with additional rules developed in consultation with linguists and interviews. This research aims to create a consistent and practical system for writing Bangla in English script.

1.2 Motivation & Significance of the Study

The motivation behind this research is twofold:

i. **Standardization of Transliteration:** A rule-based phonetic approach can provide a phonetically and linguistically correct method for transliterating Bangla into English. Unlike existing approaches, which may produce inconsistent outputs, a well-defined set of rules can improve clarity and usability.

ii. **Enhancing User Experience with a Suggestion System:** Many users make errors while transliterating Bangla words. A suggestion system can detect incorrect transliterations and provide alternative suggestions, improving the accuracy and efficiency of transliterated text in real-world applications such as chat applications, search engines, and text input systems.

The standardized approach for transliterating Bangla using English alphabets can have several practical applications across different fields. Some of the applications are described below:

i. **Text Input Systems**

- **Virtual Keyboards:** It can be integrated into virtual keyboards to help users type Bangla using English alphabets, providing standardized suggestions and auto-corrections.
- **Browser Extensions:** Can be integrated into browser extensions to help users type Bangla using English alphabets, providing standardized suggestions and auto-corrections.
- **Speech-to-Text Applications:** Can enhance speech-to-text systems by improving the recognition of Bangla words when spoken in English or transliterated format.

ii. **Natural Language Processing (NLP)**

- **Search Engines:** Improve the search results for Bangla content written in English alphabets by mapping different transliterations to standardized forms.
- **Machine Translation:** Enhance translation tools by providing more accurate transliteration between Bangla and English, aiding in the translation process.
- **Text Analysis:** Enable sentiment analysis, language detection, and text classification for mixed-script content in social media and informal communication.

iii. **Social Media and Informal Communication**

- **Improved Communication:** Reduce misunderstandings by providing a standardized way to write Bangla informally, making conversations clearer in messaging apps, forums, and social media.
- **Content Moderation:** Aid in filtering and understanding user-generated content by identifying transliterated Bangla words more accurately.

1.3 Scope of the Study

This study focused on developing a transliteration system for converting Bangla words into English using a phonetic approach based on the International Phonetic Alphabet (IPA). The goal was to create a balance between linguistic accuracy and user-friendliness by taking IPA of Bangla words and incorporating custom rules for

problematic words that were easy to understand and apply. I transliterated a dataset of 110,000 Bangla words using this approach. For words that could not be accurately transliterated using IPA alone, additional rules provided by linguistic experts were applied. A suggestion system was also developed to offer correct transliterations for misspelled words based on this vocabulary. Feedback was incorporated from linguistic experts to ensure that the transliterated output met real-world needs. The research did not address dialectal variations or regional pronunciation differences within Bangladesh.

1.4 Research Design

This study uses a mixed-methods approach both the quantitative and qualitative technique. I processed and transliterated a large dataset of 110,000 words. Then consulted linguistic experts and conducted interviews to refine transliteration rules for problematic cases.

The study has an explanatory and descriptive purpose. It seeks to explain challenges in transliterating Bangla to English accurately while describing and refining transliteration rules to address these issues.

i. **Sampling and Data Collection**

- **Data Source and Collection:** A dataset of 110,000 Bangla words was collected from recent Bangla newspapers and chat data. This dataset represents modern language usage in formal and informal contexts. Consent was obtained from participants contributing chat data.
- **Method of Transcription:** Words were initially transliterated using the International Phonetic Alphabet (IPA) combined with equivalent English alphabets to produce a baseline transliteration.

ii. **Data Processing and Analysis**

- **Manual Verification:** Each transliterated word was manually reviewed for accuracy, with particular attention given to identifying problematic transliterations that did not align with expected pronunciation or standard transliteration practices.
- **Linguistic Expert Consultation and Interview:** For words that could not be adequately transliterated using the IPA alone, input was sought from linguistic experts. Interviews were conducted among language specialists to gather feedback and formulate transliteration rules for problematic words. These additional rules were applied to refine the transliterations.
- **Rule Development:** Specific rules were developed and applied to handle cases that commonly deviate from expected transliteration patterns, such as differentiating between similar-sounding consonants.

iii. **Suggestion System Development**

- **Algorithm Selection:** An algorithm using Python was designed to give wrong transliteration's correct suggestions by integrating Syllabic Phonetic Approximation, Consonant Skeleton, and Levenshtein Distance.
- **Testing and Evaluation:** The suggestion system was tested against 1378 misspelled words to ensure that it produced accurate suggestions.

1.5 Research Questions

To address the challenges mentioned, this thesis investigates the following key research questions:

- How can a rule-based phonetic approach be designed to transliterate Bangla words into English consistently?
- What linguistic and phonetic principles should be incorporated to ensure accurate transliteration?
- How can an intelligent suggestion system be developed to identify incorrect transliterations and recommend the most probable corrections?
- What implications does a standardized transliteration system have for digital communication and NLP applications?

1.6 Structure of the Thesis

This thesis is structured into six chapters:

i. **Introduction:** This chapter provides an overview of the research topic, motivations, objectives, and outlines the structure of the thesis.

ii. **Background:** The second chapter introduces the challenges of Bangla-to-English transliteration and discusses key concepts such as Soundex, Levenshtein Distance, word similarity measures, and phonetic matching techniques, providing the necessary foundation for the proposed approach.

iii. **Literature Review:** The third chapter reviews relevant literature and existing research in the field of Bangla language and its phonetics.

iv. **Methodology:** In chapter four, how data was collected, the methodology used in approach selection, word conversion, rules generation, and suggestion algorithm creation is discussed.

v. **Experimental Results:** Chapter five presents the rules and the suggestion system I have created, using my dataset.

vi. **Conclusion:** Chapter six provides a comprehensive discussion of the findings, interpretations, and implications of the results, as well as limitations and future research directions.

Chapter 2

Background

2.1 Introduction to Transliteration

Convertion of text from one language to another language preserving its phonetic characteristics as accurately as possible is called transliteration. Unlike translation, which represents meaning, transliteration focuses on phonetic representation. Bangla-to-English transliteration is widely used in digital communication, search engines, and NLP applications, but it lacks standardization. As a result, multiple variations for the same word is created. For example, the Bangla word "ভালো" can be transliterated as "bhalo", "valo", or "bhallo", which creates inconsistencies in text processing.

Existing transliteration systems either follow simple character mappings, phonetic approximations, or machine learning-based approaches. However, handling phonetic variations, spelling inconsistencies, and misspellings are issues here. This research introduces a rule-based phonetic approach to achieve more consistent transliteration and a suggestion system to correct errors based on phonetic similarity.

2.2 Challenges in Bangla-to-English Transliteration

Several challenges arise in Bangla transliteration:

i. **Phonetic Ambiguity:** The same Bangla sound may have multiple English representations. Example: শিক্ষা → shikkha/sikhsha.

ii. **Multiple Spellings:** Different users may transliterate the same word differently. Example: রাস্তায় → rastay/raastay/rastai.

iii. **Misspellings & Typing Errors:** Users often make errors due to phonetic closeness, such as confusing b/v, f/ph, y/i, s/sh, k/c, e/a.

iv. **Lack of Context Awareness:** Some transliteration methods rely on direct mappings without considering pronunciation variations in different contexts.

v. **Regional Variations in Bangla:** Bangla exhibits significant regional variations, particularly between Kolkata Bangla (Indian Bangla) and Bangladeshi

Bangla. While both share the same script and core grammatical structure, they differ in pronunciation, vocabulary, and phonetic representation. In this paper my focus is on Bangladeshi Bangla.

To address these issues, I incorporated various computational techniques that measure word similarity and provide intelligent correction suggestions.

2.3 Computational Techniques for Transliteration and Error Correction

To improve transliteration accuracy and develop an effective suggestion system, I used several phonetic and string similarity measures. I created two suggestion algorithms. Below are some key concepts from these algorithms that are discussed:

2.3.1 In Algorithm 1

Soundex

Soundex is a phonetic algorithm that converts words into a four-character alphanumeric code based on their pronunciation. It helps identify words that sound similar despite minor spelling differences. For example:

ভালো (bhalo) → B400
ভেলো (velo) → B400

Since Bangla transliteration varies significantly, Soundex alone is insufficient, but it helps in grouping phonetically similar words.

