Shahjalal University of Science & Technology

Software Engineering

Institute of Information and Communication Technology

Course Code: SWE450

DUAL-IPA: Bengali Transcription Modeling Dataset

Submitted By

Navid Hasan
Reg. No.: 2018831026
Software Engineering
IICT, SUST

Sourav Ahmed
Reg. No.: 2018831068
Software Engineering
IICT, SUST

Supervisor

Sayma Sultana Chowdhury
Assistant Professor
Institute of Information and Communication Technology
Shahjalal University of Science and Technology

February 11, 2024

DUAL-IPA: Bengali Transcription Modeling Dataset

A Thesis submitted to the Institute of Information and Communication Technology, Shahjalal University of Science and Technology, in partial fulfillment of the requirements for the degree of B.Sc.(Eng.) in Software Engineering.

Submitted By

Navid Hasan
Reg. No.: 2018831026
Software Engineering
IICT, SUST

Sourav Ahmed
Reg. No.: 2018831068
Software Engineering
IICT, SUST

Supervisor

Sayma Sultana Chowdhury
Assistant Professor
Institute of Information and Communication Technology
Shahjalal University of Science and Technology

February 11, 2024

Recommendation of the Thesis Supervisor

To Whom It May Concern,

This letter is to certify that, the thesis entitled **DUAL-IPA: Bengali Transcription Modeling Dataset** undertaken by the students **Navid Hasan** and **Sourav Ahmed** is under my supervision. I, hereby, agree that the thesis can be submitted for examination.

Sayma Sultana Chowdhury
Assistant Professor
Institute of Information and Communication Technology
Shahjalal University of Science & Technology, Sylhet
Date: February 11, 2024

Approval of the Thesis

Students Name: Navid Hasan, Sourav Ahmed
Thesis Title: DUAL-IPA: Bengali Transcription Modeling Dataset

This is to certify that the above-mentioned thesis, submitted by the students named *above* in February 2024 as part of the requirements of the course **SWE 450**, is being approved by the Department of Software Engineering, Institute of Information and Communication Technology as a partial fulfillment of the **B.Sc.(Eng.)** degrees of the above students.

Director of IICT	**Chairman,**	**Supervisor**
Prof Dr. M. Jahirul Islam	**Exam. Committee**	Sayma Sultana Chowdhury
PhD, PEng	Prof. Dr. M. Jahirul Islam	Assistant Professor
	PhD, PEng	

Acknowledgement

We begin by expressing our heartfelt gratitude to the Almighty Allah for granting us the strength and composure to successfully complete our research.

We extend our sincere appreciation to the Department of Software Engineering at IICT, SUST, for their invaluable support and encouragement throughout our research journey. Furthermore, we owe a debt of gratitude to our supervisor, **Ms. Sayma Sultana Chowdhury**, Assistant Professor in the Department of Software Engineering at IICT, SUST, for her exceptional guidance, insightful advice, and unwavering support. Her contribution has been instrumental in shaping the course of our research.

We are also deeply indebted to the Bengali.AI community, particularly **Asif Shahriyar Sushmit** and **Md. Rezuwan Hassan** coordinators from **Bengali.AI**, for their remarkable assistance and unwavering support in our work.

Lastly, we would like to acknowledge the numerous individuals whose kind gestures and encouragement inspired and motivated us throughout this research. Their unwavering support and encouragement were truly appreciated

Abstract

This thesis presents a significant contribution to the field of Natural Language Processing (NLP) by introducing a novel and comprehensive Bengali dataset comprising 150,000 sentences transcribed in the International Phonetic Alphabet (IPA). Bengali, characterized by its intricate phonological features, stands as a promising subject for advanced NLP research. The meticulously curated dataset aims to capture the subtle phonetic nuances of Bengali, serving as a valuable resource for training and evaluating NLP models. The abstract outlines a detailed statistical analysis, annotation methodologies, and explores potential applications within the NLP domain, emphasizing the dataset's pivotal role in advancing the understanding of the Bengali language.

The discussion delves into the dataset's significance, pushing the boundaries of Bengali language comprehension and broader NLP research. Furthermore, the abstract highlights the dataset's versatility by showcasing its potential in enhancing various NLP tasks such as speech recognition, language modeling, sentiment analysis, and named entity recognition. This research not only contributes to the existing body of knowledge but also provides a foundation for future developments in Bengali NLP research, underscoring the importance of the dataset in fostering advancements in language technology.

Keywords: IPA, Bengali, Linguistics

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Introduction

Motivated by the captivating linguistic tapestry and ubiquitous utilization of the Bengali language, with a specific focus on Bangla as the official language of Bangladesh, this research endeavors to address a discernible lacuna in the realm of Natural Language Processing (NLP) resources tailored for Bengali. Against the backdrop of a staggering population tallying 272.7 million individuals in Bangladesh and a diaspora of diverse Bengali communities dispersed globally, the exigency for a comprehensive and nuanced linguistic dataset becomes palpable. Notwithstanding the nuanced morphological variations discernible among dialects spoken in distinct regions, the salient divergences in sounds and phonology emerge as noteworthy.

In response to this multifaceted linguistic milieu, this thesis sets forth a pioneering model with the explicit objective of transmuting written expressions of Standard Bangla into the International Phonetic Alphabet (IPA), an established system for the phonetic and phonemic transcription of spoken languages. The impetus for this endeavor derives from a conspicuous dearth of extensive datasets in Bengali IPA, thus motivating a robust initiative to bridge this lacuna. This ambitious undertaking encompasses the creation of a groundbreaking dataset, an expansive corpus comprising no less than 70,000 words. The meticulous annotation process enlisted the expertise of three native Bangla speakers, strategically chosen for their combined acumen in linguistics and engineering.

The protocols governing the conversion of Bengali to IPA underwent rigorous development, orchestrated by a seasoned team possessing specialized knowledge in linguistics. This intricate framework was further subjected to scrupulous scrutiny and validation by **Dr. Syed Shahrier Rahman**, an esteemed linguist and erudite professor situated in the Department of Linguistics at the University of Dhaka. The resultant transcribed dataset, a veritable linguistic tapestry, encapsulates an extensive array of morphological forms, numbers, acronyms, abbreviations, as

well as the nomenclature of places and individuals. Noteworthy in its design is the deliberate emphasis on Standard Bangla and the assimilation of borrowed words, reflecting the linguistic adaptability of native speakers. Significantly, the annotation process extends beyond mere linguistic formality, delving into the nuanced intricacies of how native speakers organically integrate Standard Bangla into their everyday discourse, particularly in the nuanced pronunciation of adapted loanwords within the contextual confines of Standard Bangla.

This research, animated by an earnest motivation to fill a critical void in Bengali NLP resources, culminates in the establishment of the most extensive Bangla IPA dataset to date. Beyond its sheer voluminosity, this contribution not only deepens our comprehension of Bengali phonetics but also opens up expansive vistas for groundbreaking advancements across a spectrum of NLP applications. In doing so, this thesis emerges as a beacon, illuminating pathways for further linguistic inquiry and technological innovation, all rooted in the rich tapestry of the Bengali language.

1.2 Our Contribution

This work represents a substantial contribution to the field of **Bangla IPA Transcription**, encompassing several key advancements:

1.2.1 In-Depth Study of Challenges

We initiated a meticulous investigation of the existing issues and complexities surrounding Bangla IPA transcription. This comprehensive analysis unearthed previously unidentified challenges and provided a crucial foundation for crafting our solutions.

1.2.2 Novel IPA Transcription Framework

Driven by the insights gained from our study, we meticulously designed a novel IPA transcription framework specifically tailored for Bangla. This framework addresses the intricacies of the language, offering a robust and consistent approach to IPA representation.

1.2.3 Pioneering Dataset Creation

Recognizing the scarcity of large-scale resources, we constructed a groundbreaking dataset, aptly named **DUAL-IPA**. This sentence-level parallel corpus encompasses **160,000+** samples, serving as the first-of-its-kind resource for Bangla IPA research and NLP applications.

1.2.4 Open-Source Accessibility

Firmly believing in fostering collaboration and innovation within the research community, we have chosen to open-source the DUAL-IPA dataset under the **CC BY-SA 4.0 license**, making it freely accessible to researchers and practitioners worldwide.

1.2.5 National Competition

The impact of our work extends beyond academic realms. We proudly collaborated with the Institute of Information Technology (IIT), University of Dhaka, to organize a national-level competition leveraging the DUAL-IPA dataset. Witnessing the participation of 104 teams, with over 50 actively utilizing our model, serves as a testament to the real-world application potential and community engagement fostered by this research.

Through these combined efforts, we believe we have made a significant contribution to advancing the field of Bangla IPA transcription. By addressing existing challenges, providing innovative solutions, and fostering open-source accessibility, we pave the way for future advancements in language processing and related NLP applications for Bangla.

Chapter 2

Related Work

2.1 International Phonetic Alphabet (IPA)

The requirement of such a model to transcribe the Bengali language to IPA requires a phonetic transcription scheme to represent the transcription and the pronunciation patterns for the language. The International Phonetic Alphabet (IPA) stands as the sole standard for phonetic writing systems, accounting for its significance in the scientific examination of a language's phonetics. Regardless of the language in question, the International Phonetic script predominantly relies on Roman characters as well as incorporates modified elements from diverse scripts like Greek to convey phonetic notation. The IPA-provided symbols such as (t, ɛ, ʃ, k, ˌ) are to be used for even those language that does not employ the Roman alphabet, such as Bangla, Hindi, Japanese, or Korean.

Since its establishment in 1886, the International Phonetic Association has been concerned with developing a system of symbols that maintains a balance between usability and inclusivity, which includes the wide variety of sounds present in languages all over the world. [1]The main purpose of IPA is to represent specific speech sounds rather than the abstract linguistic units known as phonemes, although it is also used for phonemic transcription. The IPA follows a common policy of using one letter for each segment. As a result, two letters are not put together to represent one single sound. For example, 'shine' - in this word 'sh' is used to convey one single sound. The IPA doesn't usually provide separate characters for sounds that aren't differentiated in known languages. Both broad and narrow transcriptions can be used using the IPA.