Word Edit Distance/Levenshtein Distance

Using Levenshtein Distance algorithm I measured word edit distances. It calculates the number of edits (insertions, deletions, or substitutions) required to transform one word into another. It helps determine how similar two words are. For example:

ভালো (vhalo) → bhalo (1 substitution: 'v' → 'b') → Distance = 1
পথ (pot) → poth (1 addition: 'h') → Distance = 1

Custom Closeness Measure

Bangla transliteration errors often result from perceptually similar sounds. I wanted to make this similar sounding alphabets distance less, in another word make them closer. Some of these errors are:

v ↔ bh, v ↔ b, f ↔ ph, f ↔ p, y ↔ i, a ↔ e

Example:
"bhalo" and "valo" are kept closer.
"pother" and "pothar" (path/road) are kept closer as people mistake between 'a' and 'e'.

This method improves the suggestion ranking by prioritizing likely corrections.

Word Length Difference

Transliteration errors often introduce extra or missing letters. This measure helps prioritize words with similar lengths when ranking suggestions. For example:

"bhalo" (5 letters) → "bhalloo" (7 letters) → Difference = 2
"valo" (4 letters) → "bhalo" (5 letters) → Difference = 1 (closer match)

This ensures that suggestions favor words with minimal length variation, reducing incorrect predictions.

2.3.2 In Algorithm 2

Syllabic Phonetic Approximation (SPA)

Syllabic Phonetic Approximation or SPA is used to normalize variations in syllable pronunciation, ensuring that words with the same root form are mapped to a common base, similar to custom closeness. It helps correct:

- Variations in suffixes and vowel changes
- Assimilation of similar sounds

For example:

আবার (abar) → (eber)
রাজপথের (rajpother) → (rejpoter)
স্বাভাবিক (shavabik) → (cebebic)

This method enhances transliteration consistency by ensuring that words with minor phonetic differences map to a standardized form.

Consonant Skeleton (CS)

Building on SPA, Consonant Skeleton or CS removes vowels from SPA except the first letter of the word, keeping only the consonantal structure. This helps in matching words despite vowel variations. For example:

ভালো (valo) (Given Word) → belo (SPA) → bl (CS)
ভালো (bala) (Given Word) → bele (SPA) → bl (CS)
ভালো (bhalo) (Given Word) → belo (SPA) → bl (CS)

By focusing on consonant structures, CS helps identify words that share a common root form despite different vowel insertions.

2.4 Summary

In this section, I discussed the fundamental challenges of Bangla-to-English transliteration, highlighting phonetic ambiguity, spelling inconsistencies, and common errors. I introduced various phonetic and string similarity measures that enhance the transliteration process and enable an intelligent suggestion system.

The next chapter will present the previous works regarding transliteration process of Bangla to English.

Chapter 3

Literature Review

3.1 Introduction

The transliteration of Bangla using the English alphabet is a relatively new area of research, with limited studies addressing this problem comprehensively. The available literature provides diverse approaches, ranging from phonetic and grapheme-based methods to hybrid models that incorporate statistical and rule-based techniques. This section reviews the key studies that contribute to the understanding and development of Bangla transliteration systems, alongside relevant works in other Indo-European languages.

3.2 Historical Perspective on Bangla Romanization

Kurzon (2010) [1] examines historical attempts to romanize Bengali and other Indian scripts, focusing on two major 20th-century initiatives. The first, during British India, was influenced by linguistic and political considerations,notably through the work of S. K. Chatterji, who explored the feasibility of a Roman-based script for Bengali. The second occurred in East Pakistan before the 1971 Liberation War, where efforts were made to replace the Bengali script with Perso-Arabic or Roman scripts as part of broader language policy debates. The study finds that both attempts faced resistance due to deep-rooted linguistic identity and cultural attachment to the Bengali script. Additionally, Kurzon highlights how script changes were often driven by political agendas rather than linguistic necessity.

These findings are particularly relevant to the transliteration of Bangla into English. The resistance to script change underscores the importance of phonetic accuracy and linguistic validation in transliteration, ensuring that it aligns with native pronunciation while maintaining readability for non-native users. Unlike past romanization efforts, which sought to replace the Bengali script, modern transliteration approaches—such as those in this study—aim to preserve linguistic integrity while making Bangla accessible in digital and multilingual contexts.

3.3 Transliteration in Machine Translation

Salam et al. (2013) [3] address the problem of unknown words in English-to-Bangla machine translation (MT). They propose a method that integrates WordNet for semantic similarity and employs IPA-based transliteration when direct translations are unavailable. If a word is missing in the IPA dictionary, the Akkhor transliteration mechanism is used as a fallback. This approach ensures that words without direct Bangla translations can still be processed effectively in MT systems. Their key findings are:

- The proposed method significantly improves translation accuracy by resolving issues with out-of-vocabulary words.
- Implementing this approach in an Example-Based Machine Translation (EBMT) system led to a 16-point improvement in translation quality.
- The combination of semantic similarity and phonetic transliteration proved to be a robust strategy for handling unknown words.

This paper highlights the necessity of phonetic transliteration in machine translation, particularly for handling unknown words. In transliterating Bangla to English, a similar challenge exists—words without direct equivalents must be mapped accurately. This study supports the importance of a systematic transliteration framework that preserves phonetic integrity while ensuring meaningful representation in the target language.

In another paper, Dasgupta et al. (2015) [6] focus on creating a back-transliteration system from English to Bangla, which is essential for applications such as named entity translation, information retrieval, and localization. They developed a parallel corpus of 83,000 English-Bangla word pairs and utilized statistical models, including a joint source-channel model and a trigram-based approach, to extract transliteration units. Their key findings are:

- The corpus creation played a crucial role in improving back-transliteration accuracy.
- The statistical models used were effective in identifying transliteration patterns across languages.
- The study demonstrated that transliteration-based approaches outperform conventional dictionary-based methods for proper nouns and domain-specific terms.
- Transliteration varies between Bangla used in Kolkata and Bangladesh. For example: ‘d’ can represent ‘দ’, ‘ধ’, ‘ড়’, ‘ঢ়’, ‘ড’, ‘ঢ’ in Kolkata but in Bangladesh it can only represent ‘দ’, ‘ধ’, ‘ড’, ‘ঢ’.

This paper reinforces the significance of corpus-driven transliteration approaches, aligning with the need for a robust dataset when developing transliteration models. Since Bangla-to-English transliteration also requires accurate mapping of phonetic patterns, the insights from this study validate the necessity of a well-structured transliteration dataset and probabilistic models for improving transliteration accuracy.

3.4 Convergence of Digital Bangla Spellings

Khan (2014) [5] examines the increasing standardization of Romanized Bangla spelling in digital communication,particularly in SMS, emails, and online chatrooms. The study analyzes a dataset of 500 to 1,500 high-impact Romanized Bangla messages collected over four years to track spelling patterns and their evolution. The goal is to assess whether Romanized Bangla is moving toward a consistent spelling system that could support the development of spell-checkers and transliteration tools. Key findings of this paper are:

- Over time, the spelling of Romanized Bangla words has become more uniform. In 2011, only about 50% of words had consistent spellings, but by 2014, this had increased to around 75%. The study predicts that this trend could reach 85-90% in the following years.
- Users tend to simplify spellings for ease of typing, especially on mobile devices with limited keyboard space.
- Romanized Bangla often incorporates English words, reflecting the natural blending of languages in digital communication.
- The increasing standardization makes it feasible to develop spell-checking and transliteration tools that can automatically correct and convert Romanized Bangla into the Bangla script.

This paper provides valuable insights into how Romanized Bangla evolves in real-world usage, which is crucial for designing a robust transliteration system. The observed trends in spelling consistency support the idea that a standardized phonetic mapping can be developed for transliteration. Additionally, the study highlights the practical considerations of user behavior, such as spelling simplifications and English word integration, which must be accounted for when developing transliteration models.

3.5 Comparative Analysis of Transliteration Models

Bangla-to-English transliteration has evolved through different approaches, ranging from rule-based phoneme mapping to statistical and neural models. This section compares three key studies: Sen and Garg (2015) [7], Akan (2007) [17], and Akan (2018) [8],highlighting their methodologies, findings, and relevance to the present research.

Sen et al. (2015) provide a broad review of transliteration techniques, categorizing them into rule-based, statistical, and hybrid models. While rule-based methods offer phonetic consistency, they struggle with exceptions. Statistical models improve accuracy through data-driven learning but require large corpora. Hybrid approaches balance both, making them ideal for practical applications.

Akan (2007) focuses on a rule-based phoneme mapping system, laying the foundation for transliteration research. While effective for standard words, it struggles with loanwords and phonetic variations. Akan (2018) expands on this by integrating transliteration with machine translation, demonstrating how context-aware models enhance accuracy, particularly for proper nouns and technical terms. The study finds that statistical and neural models outperform purely rule-based approaches in real-world applications.

Key takeaways from these papers:

- **Rule-Based vs. Statistical vs. Neural Models:** While rule-based approaches ensure consistency, they struggle with exceptions. Statistical and neural methods offer flexibility but require large datasets. A hybrid approach may be optimal.
- **Phonetic Challenges:** Properly representing aspirated consonants, nasalization, and diphthongs remains a core challenge across all models.
- **Contextual Awareness:** Modern transliteration systems benefit from incorporating context, particularly for handling proper nouns and loanwords.
- **Integration with Translation Systems:** Transliteration is not just a standalone process but a crucial component in multilingual communication, especially in machine translation.

The comparative analysis of these three papers provides a historical and methodological perspective on Bangla-to-English transliteration. The findings reinforce the necessity of combining linguistic insights with machine learning techniques to develop an effective transliteration model. By leveraging rule-based phonetic mappings while integrating statistical learning, the current research aims to improve accuracy and adaptability, ensuring that transliterated words retain their intended pronunciation and meaning.

3.6 Autocorrection in Transliteration Systems

Transliteration systems often struggle with spelling inconsistencies, especially when users input phonetic variations of words. Hossain et al. (2019) [10] address this issue by introducing an auto-correction mechanism for English-to-Bangla transliteration, leveraging Levenshtein distance to detect and correct errors.

The study highlights that minor misspellings—such as missing, extra, or swapped letters—can significantly affect transliteration accuracy. By calculating the edit distance between the user's input and a predefined lexicon of correct transliterations, the system intelligently suggests the closest match. This approach enhances accuracy, particularly for words with ambiguous phonetic mappings. Their key findings are:

- Levenshtein distance effectively corrects small input errors, improving transliteration accuracy without requiring extensive manual intervention.

- The method works well for common spelling variations but may struggle with highly ambiguous or context-dependent words.
- Integrating auto-correction into transliteration systems reduces user frustration and enhances usability, making transliterated text more reliable.

Hossain et al.'s work aligns closely with the goal of refining transliteration by ensuring phonetic consistency and error correction. Since misspellings are common in real-world transliteration scenarios, incorporating Levenshtein-based auto-correction can significantly improve the reliability of transliterated text. This research reinforces the importance of error-tolerant transliteration models, which is a key consideration in the current study.

3.7 Transliteration in Other Indo-European Languages

Transliteration challenges are not unique to Bangla, other Indo-European languages face similar phonetic complexities. Studies on Punjabi, Hindi, and Kurdish transliteration provide valuable insights into the diverse approaches used to handle these challenges.

Deep et al. (2011) [2] developed a Punjabi-to-English transliteration system that relies on a rule-based approach to map Punjabi phonemes to their closest English equivalents. The study highlights the difficulty of handling Gurmukhi script's inherent vowel sounds and aspirated consonants, which often require contextual adjustments. While rule-based methods work well for structured phonetic systems, they struggle with loanwords and dialectal variations.

Kaur et al. (2014) [4] explored a hybrid approach for Hindi-to-English transliteration, specifically for proper nouns. By combining statistical models with linguistic rules, their system improved accuracy compared to purely rule-based or statistical methods. The research emphasizes that proper nouns require special handling, as their transliterations often deviate from standard phonetic rules.

Ahmadi (2019) [9] developed a rule-based Kurdish transliteration system, tackling challenges posed by multiple Kurdish dialects and script variations (e.g., Sorani and Kurmanji). The study highlights the importance of dialect-aware transliteration models, as a single set of rules cannot capture all phonetic variations across dialects.

Key takeaways from these papers:

- Rule-based approaches work well for structured phonetic systems but struggle with loanwords and dialectal differences.
- Hybrid models improve transliteration accuracy by balancing linguistic rules with data-driven learning.
- Proper nouns and dialectal variations require specialized handling to ensure correct transliteration.

These studies reinforce the idea that a one-size-fits-all transliteration model is insufficient, a combination of phonetic rules, statistical learning, and contextual awareness is necessary. The challenges faced in Punjabi, Hindi, and Kurdish transliteration closely parallel those in Bangla-to-English transliteration, especially regarding aspirated consonants, loanwords, and proper noun handling. This comparative analysis informs the need for adaptive, hybrid transliteration systems in Bangla as well.

3.8 Conclusion

The existing literature on Bangla transliteration, while limited, offers diverse methodologies ranging from historical perspectives to computational approaches. Studies have explored phonetic preservation, machine translation integration, digital spelling convergence, and autocorrection mechanisms. Additionally, transliteration research in other Indo-European languages provides comparative insights that can inform Bangla transliteration system development. Future research should focus on refining hybrid approaches, integrating deep learning techniques, and standardizing transliteration rules for widespread adoption.

Chapter 4

Methodology

4.1 Data

My data collection process was one of the most crucial part of the thesis. As collecting correct Bangla words in of itself is a challenge. I collected data from various sources, like online news portals, websites, social media chats etc. I have described every part below:

4.1.1 Initial Data Collection

Firstly, I collected 80,000 Bangla words from Kaggle [11]. These words were then transliterated from Bangla to English using both an automation process using IPA and manual correction by us. But, during the transliteration process, I found out that most of these words were ancient and not commonly used in recent Bangla. But my primary goal was to create a system to properly transliterate Bangla to English for modern Bangla which can be used in recent times. So, I stopped transliterating after converting 10,000 words and focused on collecting for colloquial Bangla words.

4.1.2 Modern Word Collection

There are many ways and sources to collect modern Bangla words. As I want the most recent and most common words I used two sources to collect data. These are described below:

Social Media Chats

To collect words from social media chats, a website [14] was developed that allowed users to upload their social media chat files. The site was designed to automatically extract Bangla words from the chat files. It than ranked them from most to least used with. Users were given the option to deselect any word they did not wish to share before submitting the data. After they had submitted their words, those were sent to my database. But, these were not enough for my research. Also, me and my team had to manually verify if those words were genuine Bangla words or there were any wrong words included as people tend to misspell words when texting.

Online News Portals

Lastly, I collected Bangla words from online Bangla news portals. I only collected words of the past year (2023). This provided us with a huge corpus of words that are currently in use. One of the main reason behind chosing this source is these words were already corrected by news portal authors. That gave us some validation not to go through the whole dataset.

4.1.3 Data Merge

I then combined all the words from both social media chats and online news portals as sown in Figure 4.1. These were more than 3 lakhs words combined. I took the top 100,000 most used words from this combined dataset, excluding any words that were part of the initial 10,000 words transliterated from the Kaggle dataset. Lastly, the initial 10,000 words and later collected 100,000 words were merged. This set of 110,000 words was used as my total dataset.

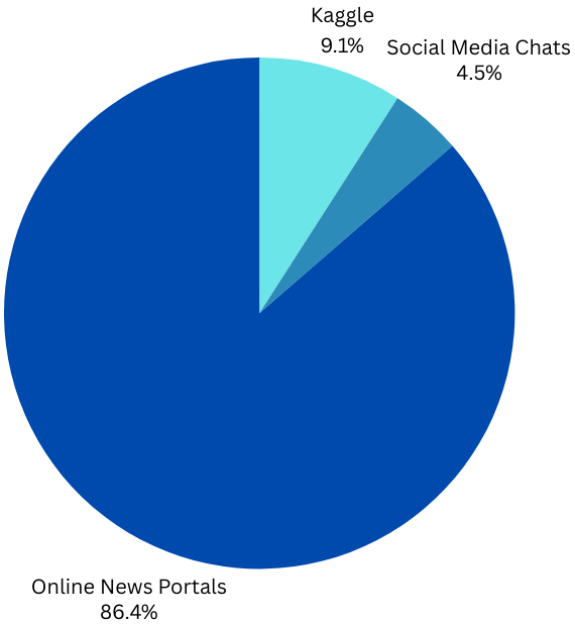


Figure 4.1: Collected Word Distribution

4.1.4 Ethical Considerations

My data collection process was designed with privacy and ethical considerations in mind. To collect chat data, I broke every sentence into words and ordered those words based on count in descending order. Also, users could deselect any word they did not wish to send. This method did not expose any confidential information of users. So, users' privacy and data integrity were kept intact. And, I took words from news portals that were open for public uses. So, there were no copyright issues.