2.1.1 Bangla Vowels

Chatterji [2] used Jones [3]'s cardinal vowel system to explain the Bangla vowel system. He claimed that the Bangla language has seven primary vowels ই/i/, এ/e/, অ্যা/æ/, আ/a/, ও/o/, অ/ɔ/, and উ/u along with their corresponding nasal counterparts /ĩ ẽ æ̃ ã õ ɔ̃ ũ/. Chatterji also noted that Bangla vowels are generally articulated in a lax manner, imparting the characteristic 'timbre' to the vowel system. Morshed [4] categorized the vowels as /i, u, e, o, ae, ɔ, and a/, including two high, two high-mid, two low-mid, and one low vowel. Ali [5] investigated vowel contrasts, defining phonological properties, and reported the same number of vowels, with a subtle distinction. He employed the symbol /ɛ/ to represent the vowel /æ/ as described by Morshed [4].

In a separate study, Hai [6] analyzed the vowels of Standard Bangla using the concept of cardinal vowels. He claimed that there are eight vowels ই/i/, এ/e/, অ্যা/æ/, আ/a/a, ও/o/ ও'/o'/, অ/ɔ/, and উ/u/ in the Bangla language. He categorizes ই/i/, এ/e/, অ্যা/æ/ as front vowel and ও/o/, ও'/o'/, অ/ɔ/, and উ/u/ as back vowel. In contrast to Morshed [4], Hai did not classify the Bangla vowel আ/a/a as occupying a central position. He explained that the Bangla আ/a/ sound differs from the neutral quality of the English /a/ and is distinct from the Urdu close /ə/ sound. Instead, he characterized it as an open vowel. Hai also pointed out the presence of an additional vowel in the Bangla vowel system, denoted as /o'/. He explained that when producing the /o'/ sound, the lips are slightly less rounded compared to the /o/ sound. However, there isn't a significant difference in the gap between the jaws, and the back of the tongue is not raised as much as it is when articulating the /o/ sound. This led him to term it as yotized o (o^{y}), known in Bangla as অভিশ্রুত /ob^{ɦ}ɪsɾuṭo/ ও /o/ or ও' /o'/. This observation was supported by Huq [7]. An example provided for this distinction is between বিয়ের ক'নে/bɨer ko'ne/ and ঘরের কোণে /g^{ɦ}ɔrer kone/. Nevertheless, it's worth noting that there is limited empirical evidence to support this concept. On the contrary, the claim that the number of vowels is seven is backed by Pobitro Sorkar (1992) and Puny Sloka Ray (1997) as noted in Ali [5].

2.1.2 Bangla Semi-Vowels

According to Chatterji [2] and Sen [8], there are two Bangla semivowels, namely অন্তস্থ ব/w/ and অন্তস্থ য /y/. Hai [6] contends that there are three semivowels: অন্তস্থ ব/w/, অন্তস্থ য /y/, and অন্তস্থ ই/i/. Morshed [4] argues that while অন্তস্থ ব/w/ and অন্তস্থ য /y/ are considered semivowels in English, they do not possess similar status in Bangla. A different perspective was presented by Ferguson and Chowdhury [9], who claim that there are four semivowels: /i e o u/. It is noted in Ali [5] that this assertion was supported by Pobitro Sharker and Ghonesh Boshu (1998). Along with the ই/i̯/, উ/u̯/, and ও/o̯/, there is a fourth semi-vowel which is এ/e̯/ that is found at the end of the word in the form of 'য়' such as হয়/hɔe̯/, যায়/jae̯/ [5].

2.1.3 Bangla Diphthongs

Sen [8] noted that the Bangla has two diphthongs: ঐ(oɪ) and ঔ(ou). These combinations of two sounds do not fit the conventional definition of diphthongs but are represented in written form. In linguistic terms, they are referred to as digraphs [5]. On the contrary, Chatterji [2] claimed that there are 25 diphthongs in standard Bangla. Hai [6] asserted that there are a total of 31 diphthongs, categorizing them into 19 regular and 12 irregular ones. However, he also once argued that there are only 18 diphthongs, as noted by Ali [5], who in turn asserts that there are 17 diphthongs in Bangla. The government-approved IPA website acknowledges the regular 19 diphthongs, but they have used the diphthong /ui̯/ two times and did not consider the /eo̯/ diphthong.

2.1.4 Bangla Consonants

There have been numerous past studies, primarily rooted in articulatory phonetics, that have examined the articulatory and acoustic characteristics of Bangla consonants. It is described in Hai [6] that Bangla consonant has 20 stops, 7 fricatives, 4 nasals, 1 lateral, 1 trill, 2 flaps, and 1 glide; totaling 36 consonants. Hai [6] claims that there's only on phone close to /ʃ/ in Bangla. Huq [7] presented a slightly different categorization of a total of 35 consonants, presenting 21 stops, 5 fricatives, 3 nasals, 1 lateral, 1 trill, 2 flaps, and 2 glides. Morshed [4] stated that Bangla includes 20 stops, 4 nasals, 4 fricatives, 1 lateral, and 2 flaps, totaling 31 consonants. On the other hand, Ali [5] argued that Bangla has 20 stops, 3 nasals, 3 fricatives, 1 lateral, 2 flaps, 1 trill, and 2 glides, resulting in a total of 32 consonants.

Chapter 3

Bengali IPA Transcription Framework

Despite the widespread use of the Bengali language worldwide, there's a notable absence of a comprehensive IPA transcription framework and modeling. While the government-endorsed IPA system exists, it doesn't always offer clear explanations for specific diacritic usage, nor does it provide consistent reasonings for transcribing loaned words, accounting for morphological variations, or giving accurate IPA transcriptions. Besides, there remain unresolved debates among linguists regarding the inventory of vowels, semi-vowels, diphthongs, and consonants in Bengali. Scholars like Abdul Hai [6] have observed that the existence of long vowels in the language does not make a difference in the meaning and specific tongue positions for vowel/a/, which leads us to questions about the articulation manner of morphological suffixes and accurate numbers of pure vowels in the language.

Regional variations in Bengali further complicate matters, impacting not only the pronunciation variation among individual speakers but also how sounds are produced based on different regions and dialects. Noting all these drawbacks of the Bengali language, we propose an IPA framework that we've employed to create a dataset of 70,000 words, alongside a modeling approach for accurate Bengali-to-IPA transcription. It's worth mentioning that our suggested phonetic representations may not be universally accepted, and users are encouraged to substitute specific phonemes with alternatives that better align with their linguistic preferences. With the readily available IPA chart, individuals can readily determine which sounds best match the intended IPA representation.

3.1 Vowels

In our proposed IPA, we conducted a thorough review and made some revisions that were then incorporated into our dataset. It's important to note that the vowel sounds in Bengali are articulated in a lax manner. After carefully listening to the IPA sounds provided by Peter Ladefoged (1975), we devised a chart where these two sounds are considered true equivalents. We recommend substituting /ɐ/ for /a/ when representing the Bengali letter 'আ'. The /a/ is an open vowel and it's produced towards the front of the mouth. On the other hand, /ɐ/ is produced at the center of the mouth and the mouth is slightly less open while articulating this which is more suitable for the Bangla letter 'আ' rather than the /a/sound. Similarly, for the Bengali letter 'ই', we propose representing it as /ɪ/. The position of /ɪ/ is a near-high, front vowel in comparison to /i/ which is a high, front vowel. While producing the /ɪ/sound, the position of the tongue remains slightly lower and back in the mouth in comparison to the /i/. The reason we propose /ɪ/ for the Bangla letter 'ই' is that the /ɪ/is a lax vowel and when we produce the 'ই' sound, there is less muscular tension in the tongue. This adjustment better aligns with the articulation of native Bengali speakers, where the /ɐ/ and /ɪ/ sounds are more appropriate. Regarding the অ্যা sound, both /æ/ and /ɛ/ are true equivalents. However, for consistency in our dataset, we have chosen to use /ɛ/ exclusively.

	Front	Central	Back
High	ɪ		u
High-mid	e		o
Low-mid	æ/ɛ		ɔ
Low		ɐ	

Table 3.1: Bengali Proposed Vowel Chart

3.2 Semi-vowel

Semi-vowels, often referred to as glides or semi-consonants, are phonetically identical to vowels but function at the syllable's boundary rather than as the syllable's central component known as the nucleus. The glide diacritic which is an inverted breve (◌̯) is used beneath semi-vowels in the International Phonetic Alphabet (IPA) to denote their dual nature, exhibiting features of both vowels and consonants. We have proposed four semi-vowels that have been incorporated into the dataset.

Semi-vowel	
bangla	ipa
ই	/i̯/
উ	/u̯/
ও	/o̯/
এ	/e̯/

Table 3.2: Proposed Semi-vowel Chart

3.3 Diphthongs

To maintain clarity, it's wise to include all 31 diphthongs, especially considering the presence of regional dialects that might feature words absent in standard Bengali. Moreover, accurately discerning diphthongs requires audio reference rather than relying solely on written text. It's essential to acknowledge irregular diphthongs, particularly those involving the /a/ sound, which lacks a semi-vowel counterpart in Bengali. Therefore, the determination of whether a diphthong is rising or falling as well as whether is a vowel cluster or actually a diphthong hinges on careful consideration. According to Dr. Syed Shahrier Rahman, Bengali diphthongs are quite contextual. So, we should look into the possible combinations instead of the available combinations.

3.4 Consonants

[Note, in the chart 3.4, Unasp. is used to convey, unaspirated, and Asp. is used to convey, aspirated]

*In certain contexts, the 'হ' /h/ have extra careful articulation. For example, the word 'হ্রাস' in normal conversation would be pronounced as /ɾɐʃ/ but a news presenter or a person reciting a poem would articulate with an aspiration sound in the initial position of the word such as /ʰɾɐʃ/.

*In the Bangla language, the য় /j/ is not articulated as a phoneme but is commonly used in the co-articulation. For example, দেউলিয়া /d̪eulɪʲɐ/, নিয়তি /nɪʲɔt̪ɪ/, নিয়ম /nɪʲom/- in these three words the Bangla letter 'য়' is pronounced as palatalized /ʲ/. দাবায় /d̪ɐbɐe̯/, জয় /ʄɔe̯/ - 'য়' is pronounced as diphthong.

There are a few disputes among linguists regarding Bengali consonants. We have discussed the issues and provided a solution which we have followed in this consonant chart and in the curated dataset.

Semi-vowel (Regular)		
Bangla	IPA	Examples
আই	a̯i̯	চাই
আএ	a̯e̯	যায়
আউ	a̯u̯	দাউ
আও	a̯o̯	যাও
অ্যাএ	æ̯e̯	দ্যায়
অ্যাও	æ̯o̯	ম্যাও
অএ	ɔ̯e̯	কয়
অও	ɔ̯o̯	কও
এই	e̯i̯	সেই
এউ	e̯u̯	কেউ
ওই	o̯i̯	বই
ওএ	o̯e̯	ধোয়
ওউ	o̯u̯	নৌকা
ওও	o̯o̯	শোও
ইই	i̯i̯	দিই
ইউ	i̯u̯	মিউ
উই	u̯i̯	রুই
উউ	u̯u̯	কুউ
এও	e̯o̯	শোও

Semi-vowel (Irregulars)		
Bangla	IPA	Examples
ইয়ে	i̯e̯	বিয়ে
ইয়া	i̯a	টিয়া
ইও	i̯o̯	নিও
এআ	ea	দেয়া
এয়ো	e̯o̯	দেও
অ্যায়া	æa	দ্যায়া
ওয়া	o̯a	ধোয়া
ওএ	oe	কয়ে
উয়ে	u̯e	শুয়ে, ধুয়ে
উয়া	u̯a	নুয়া, ধুয়া
উয়ো	u̯o	কুয়ো

Table 3.3: Proposed List of Diphthongs (Regular and Irregular)

Place		Bilabial		Dental		Alveolar		Post-Alveolar	Palatal		Velar		Glottal
Manner													
		Unasp	Asp	Unasp	Asp	Unasp	Asp		Unasp	Asp	Unasp	Asp	
Stop	Voiceless	প/p/	ফ/pʰ/	ত/t̪/	থ/t̪ʰ/	ট/t/	ঠ/tʰ/		চ/c/	ছ/cʰ/	ক/k/	খ/kʰ/	
	Voiced	ব/b/	ভ/bʱ/	ত/t̪/	ধ/d̪ʱ/	ড/d/	ঢ/dʱ/		জ, য /ɟ/	ঝ/ɟʱ/	গ/g/	ঘ/gʱ/	
Nasal		ম/m/				ন, ণ/n/					ঙ, ং/ŋ/		
Tap						র /ɾ/							
Flap						ড়/ɽ/, ঢ়/ɽʰ/							
Fricatives						শ, স/s/		শ, ষ, স/ʃ/					*হ/h/
Lateral						ল/l/							
Approximant									*য়/j/				

Table 3.4: Proposed Consonant Chart

3.4.1 Disputing the Plosive and Affricate Argument

	চ	ছ	জ	ঝ
Plosive	c	c^h	ɟ	$ɟ^h$
Affricate	tʃ	$tʃ^h$	dʒ	$dʒ^ɦ$

Table 3.5: Phonetic Transcription for Plosive and Affricate

There has been a longstanding dispute among linguists about whether certain Bengali sounds, particularly those represented by চ, ছ, জ, and ঝ, should be classified as affricates or plosives. Dr. Muhammad Abdul Hai [6] agreed with this discussion and sided with the view that these sounds are best described as palatal plosives. In this proposal, we agree with this perspective, as when we consider how we articulate these words, they seem to align more closely with plosives rather than affricates.

3.4.2 ট - Alveolar or Retroflex

	ট	ঠ
Alveolar	t	t^h
Retroflex	ʈ	$ʈ^h$

Table 3.6: Phonetic Transcription of Consonants

The ট sound in Bengali is produced with the alveolar ridge acting as the fixed point in the mouth. The active part, which usually includes the tip of the tongue, interacts with this ridge during articulation [6]. Abdul Hai [6] acknowledges that while articulating words, the tip of the tongue curls up and back. This is why he categorizes it as an alveolar-retroflex-plosive sound [6].

3.4.3 ফ - /p^h/ and /f/ both or only /p^h/

The pronunciation of the sound represented by ফ in Bengali can vary regionally. While it is generally considered a plosive sound, in some regions, it may be perceived as a labio-dental fricative /f/ [6].

As a native speaker, when I articulate words like ফরি, ফাইজলামি, ফরালেহা, I bring my bottom lip close to the upper teeth, creating a narrow passage for the air to flow through. This suggests

that □ can indeed resemble a labio-dental fricative sound. However, it's important to note that this can still be a subject of debate, with variations observed from region to region and from person to person. As for written transcription, without the aid of audio from a regional speaker, accurately determining whether ফ is pronounced as a plosive or a labio-dental fricative can be challenging. But if we have audio data from regional speakers, we can transcribe words that are pronounced with dialectal accents with /f/ sound (such as fɔralæha) and other words that are also found in standard Bengali with /p^h/ (such as p^hul, p^hɔʃol)

Another concern with the /p^h/ sound is when dealing with borrowed foreign words, there can be further variations in pronunciation. The choice between considering them as labio-dental fricatives or recognizing potential individual differences ultimately depends on the availability of audio data. In cases where only written text is available, the decision is typically based on IPA transcription without the benefit of audio confirmation. But as a native speaker when I articulate these borrowed words, I sometimes pronounce them with their English accent and sometimes I might produce them with a Bengali native accent. For example, when I pronounce the words ফরজ, ফারসি, ফিউচার - I produce the labio-dental /f/. However, when I am producing the word ফেইক, I sometimes pronounce them with the plosive /ph/ sound and when I am producing the word in a certain context, I might produce the /f/ sound. Without enough audio and video data, it isn't easy to come to a certain generalization.

3.4.4 Trill r or Tap ɾ

The government website employs the trill 'r' sound, but in Bengali words like রাজা, রাজ্য, and রাগ we don't naturally use the trill sound. To ensure better pronunciation, the tap sound (ɾ) would be more suitable for Bengali.

3.4.5 Contextual Substitution of phoneme

The Benglai /ɟ/ is a voiced palatal stop and in standard Bengali, there is no voiced alveolar fricative /z/. Furthermore, in the Bengali language, the closest phoneme with the labio-dental fricatives such as /f/ and /v/ are aspirated labial stops /p^h/ and /$b^ɦ$/. However, many words in standard Bengali are adapted from foreign languages such as English, Arabic, Farsi, and so on. When native speakers articulate these loaned words they do not pronounce them in the same way a native English or native speaker Arabic does, but pronounce these with a native influence. Hence, for loaned words where the speaker articulates these foreign phonemes in a certain word context, we will consider these phonemes (/ɟ/, /f/, /v/) in the IPA transcription.

3.4.6 Voiced Aspiration

Aspiration is a significant distinctive feature in the Bengali phoneme. It can be noted from the chart above, that in Bengali, ভ, ধ, ঢ, ঝ, and ঘ are voiced aspirated stops. Aspiration is about how much air leaves your mouth while articulating the phoneme. If an unvoiced consonant is aspirated, then an extra puff of air leaves the mouth after the primary articulation is complete. For example in /pʰ/, /t̪ʰ/, /cʰ/, /tʰ/, and /kʰ/ voiceless aspiration occurs, hence for the secondary articulation of the aspiration, we use /ʰ/ which is voiceless. On the other hand, /bʱ/, /d̪ʱ/, /dʱ/, /ɟʱ/ and /gʱ/ are voiced stops and for that reason, it is suitable to use a voiced aspiration /ʱ/ for the secondary articulation. In the govt-IPA, the aspiration suggestions for voiced stops have both voiced /ʰ/ aspiration and voiceless /ʱ/ aspiration as their secondary articulation. For instance, they kept both /bʰ/ or /bʱ/ for the transcription of the letter 'ভ' despite that the /ɦ/ should be voiced after voiced consonants.

3.5 Diacritics

Our proposed diacritics for standard Bengali.

Diacritics	
ʷ	Labialized
ʲ	Palatalized
◌̃	Nasalized

Table 3.7: Phonetic Transcription of proposed Diacritics

- **Labialized:** The use of labialized diacritics is found in Bengali words such as উপরওয়ালা /uporoʷɐlɐ/, দেওয়া /d̪eoʷɐ/, নেওয়া /neoʷɐ/, etc where the consonant sounds indicate that they are pronounced with rounded lips. In certain cases, diphthongs are pronounced with simultaneous lip rounding, such as রওশন /rɔ͡o̯ʷ.ʃon/.

- **Palatalized:** To determine the use of palatalized ʲ, we have followed two phonological rules. The rule for determining whether the Bengali consonant য় (y) is palatalized or functions as a diphthong is as follows:

  When the position of the য় is in the syllable-final, without a following vowel, it remains unpalatalized. For example, in compound words like মামলায় /mɐmlɐ͡e̯/, নিরাপত্তায় /nɪrɐpoẗ̪ɐ͡e̯/, etc.

Conversely, if a word with য় concludes with a vowel in the syllable's final position and does not have য় in the word's final position, it will be pronounced as a palatalized ^{j}. For instance, this can be observed in words like ছেলেমেয়ে /c^{h}eleme^{j}e/, খায়রুল /k^{h}ɐ^{j}erul/, and নিয়ক /nɨ^{j}om/.

- **Nasalized:** It was mentioned earlier that in Bangla, all seven oral vowels have their seven nasal counterparts, which is described using the nasalized diacritics /ĩ ẽ õ ã ɔ̃ ũ/. This nasalization of vowels in Bengali text is consistently indicated by a diacritic known as 'chandarabindu' (ঁ) placed above the relevant segment, and this occurrence is a common feature in Standard Bengali text.

3.6 Loan Words Consideration: Vowel and Consonant

In the Bangla language, using loaned words from foreign languages and using them with a different pronunciation in comparison to their native pronunciation is quite common. In the case of vowels, no foreign phonemes are produced by native speakers. For example, the English word 'foam', 'cloud', and 'flower' is pronounced as /foʊm/, /klaʊd/, and /flaʊə/ by native English speakers. However, /ʊ/ and /ə/ are not articulated by the Bengali native speakers. Instead, they pronounce these words using the existing vowel phonemes of the Bangla language.

On the contrary, there are a few cases where foreign words are pronounced using consonant phonemes, which does not exist in Bangla.

	Bilabial	Labio-dental	Example
Plosive	ফ /p^{h}/		
Plosive	ভ /b^{h}/		
Fricative		/f/	ফেইল /fe͡i̯l/
Fricative		/v/	ভিউ /vɪu/

Table 3.8: Trancription of foreign words.

Labio-dental fricative sounds such as /f/, and /v/ do not exist in the Bangla language but they are articulated by the native speakers when they produce loaned words with these phonemes.