This approach ensured a good mix of recent words with archaic words in my corpus of Bangla words.
In this paper, I have tried two different approaches to transliterating Bangla to English. Both are described below.

4.2 Approach Evolution

4.2.1 Alphabet-by-alphabet Transliteration Approach

Initially, I tried an alphabet-by-alphabet transliteration method. In this approach, individual Bangla characters were mapped directly to their English equivalents, such as 'ক' to 'k' and 'ঘ' to 'gh'. However, this method had significant limitations. Many Bangla words did not sound as they were spelled. For example, the word 'ধ্বনি' would be transliterated as 'dhboni', but its actual pronunciation is closer to 'dhoni'. This creates some issues to follow this approach.

To address this, I had to create various rules to handle many exceptions. For example, one rule was, 'জ' and 'য' should be transliterated as 'j', and silent 'য' at the end of words was omitted (e.g., 'স্বাচ্ছন্দ্য' was transliterated as 'shacchondo'). I had to add many rules and it was getting very complex and unintuitive.

4.2.2 Phonetic Approach

After reconsidering I changed my approach to a phonetics. This method prioritizes the pronunciation of Bangla words rather than their spelling, aiming to provide a more accurate and natural transliteration. By focusing on the sounds of the words, the phonetic approach offers a more intuitive and user-friendly solution.

In the phonetic approach, I first converted each of the 110,000 Bangla words to the International Phonetic Alphabet (IPA) using the IPA conversion tool provided by the Bangladeshi government [12]. The accuracy of the website seemed good enough as most of my transliteration were correct. I then used the International Phonetic Association's [13] chart and phonetic help from Vocabulary.com [15]to map IPA symbols to corresponding English alphabets.

Despite this method's advantages, I encountered several issues. For example, the Bangla word 'টমেটো' was converted to 'tometo' based on its phonetic pronunciation. However, since 'টমেটো' is an English word (i.e., 'tomato'), spelling it phonetically could create confusion. To address such issues, I implemented a rule to retain the original English spelling for English words, ensuring clarity and avoiding misinterpretation. To verify these issues I hired a team of undergraduate students to go through every word manually and find if any word's trnaliteration has any issues or not.

Annotator Selection and Quality Control

To ensure the accuracy of transliterations, I conducted a screening process for data annotators using a Google Form-based evaluation. The form contained 10 Bangla words, and participants were required to provide their correct transliterations in English based on phonetic consistency. Annotators who provided correct responses were selected, ensuring that only those with a proper understanding of transliteration rules contributed to the dataset.

The following Bangla words were used in the screening process:

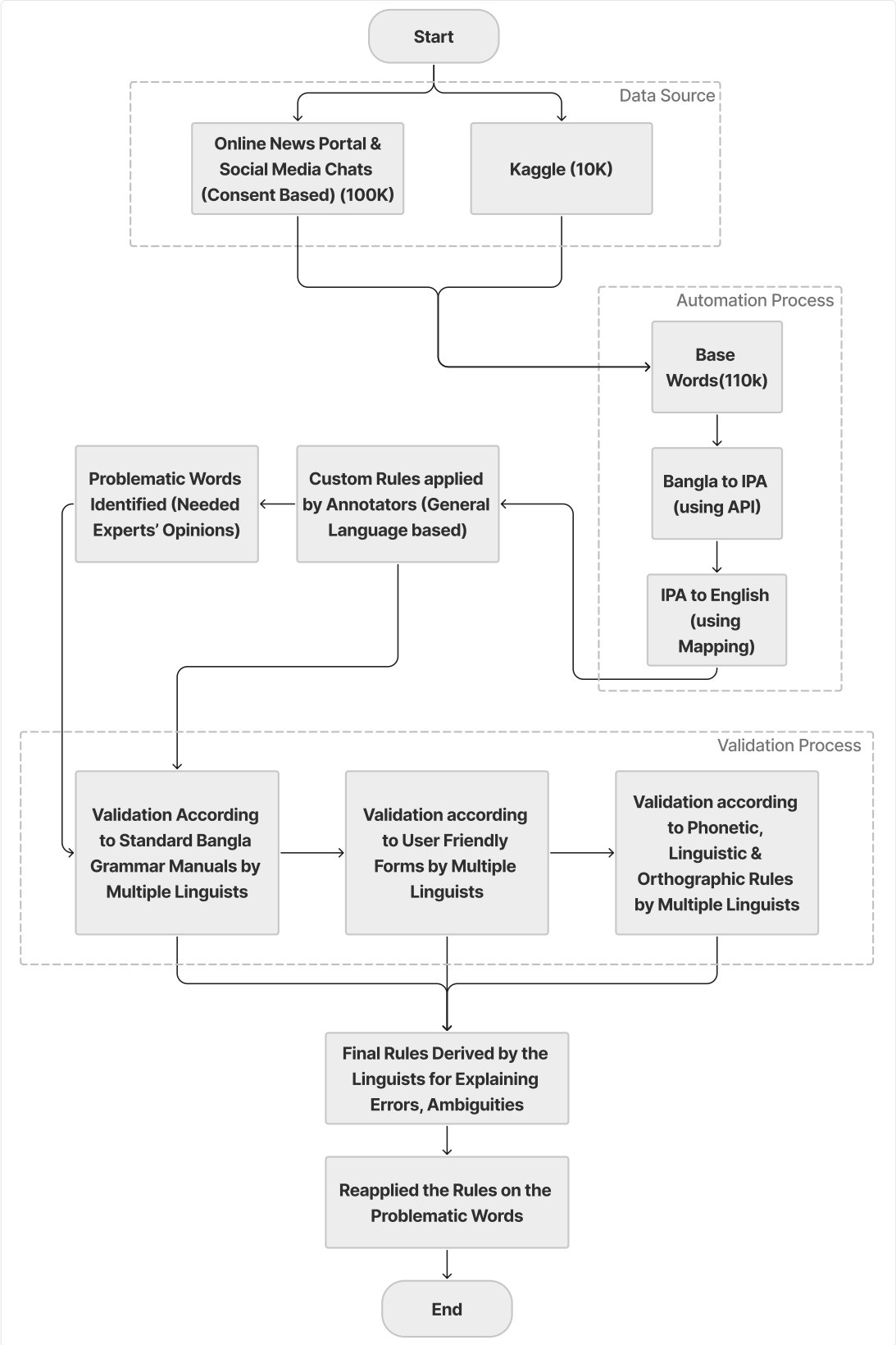


Figure 4.2: Flowchart of the Transliteration Process

- হ্যাঁ → hae
- অভিযোজনক্ষম → obhijojonkhomo
- কল্পনাপ্রবণ → kolponaprobon
- বাদ্যযন্ত্র → baddojontro
- সত্যান্বেষী → shottanneshi
- শিখরচূড়া → shikhorchura
- দৃষ্টিভঙ্গি → drishtibhongi
- মাঝি → majhi
- কৈফিয়ত → koiphiyot
- অঙ্গভঙ্গি → ongobhongi

These words were chosen to assess different phonetic patterns, including consonant clusters, aspirated sounds, nasalization, and vowel variations, ensuring that annotators could handle diverse transliteration challenges.

Data Annotator Manual

After selecting my data annotators, their main task was to identify words where pronunciation and transliteration did not align. Initially, me and my linguists went through 10000 words and created basic guidelines for my data annotators to follow and find out issues with the transliterations. These are given below:

- **Phonetic Consistency:** Ensuring that transliteration follows a standardized phonetic approach rather than personal preferences.
- **Handling of Aspirated and Unaspirated Sounds:** Differentiating between sounds like 'ত' (t) vs. 'থ' (th), ensuring correct representation.
- **Retention of Proper Noun Capitalization:** Maintaining capitalization for names, places, and established terms.
- **Dealing with Loanwords:** Identifying words borrowed from other languages and ensuring their transliteration follows established norms.
- **Consonant Clusters:** Correctly representing combinations like 'স্ক', 'গণ', 'স্ফ' without adding or omitting letters.
- **Nasalization and Vowel Length:** Recognizing nasalized sounds and long vowels to ensure phonetic accuracy.
- **Avoiding Overcomplication:** Keeping transliterations intuitive and readable instead of adding unnecessary complexity.
- **Handling of Compound Words:** Ensuring that compound or conjunct words (e.g., বাচ্চা, শিল্প) follow consistent rules.