	Palatal	Alveolar	Example
Plosive	জ, য/ɟ/		
Fricative		/z/	ম্যাগাজিন /mɛgɐzɪn/

Table 3.9: Transcription of foreign words.

Same case for the alveolar fricative phoneme /z/. Loaned words from Arabic and English languages such as মেরাজ /meɾɐz/, ম্যাগাজিন /mɛgɐzɪn/, মোনাজাত /monɐzɐt̪/ are continuously used in the Standard Bangla.

English words such as judge /ʤʌʤ/, and justice /dʒʌstɪs/ have voiced postalveolar affricate /dʒ/ which is not used by native Bengali speakers. They turn this affricate sound into the plosive sound /ɟ/ and articulate it as /ɟudɟ/ and /ɟustɪs/.

The English language has a voiceless dental fricative sound /θ/ which is not found in the Bangla language. They turn this phoneme into a voiceless aspirated dental plosive sound /t̪ʰ/. So 'think' is pronounced as /t̪ʰiŋk/ in its Bangla adaptive form.

	Alveolar	Alveolar	Example
Fricative	শ, স/s/	/s/	স্টপ/stɔp/

Table 3.10: Transcription of foreign words.

The /s/ is a voiceless fricative alveolar sound that is found in both Bangla and foreign languages such as English.

Chapter 4

Validation and Linguistic Challenges of Standard Bengali IPA

4.1 Morphological Variations in Words

The Bengali language exhibits an extensive array of morphological variations, presenting a challenge in accurately contextualizing the meaning of words in light of their morphological alterations. It poses a challenge to accurately represent these subtle morphological variations within the framework of the International Phonetic Alphabet (IPA). Consider the Bengali word আজকেই, transcribed as /ɐjkeɪ̯ː/, or loaned words with Bengali morphological extensions like মেক্সিকোতেও /meksɪkoto:/ and মেক্সিকোও /meksɪkoo:/. While these all end with a vowel, without a syllabic marker, it may not be immediately clear that these suffixes are part of the base word. However, by incorporating the lengthening diacritic after the word (the long vowel diacritic /ː/), this distinction becomes more apparent to the reader. The reason for utilizing this diacritic is rooted in certain linguistic contexts. In some cases, when producing specific vowels, some individuals perceive a long i: as merely an extended version of the short vowel, without any discernible difference in quality, i.e., without raising the tongue for the long sound. For instance, Bengali e: is slightly higher than Bengali e, and Bengali e̝ (short) falls midway between cardinal e and ɛ. This concept is supported in the work of Suniti Kumar Chatterji as well. Furthermore, this long vowel diacritic also clears out the confusion that no case of diphthongs is present here (মেক্সিকোও /meksɪkoo:/)

The issue with morphological suffixes may create confusion to distinguish them from diphthongs such as the above word গরুগুলোও /goruguloo:/, some might transcribe it as গরুগুলোও /goruguloo̯/ because there are two vowels together in the word. But if we notice carefully and break into the syllable of the /go.ɾu.gu.lo.oː/, both of the vowels belongs to different syllable, even if both of the vowels are beside each other the last vowel o is pronounced with a long sound. This is the reason we have annotated morphological variation in such cases with long vowel marks.

শুটিংয়ে /ʃu.tɪŋʲ.eː/
শুটিংও /ʃu.tɪŋ.oː/
গরুগুলোও /goɾuguloːː/

Table 4.1: Phonetic Transcription of morphological suffixes

4.2 Diphthongs

Our dataset contains cases of Bengali diphthongs. To accurately transcribe them, it's crucial to first identify whether they are indeed diphthongs. Syllabification serves as a method to recognize diphthongs, which makes the process easier. However, due to the shortness of time, we decided to avoid the process of syllabication of each word just to identify diphthongs. Another significant aspect in distinguishing diphthongs is the use of the glide. The upper diphthong glide (͡) is used to describe the movement of the articulatory vocal organs, particularly the tongue, from a higher position to a lower one during diphthong production. This downward movement contributes to the distinct sound of the diphthong. Each language possesses its own set of unique diphthongs. We've provided a diphthong chart, from which standard Bengali focuses primarily on the regular diphthongs. Understanding the role of the glide and accurately using it ensures the correct pronunciation of words in a given language.

Examples

পরিচর্যায়	/poɾɪcɔɾɟɐ̯͡e̯/
ভাই	/bʱɐ̯͡ɪ̯/
যাচাই	/ɟɐ.cɐ̯͡ɪ̯/
চাই	/cɐ̯͡ɪ̯/
দুই	/du̯͡ɪ̯/
বোঝাই	/bo.ɟʱa̯͡ɪ̯/

Table 4.2: Phonetic Transcription of Dipthongs

Sometimes, a few cases of standard Bengali are found which may confuse the reader, if a certain word has a diphthong or vowel cluster. For example, শিরোইলে is transcribed as /ʃɪɾoɪle/, here the ɾoɪ constitutes one single syllable, but the question remains if it is a vowel cluster or diphthong. Bengali native speakers articulate this word in this way where a downward movement of tongue position from o to ɪ occurs. As a result, the o stays as a pure vowel and glides toward ɪ̯

which creates a diphthong. Hence, the final transcribed text is /ʃɪro͡i̯le/. If the pronunciation of the word were something such as /ʃɪ.ro.ɪ.le/ where the ই letters are pronounced as a pure vowel and separately from the syllable then the final result might have been something different.

4.3 Loan words

Native Bengali speakers commonly integrate vocabulary from English, Arabic, Farsi, and Portuguese into their speech. As a result, distinctive phonemes of these languages, which may not be common in standard Bangla, are spoken by native speakers. Due to their frequent usage, these phonemes may not be distinctly differentiated from the standard Bangla phonetic inventory. This challenges IPA models in accurately recognizing and transcribing these foreign phonetic elements.

In our dataset, we have a significant number of English and Arabic words. To transcribe these words, we consider how native Bengali speakers, adhering to the standard Bengali form, would pronounce them. Since standard Bengali users often employ a more received pronunciation when uttering these words, we have annotated them accordingly. Hence, we have used /z/, /f/, /v/ /s/ phonemes for the letters, জ/য, ফ, ভ, শ/স respectively. These sounds are not commonly present in the native Bengali language, but to transcribe the borrowed foreign words, we have employed these.

For example,

ফেইক	/feɪk/
শিডিউল	/ʃɪ.dɪ.ul/
মোস্তাফিজ	/mosʈɐfɪz/
যারহাদ	/zɐɾhɐɖ/
ফজর	/fɔzoɾ/
রদ্রিগেজ	/ɾɔɖɾɪgez/

Table 4.3: Phonetic Transcription of borrowed foreign words

4.4 English Diphthong and Triphthong in Bengali Adaptive Form

In English words with diphthongs, the presence of schwa/ə/ can influence the pronunciation. It appears in unstressed syllables usually containing the neutral, unstressed vowel sound. This

leads to subtle variations in how diphthongs are articulated. For example, ‘power’- in the word, the diphthong /aʊ/ is followed by the schwa sound in the unstressed syllable. Or for the word ‘water’, the first syllable may be reduced to a schwa sound, especially if it’s unstressed. It might sound like ”wuhAbbreviation-ter.” However, when these words are adapted by the Bengali speaker they will be pronounced like /pa.o̰͡ʷaɾ/ /o̰͡ɐ̯.teɾ/.

Bengali speakers adopt English diphthongs that do not contain schwa and the pronunciation tends to align with the native English pronunciation. For example, ’high’ is transcribed in the Bangla as /hɐ̯͡ɪ̯/, boil as /bɔ͡ɪ̯l/, and time as /tɐ̯͡ɪ̯m/.

The English language contains triphthongs, which is a rare case in the Bangla language. In the case of English triphthongs, native Bengali speakers tend to avoid pronouncing the word as a triphthong. Instead, they convert it into a diphthong and therefore avoid pronouncing the triphthong word. For example, in English, the word ‘fire’ is pronounced as /fʌɪə/, which in Bangla is transcribed as /fɐ͡ʲe̯.ɐɾ/. Cases like these are found in these words as well - ‘hour’ /aʊər/, which is pronounced as /ɐ.o̰͡ʷɐ̯ɾ/, ‘prayer’ /preɪər/, pronounced as /pɾe.ɐɾ/, ‘pure’ /pjʊr/ pronounced as /pɪo̰͡ɾ/.

Hence the only concern while transcribing these words is how a native speaker pronounces them.

ফায়ার	/fɐʲe.ɐɾ/
ফাইনাল	/fɐ̯͡ɪ̯.nal/
শুটআউটে	/ʃut.ɐ̯͡ʊ̯teː/

Table 4.4: Phonetic Transcription of adaptive English words

In the first example, fɐʲe.ɐɾ is transcribed for the English word ’fire’. The native English speaker pronounced it as faɪər where the diphthong aɪ glides into schwa ə in the second syllable. However, the Bangla language does not have a schwa ə sound as a result for this English diphthong word native Bangla speakers use the existing sound to produce the loaned word as fɐʲe.ɐɾ which does not have a diphthong in the adaptive form.

The pronunciation of words by Bengali speakers can vary based on regional accents and specific contexts. Even a standard native speaker may pronounce certain words differently depending on the situation, which could lead to variations in IPA transcription. Unless the transcription is based on audio data, ensuring accurate contextual transcription can be a challenge.

4.5 Transcribing Numbers

In the dataset, there are numbers represented in various forms like "19টা", "১ম", "১৯৮৯", "১০০০", or in the context of phone numbers and house numbers. To transcribe these, we followed an IPA transcription based on how we naturally pronounce them. For instance, "২০৬" is transcribed as "d̪uɪ̯ʃo cʰoe̯". When numbers are pronounced individually, they are transcribed accordingly, for example, "২০৫০" as "d̪uɪ̯ ʃunno pãc ʃunno".

4.6 Handling the cases of Abbreviations and Acronyms

To ensure dataset accuracy and disambiguate between abbreviations and acronyms, we established a specific protocol. When transcribing abbreviations like "ম., ড., মো.", we referred to the context to identify their full forms, which in this case were "মহাম্মদ", "ডাক্তার", and "মোহাম্মদ". We then proceeded to transcribe the entire words. In the case of acronyms like "এসএসসি", "মূসক", and "পিডিডি", we applied IPA notation for accurate representation. Handling these types of transcriptions poses certain challenges. Sometimes মহাম্মদ might be spelled and pronounced as মহাম্মাদ or only স. is only given in a sentence and the transcriber has to assume the words if proper indication is not given in the sentence. So with abundant acronyms and abbreviations in a language, the transcription of IPA for these may produce incorrect transcriptions.

Acronym Examples

No	Acronym	IPA
1	এসএসসি	esessɪ
2	পিডিডি	pɪdɪdɪ
3	মূসক	muʃɔk

Table 4.5: Phonetic Transcription of Acronym

Abbreviation Examples

No	Abbreviation	Bangla Word	IPA
1	ম.	মহাম্মদ	mɔhɐmmɔd̪
2	মো.	মোহাম্মদ	mohɐmmɔd̪
3	ডা.	ডাক্তার	dɐktɐr

Table 4.6: Phonetic Transcription of Abbreviation

4.7 Orthographic Challenges

Bengali orthography may not always align perfectly with phonetic transcription, requiring careful interpretation. Our dataset has been curated from written texts, based on the specific annotator's pronunciation intuition, as pronunciation sometimes varies from individual to individual. In spite of this, the pronunciation of a word might match word to word in the IPA transcription. Such as হ্রাসমান /ɾɐʃmɐn/, the হ letter here is not pronounced the way it is pronounced in the word হলুদ /holud̪/. Also in the spelling of the word হলুদ, there is not any 'ও' visible but while articulating the word an /o/ sound has been produced and that's how the word has been transcribed.

4.8 Placement of Diacritics

IPA transcription involves a meticulous and time-consuming manual process. Accurate placement of diacritics and special characters is critical for correctly representing sounds. For instance, if we were to transcribe the Bengali word দোয়েল as /doel/ or /d̪o͡e̯l/, rather than /d̪oʲel/, it would lead to an inaccurate pronunciation.

Chapter 5

Methodology

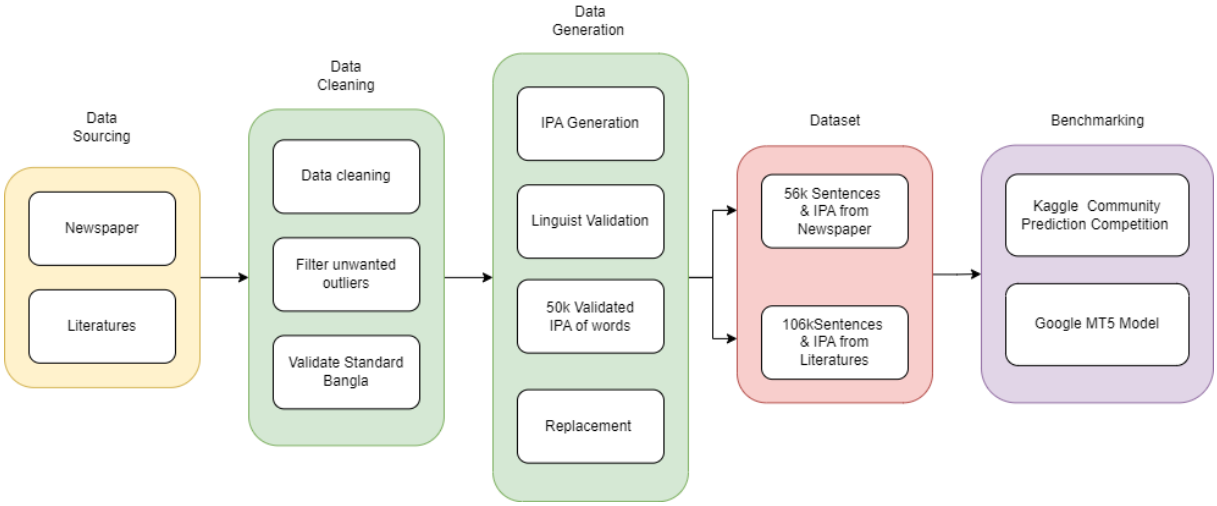

Figure 5.1: Workflow Diagram of the whole Project

The realm of Bangla IPA transcription has long awaited a comprehensive and robust dataset to fuel advancements in research and applications. In this work, we proudly present the DUAL-IPA dataset, a pioneering resource meticulously crafted to address this need. Standing as the first-of-its-kind large-scale dataset specifically designed for Bangla IPA, DUAL-IPA sets a new benchmark for quality, diversity, and accessibility.

This chapter meticulously documents the journey of constructing DUAL-IPA, encompassing every step from data collection and expert annotation to comprehensive statistical analysis. We unveil the intricate processes involved in sourcing diverse Bangla sentences, harnessing the expertise of linguists for accurate IPA transcription, and streamlining the annotation workflow through innovative techniques. Furthermore, we delve into the dataset's statistical composition, offering valuable insights into its vocabulary breadth, sentence length distribution, and phoneme landscape. Through this in-depth exploration, we not only illuminate the characteristics of DUAL-IPA, but also demonstrate its immense potential to empower researchers and practitioners in the field of Bangla NLP.

5.1 Data Collection and Preparation

5.1.1 Data Sources

- Data were collected from newspapers, online platforms, e-books.
- We wanted to make sure we cover all the conventional sources of standard Bangla for both word and sentence level that we use in our daily life.

5.1.2 Data Scraping Techniques

- To scrape all these data, we wrote multiple python scripts specific to each source.
- As different platform and formats have different data structures and most of them are not in a standard format when scraped, we had to study the data types first properly in order to find patterns so that we can maximize our output in terms of number and quality.

5.1.3 Data Cleaning and Preprocessing

- Removing duplicates, irrelevant content, and noise.
- Addressing missing values with justification.
- Language identification and filtering for non-promito bangla contents.

5.2 IPA Transcription Process

We were given a few words and IPA. We constructed a simple septa graph to identify unique phonemes. Using these phonemes and septa graph we made a Jupiter notebook that can generate IPA given a word. The notebook generates 116 IPAs.

5.2.1 Word-Level Transcription

Using this notebook, we generated around 50k IPAs for 50k unique words from 56k sentences provided by Bengali.AI These sentences were taken from newspapers.

5.2.2 Sentence-Level Transcription

We generated around 162k IPA from online sources and e-books after a rigorous process of collecting and cleaning the data. Then, using the same model as the word-level data, 162k IPA was generated for the 162k sentences.

5.3 Linguistic Review and Adjustments

- After every transcription, a team of linguists from University of Dhaka thoroughly reviewed each word and sentence to make sure they are on par with the standard rules and regulation.
- Then, based on the feedbacks of the linguists, changes were made to the data to make sure the dataset it as accurate as possible.
- There were some debate about some transcriptions, that were mitigated by taking opinion from **Professor Dr. Syed Shahrier Rahman from Department of Linguistics, University of Dhaka**, who also led the team of linguists for this project.

5.4 Dataset Creation and Finalization

- The final dataset is **two CSV** file, one for newspaper sentence IPA transcription and one for Literature sentence IPA transcription, containing two columns which are: sentences, clean_validated_ipa sentences, IPA.
- The whole dataset contains 160,000+ Bengali words and sentences with their corresponding IPAs.
- we have chosen to open-source the DUAL-IPA dataset under the **CC BY-SA 4.0 license**, making it freely accessible to researchers and practitioners worldwide.

5.5 Achievements

Through this work, we achieved

- 160,000+ word and sentence level dataset for Bengali to IPA transcription.
- Human and Machine Validation of the dataset.
- Arrange a national level competition on the dataset in collaboration with **IIT Software Engineers' Community (IITSEC), University of Dhaka**, where this dataset was put into test and different models were created based on it.
- Submitting our paper in the prestigious **LREC-Coling - 2024 conference**.

Chapter 6

Dataset Preparation

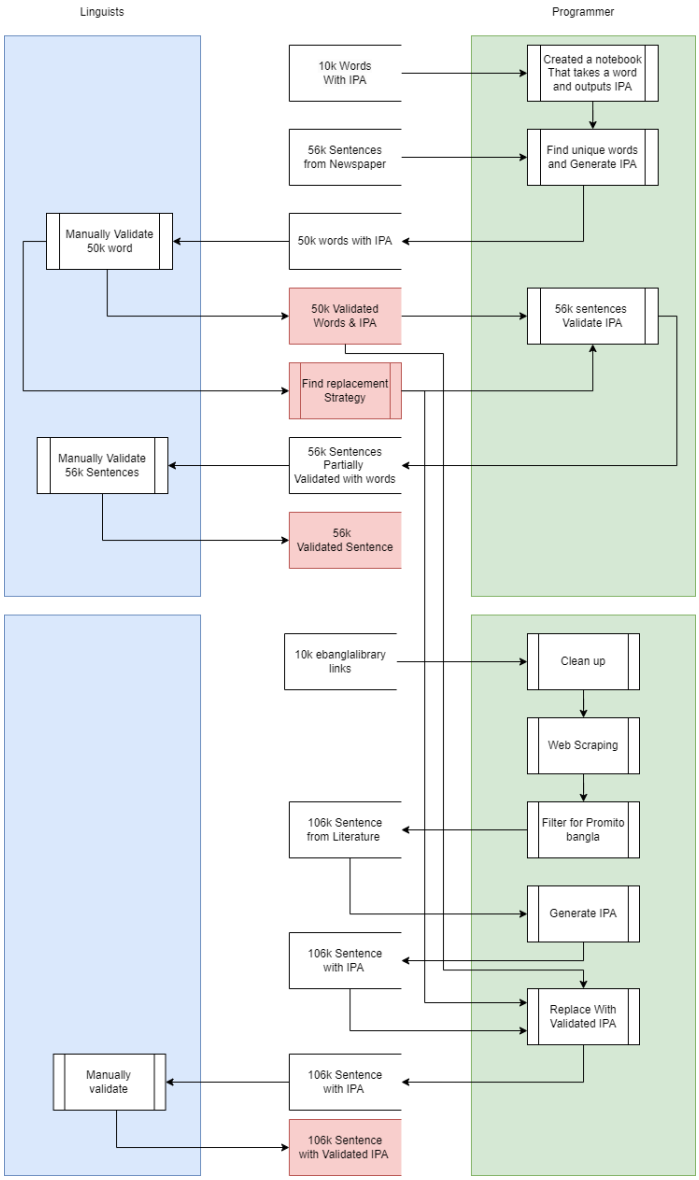

Figure 6.1: Dataset Preparation Steps

Here we're going to describe the rigorous process of preparing the DUAL-IPA dataset step by step, explaining the whole process end to end which is shown above in the Figure 6.1.