- **Intuitive Findings:** If an annotator encountered a confusing or ambiguous word, they were required to consult the research team for clarification. This ensured that no subjective or inconsistent transliterations were included in the dataset.

If any transliteration had any issues with these rules my data annotators marked those and I consulted with my linguist team and made necessary rules to tackle with the issues.

Validation and Rule Generation Process

After identifying problematic words through my data annotation process, my linguists developed a systematic set of transliteration rules. These rules were formulated based on established linguistic principles while ensuring usability and accuracy. The process involved the following key steps:

i. **Validation of Problematic Words**

- Linguists thoroughly reviewed the identified words to confirm inconsistencies in pronunciation and transliteration.
- Words were cross-referenced with standard Bangla grammar manuals to ensure alignment with linguistic norms.
- Common user errors were analyzed to understand recurring mistakes in transliteration.

ii. **Consideration of Phonetic and Linguistic Principles**

- Each rule was designed to preserve phonemic accuracy, ensuring that the transliteration reflects the actual pronunciation of the word.
- Special attention was given to aspirated vs. unaspirated consonants, nasalization, and long vs. short vowels (e.g., $\rightarrow$ Dhaka, $\rightarrow$ baba).
- Loanwords and proper nouns were examined to maintain their original pronunciation while following Bangla phonetic patterns.

iii. **Balancing User-Friendly Forms and Orthographic Rules**

- Rules were designed to be intuitive and easy to apply, avoiding excessive complexity that might confuse users.
- Linguists ensured that conjunct consonants (e.g., বাচ্চা $\rightarrow$ baccha, স্কুল $\rightarrow$ school) followed natural pronunciation patterns.
- Word-final consonants were handled carefully to avoid distortions in pronunciation (e.g., পত্র $\rightarrow$ potro, not patro).

iv. **Conflict Resolution Between Linguists**

- When differing opinions on transliteration rules arose, they were documented and analyzed based on phonetic consistency, linguistic precedent, and practical usability.

- Additionally, 53 interviews were conducted from different universities linguist department to resolve conflicts.

v. **Iterative Refinement and Testing**

- The proposed rules were tested on a diverse set of words to ensure consistency and accuracy.
- Multiple iterations were conducted, refining rules where necessary to handle edge cases effectively.
- Final validation was performed by linguists and experienced transliteration experts before implementation.

This structured approach ensured that my transliteration rules were both linguistically sound and practically applicable, balancing grammatical accuracy with real-world usability. This rules generation process is shown in Figure 4.2 and the generated rules are shown in Table 5.1.

4.2.3 Word Conversion

Using the phonetic approach and by applying the rules generated to the problematic words, we transliterated all 110,000 words from Bangla to English. This extensive process ensured that each word followed the phonetic principles we established, providing consistency and accuracy in the transliteration. The resulting dataset reflects contemporary usage, offering a practical and intuitive solution for writing Bangla using English alphabets.

4.3 Algorithms for Generating Correct Transliteration Suggestions

To improve the accuracy of transliteration correction, I developed two distinct algorithms. The first approach leveraged traditional phonetic similarity measures, while the second introduced a novel phonetic representation tailored for Bangla.

4.3.1 Initial Approach: Edit Distance and Phonetic Similarity (Limited Scalability)

The first algorithm aimed to generate accurate transliteration suggestions by comparing the given word with entries in a dictionary of correctly transliterated words. It followed these steps as shown in Figure 4.3:

i. **Preliminary Filtering:** The algorithm first checked whether the length difference between the given word and any dictionary word was less than 3. This step aimed to reduce the number of comparisons.

ii. **Phonetic Similarity Computation:** It then applied Soundex, Edit Distance, and a Custom Closeness Metric (as shown in Figure 4.4) to account for common Bangla-to-English spelling variations (e.g., v $\leftrightarrow$ bh, y $\leftrightarrow$ i, a $\leftrightarrow$ e).

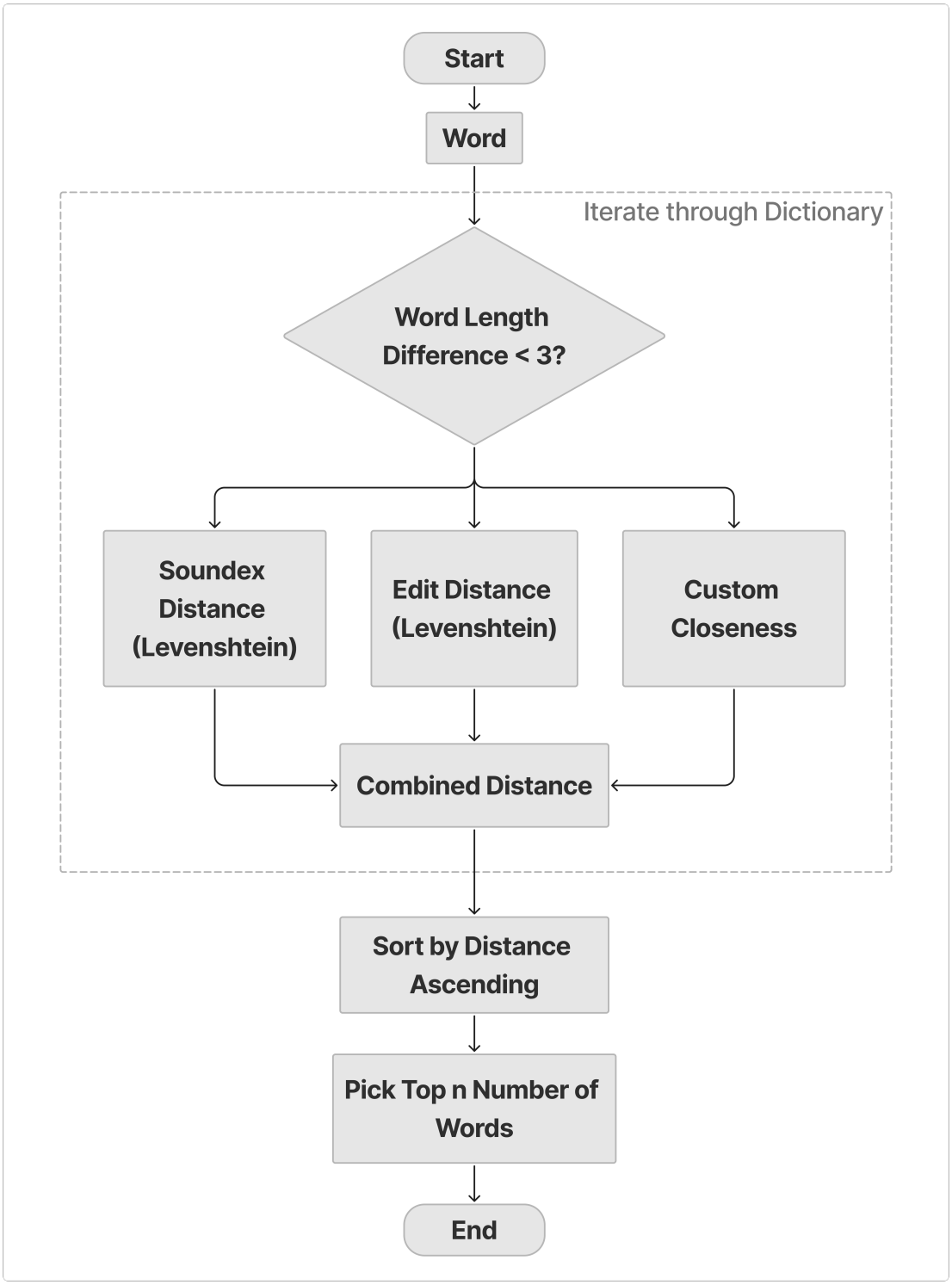


Figure 4.3: Algorithm 1 Flowchart

iii. **Ranking Candidates:** The computed distances were combined into a single score, and all dictionary words were sorted in ascending order based on this score.

iv. **Suggestions:** The top n = 5 words were returned as suggestions.