6.1 Data Collection

6.1.1 Sourcing a Diverse Corpus

- The Journey began with identifying suitable sources for Promito Bangla words and sentences.
- Meticulously selected two distinct domains as shown in the figure 6.2:

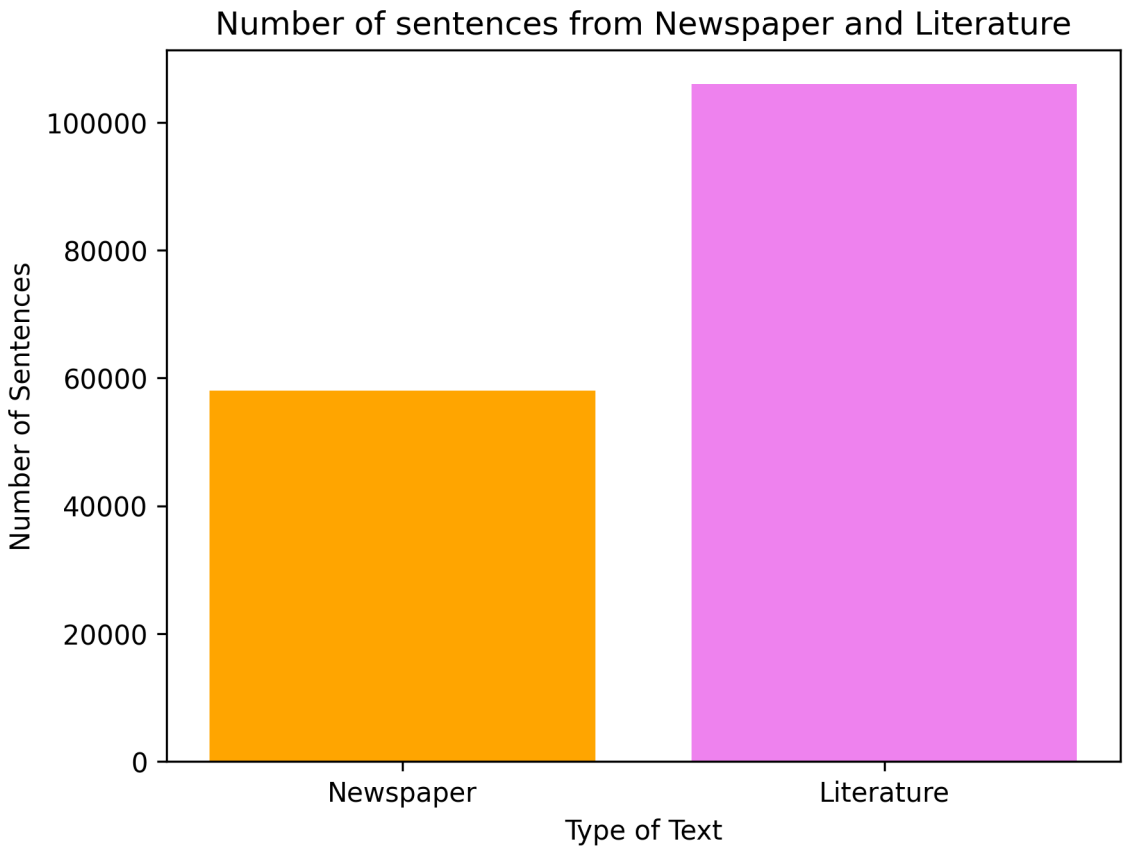

Figure 6.2: Sentence sources

  - **Newspapers (33%):** Online news articles capturing contemporary Bangla words with diverse vocabulary, which was provided to us by Bengali.AI.
  - **Literature/Books (66%):** Scraped novels, poems, stories and published works from Ebangla Library to capture nuanced beauty and historical richness.
- The Balanced approach ensures that the dataset encompasses both dynamic daily communication and the enduring essence of Bangla literature.

6.1.2 Quantity

In total, 162k+ data were curated for the DUAL-IPA dataset.

- **Words:** In total, almost 50k unique words were curated from the 56k sentences that are from newspapers, which were provided by Bengali.AI.
- **Sentences:** Total 162k sentences were curated for the dataset from newspaper and literature combined, which was scraped and curated by us. Here is the process.
  - We were given a list consisting of, 10993 unique Ebangla Library links.
  - We removed the non-promito from the list. Then we had 9971 links. Then we had 9971 links. The links contained sentences from "রমণীমোহন মল্লিক","শ্রীকৃষ্ণকীর্তন", "বাংলা বেদ - ঋগ্বেদ সংহিতা", "বাংলা হাদিস", "বাংলা মহাভারত", "বাংলা রামায়ণ", "গীতা-প্রসঙ্গ", "বাংলা গীতা", "ঈশ্বরচন্দ্র বিদ্যাসাগর", "কমলকুমার মজুমদার", "কাব্যগ্রন্থ", "গীতিগ্রন্থ", "বোদলেয়ার: তাঁর কবিতা", "মাইকেল মধুসূদন দত্ত - চতুর্দশপদী কবিতাবলী", "মাইকেল মধুসূদন দত্ত - মেঘনাদবধ কাব্য", "রাজশেখখর বসুর কবিতা", "হিন্দুধর্ম", "হুমায়ুন আজাদ - কাব্যসংগ্রহ", "শামসুর রাহমান - কাব্যগ্রন্থ","রুদ্র মুহম্মদ শহিদুল্লাহ - কাব্যসংগ্রহ","কবিতাবলী" etc.
  - We created a simple web scrapper to collect the sentences from the links.
  - We collected 120k sentences, but these contained non-promito words. In total, we collected, 106494 unique sentences from Bangla literature.

6.1.3 Diversity

To make sure the data doesn't represent a particular type or group of words, we made sure to source our data from multiple categories of literature categories and newspapers that represent our daily communication usage and also get the essence of the rich literature history we have in Bangla.
This dataset contains different lengths of words in the sentences that have been curated. From 1 to 17 character is present in words. The statistics are shown below in the figures.

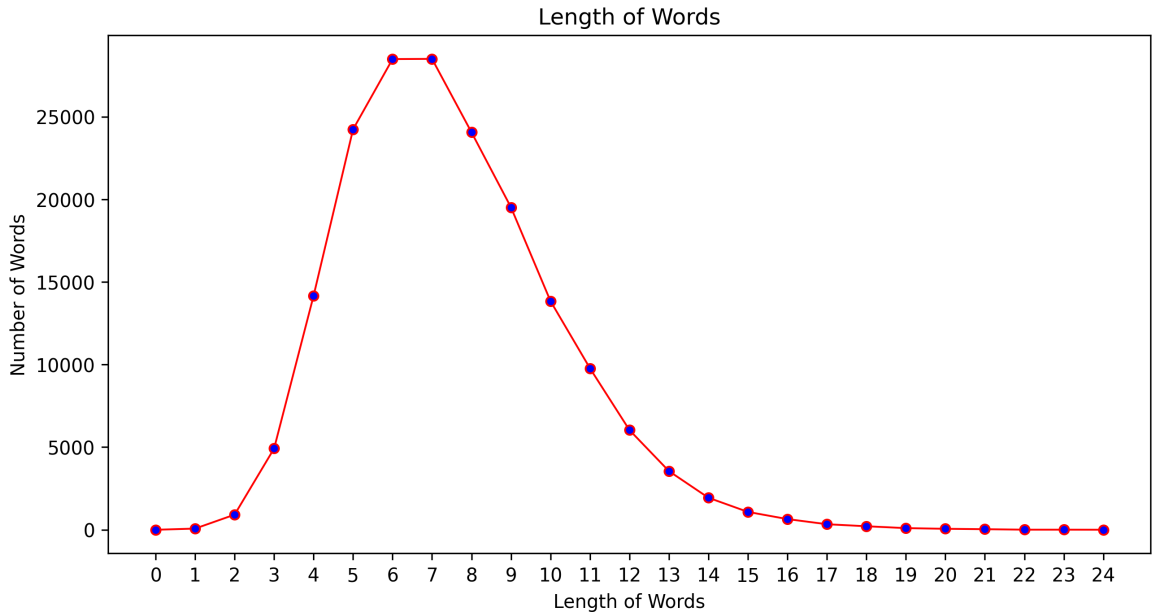

Figure 6.3: Length of words

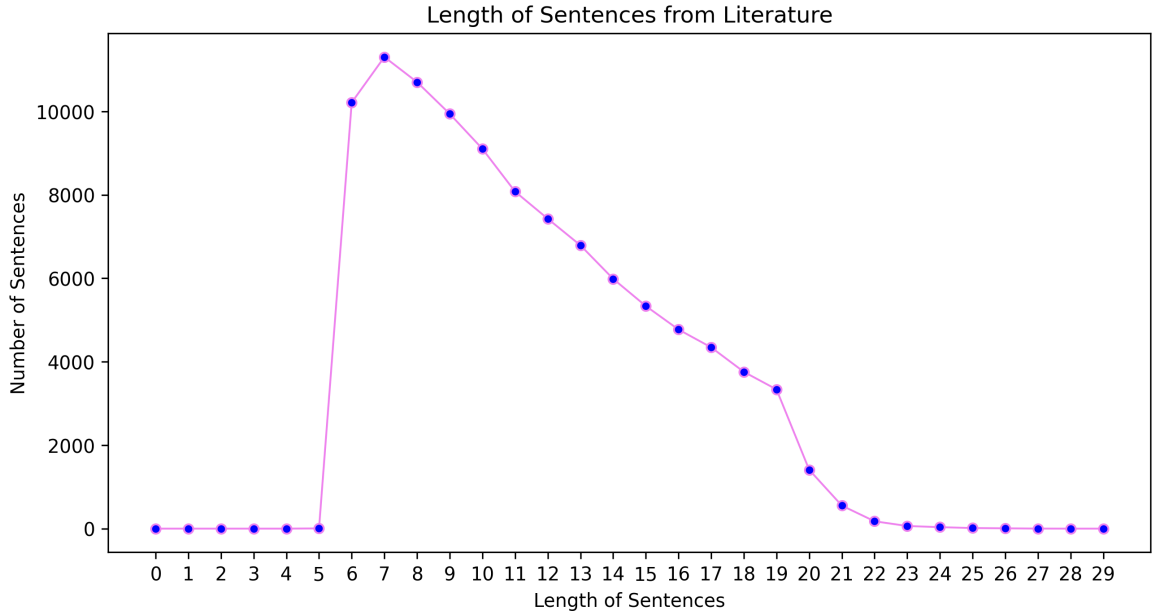

Figure 6.4: Number of words in Literature

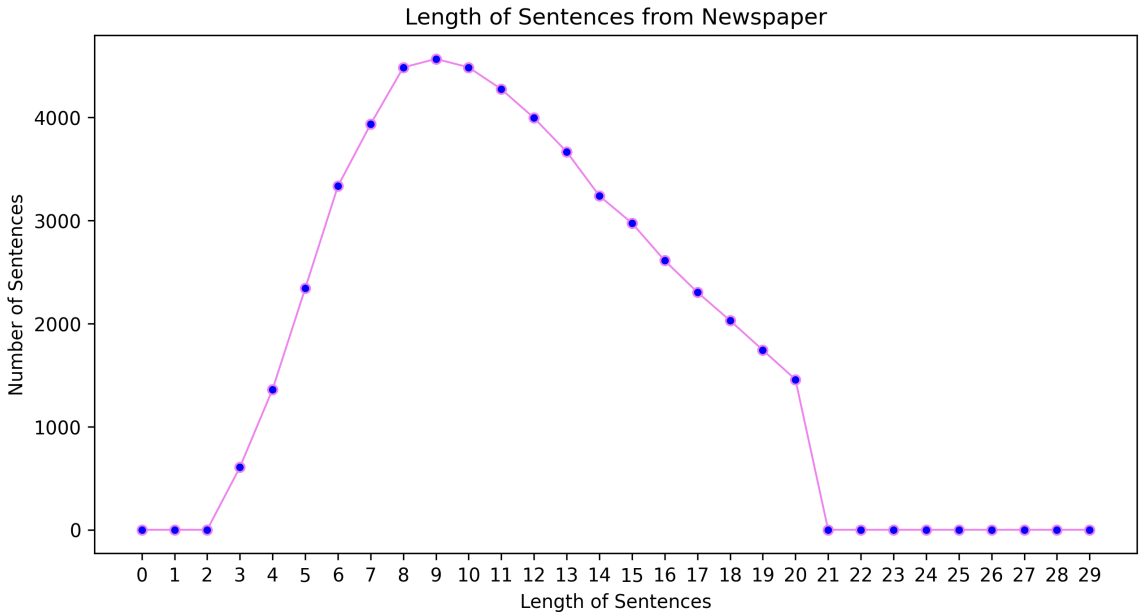

Figure 6.5: Number of words in Newspaper

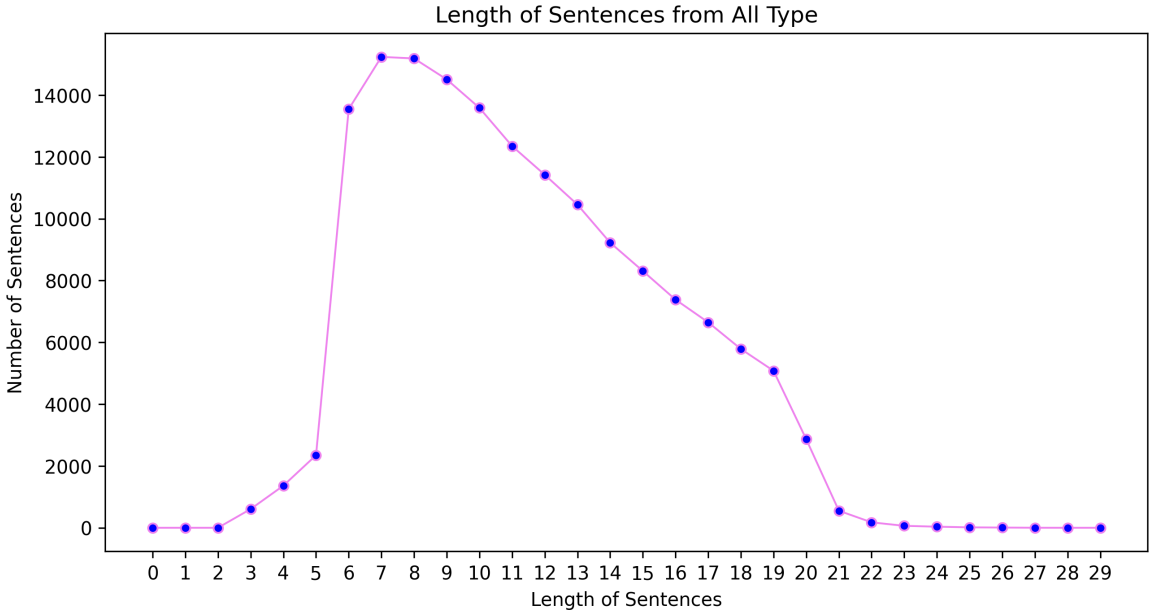

Figure 6.6: Number of words in All Type

6.1.4 Challenges in Sourcing

As different platforms and formats have different data structures and most of them are not in a standard format when scraped, we had to study the data types first properly in order to find patterns so that we can maximize our output in terms of number and quality.

6.2 Data Preparation

6.2.1 Cleaning

- As the word-level data was provided by Bengali.AI, it was already cleaned. So, we used it directly for transcription,
- On the other hand, the sentence-level data was fully sourced and curated by us from scratch. So, we had to programmatically and manually clean those from the dataset.

6.3 IPA Transcription Process

We were given a few words and IPA. We constructed a simple septa graph to identify unique phonemes. Using these phonemes and septa graph we made a Jupiter notebook that can generate IPA given a word. The notebook generates 116 IPAs.

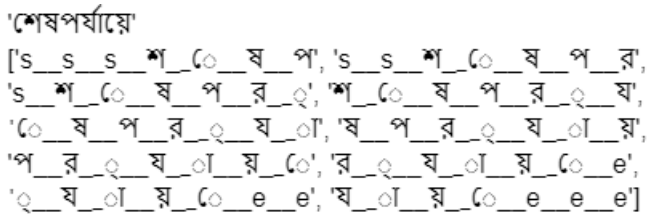

Figure 6.7: Septa graph

6.3.1 Word-Level Transcription Framework

- Using the notebook, we generated 56k IPAs for the 56k sentences provided to us by Bengali.AI which were taken from newspapers.
- These sentences had, 43356 unique words.
- We generated the IPAs for these words and gave them to the linguist team for initial validation.
- The linguist team validated these words and found some IPA mischaracterization, such as:(”a” , ”ɐ”),(”i” , ”ɪ”),(”æ” , ”ɛ”),(”r” , ”ɾ”),(”ʲ ”,”ɐ̯ ”)
- Using these validated words and new characters, we fixed the 56k sentences and gave them to the linguist team for final word-level validation.

6.3.2 Sentence-Level Transcription Framework

Process of replacing validated IPA words in generated IPA sentences.

First, we found the special characters in all sentences of the dataset.

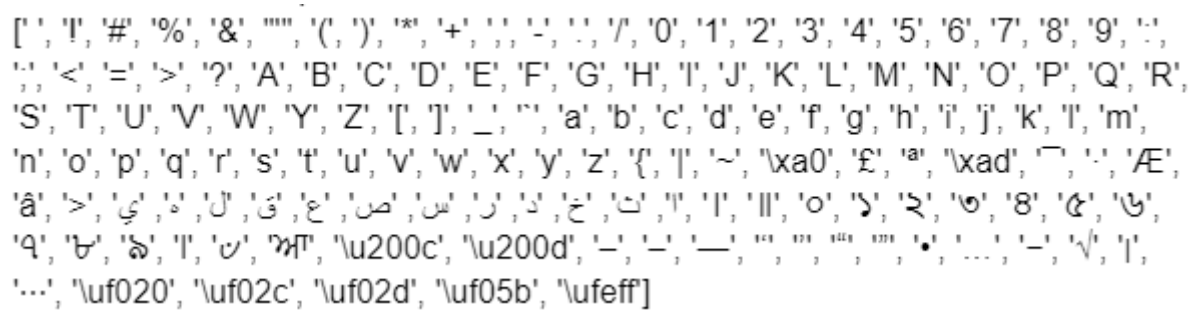

Figure 6.8: Special Characters

Then we concat all the sentences using the special character " _$_ " .

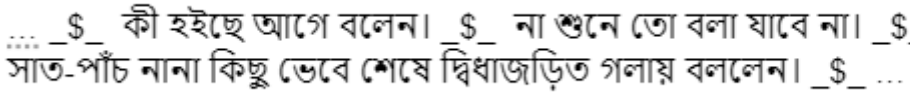

Figure 6.9: Concated sentences

We identified the words and replaced them in place with the corresponding IPS. Replaced words were marked with the prefix __##__.

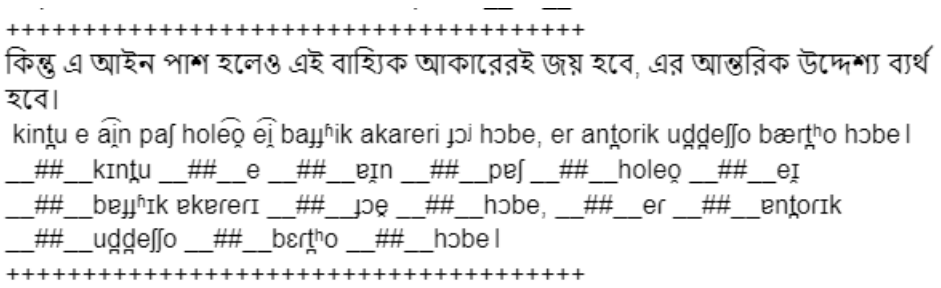

Figure 6.10: Replaced words

Then split the mixed sentence on " _$_ ". So the IPA sentence Indexes matched the main index.

Then we found the IPA sentences that contained Bengali words and re-run the sentence with the notebook.

6.4 Validation

A team of four individuals with undergraduate and graduate degrees in linguistics in the supervision of Professor Dr. Syed Shahrier Rahman from the University of Dhaka was entrusted to do the annotation and validation of the whole DUAL-IPA dataset for every step.