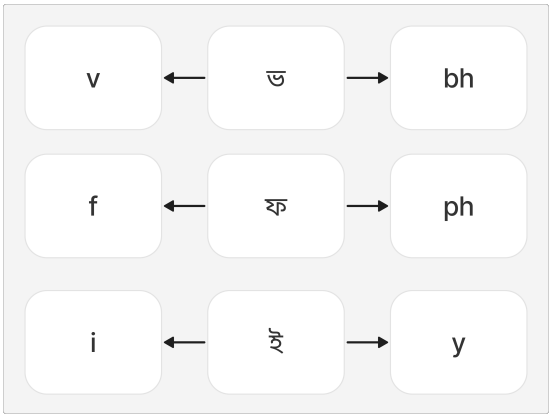


Figure 4.4: Custom Closeness of Certain Letters

Limitations

While effective on a small dataset (500 words), this approach performed poorly around 65%, when scaled to 110k words due to two key issues:

- Soundex and similar phonetic algorithms are designed for English and fail to capture Bangla phonetic nuances.
- The preliminary length filtering was insufficient to minimize exhaustive dictionary comparisons.

4.3.2 Improved Approach: Custom Phonetic Representation (Scalable Solution)

To address the shortcomings of the first approach, I designed a Bangla-specific phonetic representation that better captures spelling variations. This approach is shown on Figure 4.5 This method introduced two core transformations for each word:

- **Syllabic Phonetic Approximation (SPA):** A transformation that approximates the phonetic structure of a word based on syllables.
- **Consonant Skeleton (CS):** A simplified representation that retains only the consonant structure of the word.

The improved algorithm follows these steps:

i. **Initial Filtering:** Compute the edit distance and word length difference between the given word and dictionary words. If both values are greater than 3, the word is discarded to reduce search space.

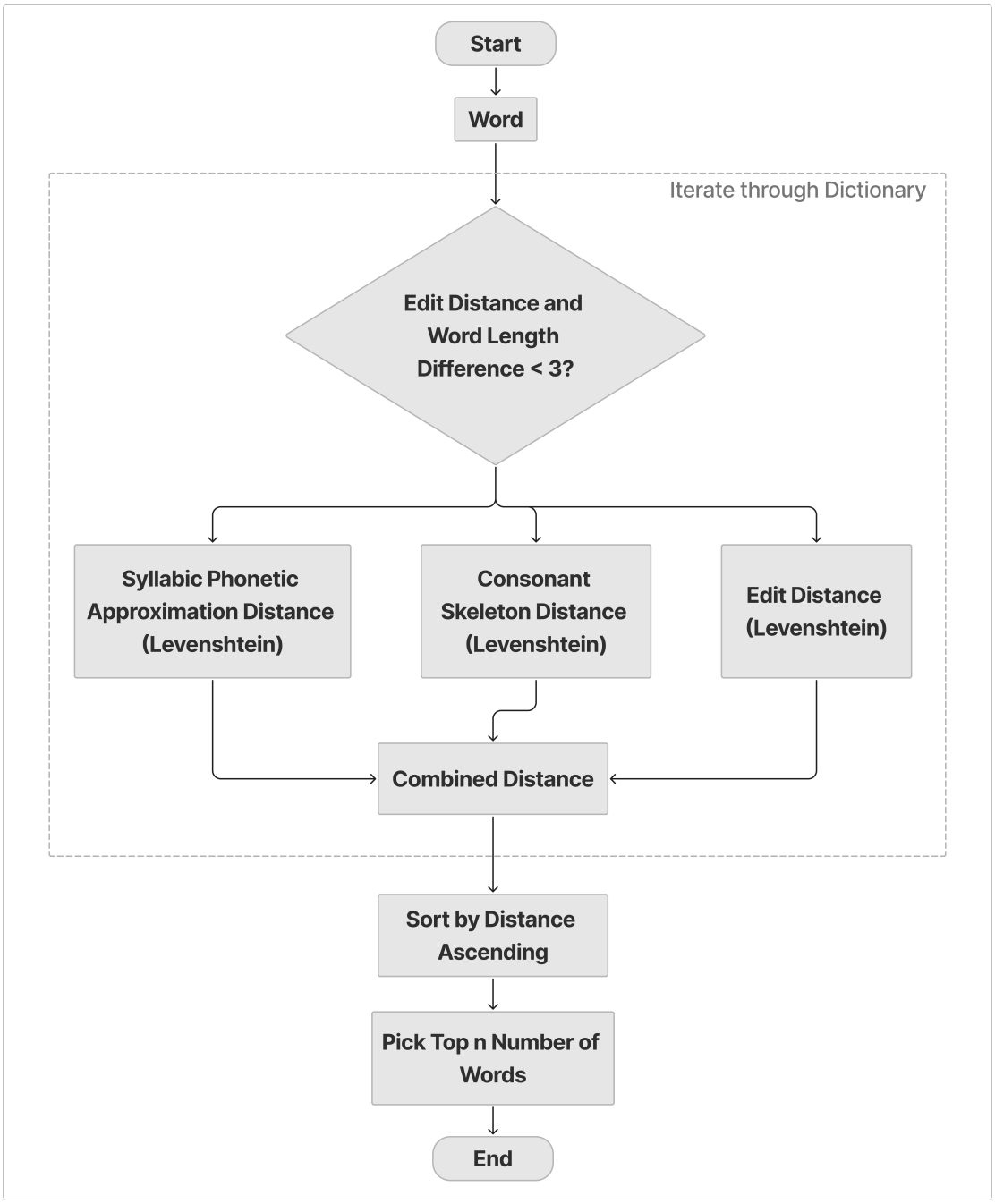


Figure 4.5: Algorithm 2 Flowchart

ii. **Phonetic Transformation:** Convert both the given word and each dictionary word into their SPA and CS forms.

iii. **Similarity Computation:** Compute three distances (elaborately discussed in Background section):

- SPA Distance: Measures phonetic similarity.
- CS Distance: Captures structural similarities.
- Edit Distance: Accounts for character-level modifications.

iv. **Combination & Ranking:** These distances are combined into a weighted score, and dictionary words are sorted in ascending order.

v. **Returning Suggestions:** The top n = 5 words are provided as likely corrections.

This approach significantly outperforms the initial algorithm, delivering highly relevant suggestions even with 110k words while maintaining computational efficiency.

4.3.3 Misspelled Word Generation

To evaluate the effectiveness of my algorithm, I utilized a set of 500 words and deliberately created 1,378 misspelled variations from those words. The misspelled words were generated based on common typing errors observed in social media conversations and frequent spelling mistakes made by individuals. These errors include typographical mistakes such as 'a' instead of 'e', 'y' instead of 'i', 'm' instead of 'n', 'j' instead of 'g', and 'o' instead of 'u', and other common slip-ups like 'b' instead of 'v', 'r' instead of 't', and 'w' instead of 'q'. These patterns were derived from my knowledge of language use on platforms like Taklami and common linguistic tendencies, ensuring a diverse and realistic set of misspellings for testing the algorithm's accuracy and suggestion capabilities.

Chapter 5

Experimental Results

5.1 Rules

Following the initial phonetic conversion and consultation with linguists, a comprehensive set of rules was established to address specific challenges and improve the transliteration process. These rules were developed to ensure that the transliteration process is both accurate and intuitive, addressing specific linguistic and practical challenges.

Table 5.1: Transliteration Rules

Sl.	Rule	Example	Linguistic Validation
1	Established names and terms should remain unchanged.	ইদ→Eid, পদ্মা→Padma, ফরিদপুর→Faridpur, চট্টগ্রাম→Chattogram	Established names and spellings should be according to the conventional rules as established.
2	Proper noun spellings should not be changed.	আফ্রিদি→Afridi, প্রত্যয়→Prattoy	Proper nouns (Names) also do not change linguistically. For example, Russian Names - 'Vladimir Putin' or Arabic Names- 'Muhammad'
3	The spelling of English words should not be changed.	টমেটো →tomato (not 'tometo'), ব্যাংক→bank (not 'baengk')	As we are transliterating Bangla to English, thus the English words should be kept intact to overcome ambiguity.

Continued on next page

Sl.	Rule	Example	Linguistic Validation
4	When a word comprises two parts (mostly compound words), with one being an English word and the other being Bangla, we should spell the English part in its original language and the Bangla part in Bangla	উইকেটরক্ষক →wicketrokkhok, মোটরগাড়ি→motorgari	As we are transliterating Bangla to English, thus the English words should be kept intact to overcome ambiguity.
5	If any word ends with a Bangla determiner, there should be a space between the main word and the determiner	অ্যাপটি→app ti, ব্যাগটি→bag ti, কাজটা→kaj ta	Bangla determiners can create ambiguity for the transliterated words (Bangla/English). E.g. সাতটি→shat ti (not 'shatti'), গ্রাফিতিটি→grafitti ti (not 'graffititi'). Here, not separaitng the main word from the determiner creates confusion. That is why, a space before the determiner is crucial.
6	When transliterating Bangla letters that are represented by two English letters, and the second letter is 'h' (e.g., 'চ'→'ch'), in conjunct forms of those letters, the first 'h' is dropped.	আচ্ছা→accha (not 'achcha'), বাচ্চা→baccha (not 'bachcha'), বিশ্বাস →bisshash (not 'bishshash'), অবিবক্ষিত→obibokkhito (not 'obibokhkhito')	In Bangla, some words have consecutive sounds that are represented in English using two letters. This rule helps simplify the spelling and maintain pronunciation clarity. (e.g. In some conjunct words, the letter 'h' comes twice, and the first of 'h' should be dropped. So, in the Bangla combined letters/ phonteic blendings, 'চ্ছ'→'cch', 'শ্ব'→'ssh', 'ক্ষ'→'kkh'. These are convenient with the older IPA conventions/usage of the 1900s used by M. Abdul Hi, Munir Chowdhury.
7	If any English word ends with Bangla suffix, there should be a space between them	স্পেসের →space er, ড্যান্সের →dance er	When some Bangla suffixes are added to English words, it can alter the meaning. E.g., 'ড্যান্সের' might be transliterated as 'dancer' which has a different meaning in English. Therefore, a space should be added before the suffix.

Continued on next page

Sl.	Rule	Example	Linguistic Validation
8	If ‘য়’ appears in the middle of a word, it should be transliterated as ‘wa’ and if it appears at the end of a word, it should be ‘ay’	বানোয়াট→banowat, খুলনায় →Khulnay	In the middle of a word ‘য়’ sounds like ‘ওয়া’ (wa) so, it should be transliterated as ‘w’ but at the end of a word ‘য়’ sounds like ‘য়’ (y) so, it should be transliterated as ‘y’
9	All acronyms should be capitalized	বিবিসি→BBC, সিইউ→CU, ঢাবি (ঢাকা বিশ্ববিদ্যালয়) →DHABI	As there are established rules for English capitalization, it is being applied to transliteration for consistency and easier understanding of Bangla acronyms.
10	If a name or location ends with a Bangla suffix, a space should be inserted between the main word and the suffix, according to Linguistic rules of Assimilation or Dissimilation	প্রত্যয়ের→Prattoy er, ঢাকার→Dhakar (not ‘Dhaka er’), সিলেটের→Sylhet er	All suffixes should be separate except for Name or Location entities, only the phonological assimilation rules should be applied for these. (1) If two consecutive vowels and vowel-like sounds (semi-vowels) come together, the affix will assimilate. E.g. ঢাকার→Dhakar (Dhaka er). (2) If a consonant or consonant-like sound (semi-vowels) appears at the word ending/ starting of the affix, the affix would need a space.
11	English transliteration for ‘ফ’ should be ‘ph’ and ‘ভ’ should be ‘bh’	ফুল →phul, ভুল →bhul	In Bangla, the letters ফ and ভ are labial plosives, not matching the IPA sounds for ‘f’ and ‘v’ respectively. ফ is a voiceless plosive, like the English ‘p’ but breathier, and is transliterated as ‘ph’. ভ is a voiced plosive with vocal cord vibration, similar to ‘b’ with breathiness, transliterated as ‘bh’.
12	For any non-Bangla words ‘f’ sounding alphabet should be ‘f’ and ‘v’ sounding alphabet should be ‘v’	মওকুফ→mowkuf	As mentioned earlier, the sound ‘ফ’→‘ph’ and ‘ভ’→‘bh’ would be transliterated in Bangla according to Bangla phonologists. Thus, the foreign words should be separated with the different letters ‘f’ and ‘v’ according to the IPA pronunciations to distinguish the spellings of the foreign words.

Continued on next page

Sl.	Rule	Example	Linguistic Validation
13	English transliteration for ‘চ’ and ‘ছ’ should be ‘ch’	চাও→chao, ছায়া→chaya	In Bangla, চ is a voiceless palatal plosive and ছ is a voiceless aspirated palatal plosive. Since both sounds are formed in a similar region of the mouth (palatal), they both involve a ‘ch’ sound in English and people mostly use ‘ch’ for both of these letters.
14	English transliteration for ‘জ’ and ‘য’ should be ‘j’	জানা→jana, যাবে→jabe	In Bangla, জ and য sound similar and are phonetically distinguished as ‘j’ for জ and ‘z’ for য. However, since both are commonly pronounced as ‘j’ by Bengali speakers, they are kept that way.
15	English transliteration for ‘ট’, ‘ত’ and ‘ৎ’ should be ‘t’	টঙ→tong, তুমি→tumi, শরৎ→shorot	In Bangla, ট is a retroflex sound and ত is a dental sound, as there are no alphabet equivalent for ত English ‘t’ should be used. Moreover, both ত and ৎ sounds are very close phonetically, so using ‘t’ for both helps maintain consistency in transliteration.
16	English transliteration for ‘দ’ and ‘ড’ should be ‘d’	দাও→dao, ডাব→dab	In Bangla, both দ and ড are dental sounds. Although, these sounds different in Bangla there are no alphabet equivalent in English. So, ‘d’ should be used in transliteration.
17	English transliteration for ‘র’, ‘ড়’ and ‘ঢ়’ should be ‘r’	রঙ→rong, পড়া→pora, আষাঢ়→ashar	In Bangla, the rhotic sounds র, ড়, and ঢ় differ phonetically but sound similar to a non-specialist. র is a soft retroflex or alveolar ‘r’, ড় is a harder retroflex ‘r’, and ঢ় is an aspirated retroflex ‘r’. To simplify, all should be transliterated as ‘r’.

Continued on next page

Sl.	Rule	Example	Linguistic Validation
18	Capitalization of first letter should be preserved in transliteration for proper nouns, foreign terms, and established names. Additionally, English capitalization rules should be applied to Bangla transliterated words	হাবিব→Habib, ঢাকা→Dhaka, নর্থ→North, উত্তর→Uttor	Capitalization makes it easy to write and makes the transliteration more accessible to general users by following conventional English rules. Linguistically, capitalization is done for the disambiguation of any complexities.
19	For nasal sounds, when the doubling occurs, they should be combined into a single letter	অনঙ্গমোহন→onongomohon (not 'ononggomohon'), অঙ্গ→ongo (not 'onggo'), ভঙ্গ→bhongo (not 'bhonggo')	Nasalization in Bangla has a very common pattern, where nasal 'n' combines with other sounds like velar 'g'. But, when transliterating according to IPA conventions, the phonological combination of the pattern 'ngg' often comes. In this case, only keeping the 'ng' by dropping the extra 'g' in transliteration is enough for Bangla pronunciations. (e.g., 'ngg' becomes 'ng')
20	In Bangla numbers, every number's transliteration should be different	সাত→shat, ষাট→shaat	Ambiguity in numbers is not acceptable. In the example, for সাত 'a' is a short vowel sound and for ষাট 'aa' sound indicating a longer vowel.

5.1.1 Ambiguities in Transliteration

While transliterating Bangla into English, some characters share the same transliterated form, leading to potential ambiguity. For instance, চ (ch) and ছ (chh), or স (sh) and ষ (sh), are represented by the same English letter combinations. These transliterations are necessary for user-friendliness and to keep the system simple and consistent. Although this could create ambiguity when these words are seen out of context, the meaning and pronunciation are often clarified by the surrounding text, as the context plays an important role in distinguishing between them. Despite these possible ambiguities, using these consistent transliterations ensures a smoother reading experience for general users and avoids the need for complex and cumbersome diacritics.

5.2 Algorithm Results

To evaluate the performance of my transliteration correction algorithms, I tested them on datasets containing both correctly transliterated and misspelled words. The results of both approaches are summarized below.

5.2.1 Approach 1: Small Dictionary-Based Matching

In the first approach, I tested my algorithm on a small dictionary containing 500 words. The key results were:

- **Dictionary Size:** 500
- **Correct Words Tested:** 500
- **Misspelled Words Tested:** 1378
- **Accuracy:** 95%

This approach performed well on a small dataset, as it efficiently filtered out incorrect transliterations and provided accurate suggestions. However, when tested on whole datatset as dictionary its accuracy rate dropped down to around 67%.

5.2.2 Approach 2: Phonetic-Based Matching with Large Dictionary

The second approach leveraged phonetic representations and consonant skeleton matching to improve scalability. I expanded the dictionary to 87,594 words after removing all the peoples names from 110,000 words, significantly increasing the search space. The results were:

- **Dictionary Size:** 87,594
- **Correct Words Tested:** 500
- **Misspelled Words Tested:** 1378
- **Accuracy:** 88.9%

Although the accuracy slightly dropped compared to Approach 1, this method successfully handled a much larger vocabulary. The reduction in accuracy can be attributed to the increased complexity of Bangla phonetics and the presence of multiple possible transliterations for a single word. However, this approach proved more viable for real-world applications, where a vast vocabulary is necessary.

5.3 Comparison and Observations

- Approach 1 performed exceptionally well on a small dataset but failed to scale due to Soundex and Custom Closeness were not working well.
- Approach 2, despite a slight accuracy drop, demonstrated practical scalability and robustness in handling large vocabulary sets.
- The phonetic-based representation in Approach 2 proved crucial in distinguishing common transliteration errors, making it more effective for large-scale applications.

These findings suggest that an optimized hybrid approach could further enhance transliteration correction, combining the precision of small dictionary filtering with the scalability of phonetic modeling.

5.4 User Interface Demonstration of Suggestions

To showcase the effectiveness of my transliteration correction algorithm, I developed a web-based application [16] where users can input incorrectly transliterated words and receive multiple correction suggestions. This platform serves as an interactive demonstration of the system's capabilities, allowing real-time testing and validation. In the Figure 5.1 a sentence's every word's suggestion is shown in transliterated form and in the Figure 5.2 a sentence's every word's suggestion is shown in Bangla.

Figure 5.1: Suggestion as Transliterated Word

This web application not only validates the algorithm's performance but also serves as a practical tool for transliteration assistance. By integrating this system into a user-friendly interface, I ensured that the algorithm's capabilities can be easily accessed and tested by a wider audience.

Figure 5.2: Suggestion as Bangla Word

Chapter 6

Conclusion

6.1 Challenges and Limitations

The process of transliterating Bangla to English had several challenges and limitations. These are:

- **Collecting Modern Words:** Collecting modern words was challenging due to privacy issues. Usually, people do not want to share their chat data, which is essential for collecting up-to-date vocabulary.
- **Verification of Automated Transliterated Words:** Ensuring the accuracy of automated transliterated words requires multiple human verifications, making it a labor-intensive task. Each word must be meticulously checked by several individuals to ensure correctness.
- **Developing Generic Rules for Exceptions:** Creating rules to handle problems after reviewing all words is complex. Linguists had to consider various issues to make these linguistically accurate and user-friendly.
- **Consensus Among Linguists:** Coming to a common ground for linguists was another challenge. They had different opinions on the correct transliteration of certain words, which complicated the whole process.

Despite these challenges, my approach has made significant progress in transliterating Bangla to English, setting a foundation for future work in this area.

6.2 Future Direction

To enhance transliteration accuracy and usability, several improvements can be explored:

- **Context-Based Current Word Prediction:** Instead of treating words in isolation, the system can use surrounding words to refine transliterations, reducing ambiguity and improving accuracy. As shown in Figure 6.1, when 'k' is written the system is suggesting 'khai'.

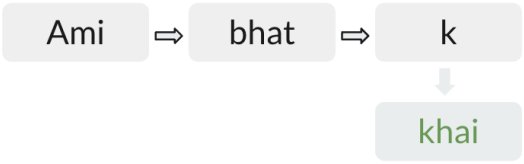


Figure 6.1: Context-Based Current Word Prediction

- **Full Sentence Correction:** Beyond individual words, a sentence-level correction model can detect and fix errors in grammar, spacing, and phonetics, ensuring smoother and more natural output. As shown in Figure 6.2, the whole sentence is corrected.

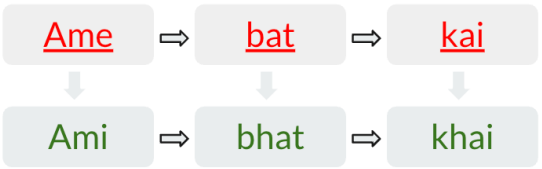


Figure 6.2: Full Sentence Correction

- **Next Word & Full Sentence Prediction:** Predicting the next word based on context can speed up typing and reduce errors. Extending this to full sentence prediction would allow for faster and more fluent transliteration. As shown in Figure 6.3, the next word 'khai' is predicted.

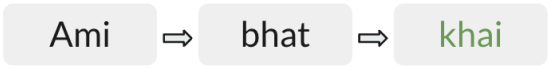


Figure 6.3: Next Word & Full Sentence Prediction

These advancements will make the system more context-aware, efficient, and user-friendly for real-world applications.

6.3 Summary

In conclusion, this study presents a novel approach to Bangla-to-English transliteration, combining human validation with automated techniques. By successfully transliterating 110,000 Bangla words and developing a highly effective correction suggestion algorithm, this research contributes significantly to improving transliteration accuracy. The future integration of machine learning will enable the system to analyze contextual information, further enhancing the precision of word suggestions. With the implementation of this method, the occurrence of misspellings can be minimized, ensuring higher quality and reliability in Bangla transliteration systems.

Bibliography

[1] D. Kurzon, "Romanisation of bengali and other indian scripts," *Journal of the Royal Asiatic Society*, vol. 20, no. 1, pp. 61–74, 2010.

[2] K. Deep and V. Goyal, "Development of a punjabi to english transliteration system," *International Journal of Computer Science and Communication*, vol. 2, no. 2, pp. 521–526, 2011.

[3] K. M. A. Salam, S. Yamada, and T. Nishino, "How to translate unknown words for english to bangla machine translation using transliteration," *Journal of computers*, vol. 8, no. 5, pp. 1167–1174, 2013.

[4] V. Kaur, A. k. Sarao, and J. Singh, "Hybrid approach for hindi to english transliteration system for proper nouns," *International Journal of Computer Science and Information Technologies*, vol. 5, no. 5, pp. 6361–6366, 2014.

[5] S. Khan, "Convergence in spelling, and spell-checker for romanized bangla in computers and mobile phones," in *2014 International Conference on Informatics, Electronics & Vision (ICIEV)*, IEEE, 2014, pp. 1–5.

[6] T. Dasgupta, M. Sinha, and A. Basu, "Resource creation and development of an english-bangla back transliteration system," *International Journal of Knowledge-based and Intelligent Engineering Systems*, vol. 19, no. 1, pp. 35–46, 2015.

[7] D. Sen and K. D. Garg, "A review on bengali to english machine transliteration system," *International Journal of Software and Web Sciences (IJSWS)*, pp. 60–64, 2015.

[8] M. F. Akan, "Transliteration and translation from bangla into english: A problem solving approach," *British Journal of English Linguistics*, vol. 6, no. 6, pp. 1–21, 2018.

[9] S. Ahmadi, "A rule-based kurdish text transliteration system," *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, vol. 18, no. 2, Jan. 2019, ISSN: 2375-4699. DOI: 10.1145/3278623. [Online]. Available: https://doi.org/10.1145/3278623.

[10] M. M. Hossain, M. F. Labib, A. S. Rifat, A. K. Das, and M. Mukta, "Auto-correction of english to bengali transliteration system using levenshtein distance," in *2019 7th International Conference on Smart Computing & Communications (ICSCC)*, IEEE, 2019, pp. 1–5.

[11] M. M. HASAN, *80k bangla words list (bangla dictionary)*, 2022. [Online]. Available: https://www.kaggle.com/datasets/mahadivai/spelling-checker-v1.

[12] I. D. EBLICT Project BCC. "ধ্বনি converter engine," Accessed: Mar. 12, 2024. [Online]. Available: https://ipa.bangla.gov.bd/.

[13] I. P. Association. “The international phonetic alphabet (ipa) - full chart,” Accessed: Mar. 12, 2024. [Online]. Available: https://www.internationalphoneticassociation.org/content/full-ipa-chart.

[14] A. M. Joaa. “Word collector,” Accessed: Apr. 15, 2024. [Online]. Available: https://word-collector-e1ad6.web.app/#/homepage.

[15] Vocabulary.com. “Ipa pronunciation guide,” Accessed: Mar. 12, 2024. [Online]. Available: https://www.vocabulary.com/resources/ipa-pronunciation/.

[16] P. Majumder. “Word suggestions,” Accessed: Jan. 15, 2025. [Online]. Available: https://infai.xyz/td-search.

[17] M. F. Akan, “Transliteration of bangla words,” *Bangla Academy Journal*,