	A	B	C	D	E	F	G
1		words	word_fix	ɪpas_old	ɪpas	sentence	assigned person
1692	14	অঙ্কন		ɔŋkon	ɔŋkon	১৬ সেপ্টেম্বর অনুষ্ঠিত 'চ' ইউনিটের সাধারণ জ্ঞান পরীক্ষায় উত্তীর্ণ মোট ১ হাজার ৫৫২ জন অঙ্কন পরীক্ষায় অংশ নেন।	Dawood
1693	10	অঙ্কুরেই		ɔŋkureɪ̯	ɔŋkureɪ	যে তাজা প্রাণ জাতীয় প্রয়োজনে রক্তাক্ত হতে চায় সেই প্রাণ অঙ্কুরেই বিনষ্ট হতে পারে।	Dawood
1694	9	অঙ্কের		ɔŋker	ɔŋker	তাদেরকে আটক করে নিয়ে এসে থানার বাইরে রেখে মোটা অঙ্কের টাকার বিনিময়ে ছেড়ে দেয়ার জন্য দফায় দফায় বৈঠক হয়।	Dawood
1695	5	অঙ্গ		ɔŋgo	ɔŋgo	মূল দলের কমিটি না থাকায় অঙ্গ ও সহযোগী সংগঠনগুলোর সাংগঠনিক কার্যক্রমে দেখা দিয়েছে গতিহীনতা।	Dawood
1696	2	অঙ্গ-প্রত্যঙ্গ	অঙ্গ-প্রত্যঙ্গ	ɔŋgo-protɔŋgo	ɔŋgo-prot̪ɔŋgo	শরীরের সব অঙ্গ-প্রত্যঙ্গ নিস্তীব হয়ে হৃদযন্ত্রের ক্রিয়া বন্ধ হয়ে যাওয়ার কারণে তিনি মারা যান।	Dawood
1697	4	অঙ্গ-প্রত্যঙ্গগুলো		ɔŋgo-prot̪ɔŋgogulo	ɔŋgo-prot̪ɔŋgogulo	ইতিমধ্যে গরুর শরীরের ভেতরের অঙ্গ-প্রত্যঙ্গগুলো ক্ষতিগ্রস্ত হয়।	Dawood
1698	5	অঙ্গ-সংগঠনগুলো		ɔŋgo-ʃɔŋgɔtʰongulo	ɔŋgo-ʃɔŋgɔtʰongulo	এরপর থেকে প্রতিবছর বিএনপি ও অঙ্গ-সংগঠনগুলো বিভিন্ন কর্মসূচির মধ্য দিয়ে এ দিনটিকে তারেক রহমানের 'কারাবন্দি' দিবস হিসেবে পালন করে আসছে।	Dawood
1699	6	অঙ্গ-সংগঠনের		ɔŋgo-ʃɔŋgɔtʰoner	ɔŋgo-ʃɔŋgɔtʰoner	বর্তমান যুগে পুলিশ ও সরকারি দলের অঙ্গ-সংগঠনের নেতাকর্মীদের এমন আচরণ কোনোভাবেই গ্রহণযোগ্য হতে পারে না।	Dawood
1700	9	অঙ্গন		ɔŋgon	ɔŋgon	নির্বাচনকে কেন্দ্র করে এখন সরগরম আদালত পাড়াসহ জেলার রাজনৈতিক অঙ্গন।	Dawood
1701	5	অঙ্গনে		ɔŋgone	ɔŋgone	তাদের কর্মকাণ্ডে বাংলাদেশ সম্পর্কে আন্তর্জাতিক অঙ্গনে ভুল তথ্য দেয়া হচ্ছে।	Dawood
1702	4	অঙ্গনের		ɔŋgoner	ɔŋgoner	মেলায় জাতীয় ও আন্তর্জাতিক অঙ্গনের ৫৫টি প্রতিষ্ঠান অংশ নিয়েছে।	Dawood
1703	9	অঙ্গপ্রতিষ্ঠান		ɔŋgoprot̪ɪ̯tʰen	ɔŋgoprot̪ɪ̯tʰen	লাফস গ্যাস বাংলাদেশ শ্রীলংকার বিখ্যাত লাফস হোল্ডিং লিমিটেডের একটি অঙ্গপ্রতিষ্ঠান।	Dawood
1704	3	অঙ্গরাজ্যটিতে		ɔŋgoreɪ̯ot̪t̪e	ɔŋgoreɪ̯ot̪t̪e	বৃষ্টিপাতের এ পূর্বাভাস অঙ্গরাজ্যটিতে 'বিপর্যয়' তৈরি করতে পারে বলে বলে জানিয়েছে যুক্তরাষ্ট্রের জাতীয় ঘূর্ণিঝড় কেন্দ্র।	Dawood
1705	1	অঙ্গরাজ্যে		ɔŋgoreɪ̯je	ɔŋgoreɪ̯je	এসব অঙ্গরাজ্যে ডেমোক্রেটের ১০১৭ ও রিপাবলিকানের ৬৬১ জন ডেলিগেট নির্বাচিত করবেন।	Dawood
1706	1	অঙ্গরাজ্যের		ɔŋgoreɪ̯jer	ɔŋgoreɪ̯jer	এসব অঙ্গরাজ্যের নিচু ও ঝুঁকিপূর্ণ এলাকা থেকে লোকজনকে নিরাপদ স্থানে সরে যাওয়ার নির্দেশ দিয়েছে কর্তৃপক্ষ।	Dawood
1707	16	অঙ্গসংগঠন		ɔŋgoʃɔŋgɔtʰon	ɔŋgoʃɔŋgɔtʰon	তবে চার বছর আগে বিভাজনের সময় একটি পক্ষ মাক্সবাদী এবং অপরটি বাংলাদেশ সমাজতান্ত্রিক দলের (বাসদ) মূলধারার অঙ্গসংগঠন।	Dawood
1708	11	অঙ্গসংগঠনের		ɔŋgoʃɔŋgɔtʰoner	ɔŋgoʃɔŋgɔtʰoner	খালেদা খানমের পরিচালনায় সমাবেশে জেলা আওয়ামী লীগ, কেন্দ্রীয় মহিলা লীগসহ বিভিন্ন অঙ্গসংগঠনের নেতাকর্মীরা উপস্থিত ছিলেন।	Tanvir
1709	9	অঙ্গার		ɔŋgɐr	ɔŋgɐr	' তিনি আরও বলেন, খামারের মালিক ঘুমন্ত অবস্থায় পুড়ে অঙ্গার হয়ে যায়।	Tanvir
1710	13	অঙ্গীকার		ɔŋgɪkɐr	ɔŋgɪkɐr	স্থানীয় ইউপি মেম্বার জামাল উদ্দিন মোরশেদাকে চাপ দিলে ছাত্রীকে শুক্রবার ফেরত দেবে বলে অঙ্গীকার করে।	Tanvir
1711	7	অঙ্গীকারবদ্ধ		ɔŋgɪkɐrbɔd̪d̪ʱo	ɔŋgɪkɐrbɔd̪d̪ʱo	তিনি বলেন, দেশের অধিকার রক্ষায় তাইওয়ানের জনগণ অঙ্গীকারবদ্ধ।	Tanvir
1712	6	অঙ্গীকারাবদ্ধ	অঙ্গীকারবদ্ধ	ɔŋgɪkɐrebɔd̪d̪ʱo	ɔŋgɪkɐrbɔd̪d̪ʱo	এ ক্ষেত্রে বাংলাদেশ তার ভূমিকা পালনে অঙ্গীকারাবদ্ধ।	Tanvir
1713	10	অঙ্গীকারের		ɔŋgɪkɐrer	ɔŋgɪkɐrer	এই প্রকল্পটি প্রধানমন্ত্রী শেখ হাসিনার নেতৃত্বে বর্তমান সরকারের সাহস ও অঙ্গীকারের বিরাট পরীক্ষা।	Tanvir
1714	13	অঙ্গে		ɔŋge	ɔŋge	এরপর তার স্বামী সোনা মিয়া তাকে মারধর করে তার বস্ত্র খুলে তার গোপন অঙ্গে মরিচ দিয়ে নির্যাতন চালায়।	Tanvir
1715	2	অচল		ɔcɔl	ɔcɔl	অবরোধের ফলে অচল হয়ে পড়েছে সদরসহ গোটা রাঙ্গামাটি জেলা।	Tanvir
1716	1	অচলাবস্থা		ɔcolebɔsʈʰe	ɔcɔlebosʈʰe	এ অচলাবস্থা নিরসনে উদ্যেগ করার মতো উদ্যোগ ও তৎপরতা দৃশ্যমান নয়।	Tanvir
1717	17	অচলাবস্থার		ɔcolebɔsʈʰer	ɔcɔlebosʈʰer	দলের মুখপাত্র জিডিএল নরসিমহা রাও বলেন, 'রাহুল গান্ধী নিজেই নিজেকে হাস্যাস্পদ করে তুলছেন, তার দল কংগ্রেসই সংসদের অচলাবস্থার জন্য দায়ী।	Tanvir
1718	8	অচিন্ত	অচিন্ত্য	ɔcɪnto	ɔcɪnto	শাহরাস্তি উপজেলা স্বাস্থ্য কমপ্লেক্সের আবাসিক মেডিকেল অফিসার ডা. অচিন্ত কুমার চক্রবর্তী জানান, অসুস্থ শিক্ষার্থীদের শরীরে কোনো উপসর্গ পাওয়া যায়নি।	Tanvir
1719	14	অচিন্ত্য		ɔcɪnto	ɔcɪnto	কিন্তু প্রধান শিক্ষক আনিচুর রহমান তাকে সভাপতি না করে ওয়ার্ড আওয়ামী লীগের সাধারণ সম্পাদক অচিন্ত্য হালদারকে নির্বাচিত করেন।	Tanvir

Figure 6.11: Validation stage of the linguistics team

6.4.1 Metrics

Linguists made sure to follow all the latest Bangla Academy conventions for spelling and pronunciation of the words that are included in the dataset, as well as they made sure to follow all the standards of the IPA conventions in the whole process. The dataset had to go through multiple validation steps. Each step made sure to increase the quality of this work.

6.4.2 Results

With the help of the linguists team, the final output which is the preparation of DUAL-IPA dataset was possible with the highest level of accuracy from our end. Here is figure 6.12 showing Number and Types of Words in Dataset.

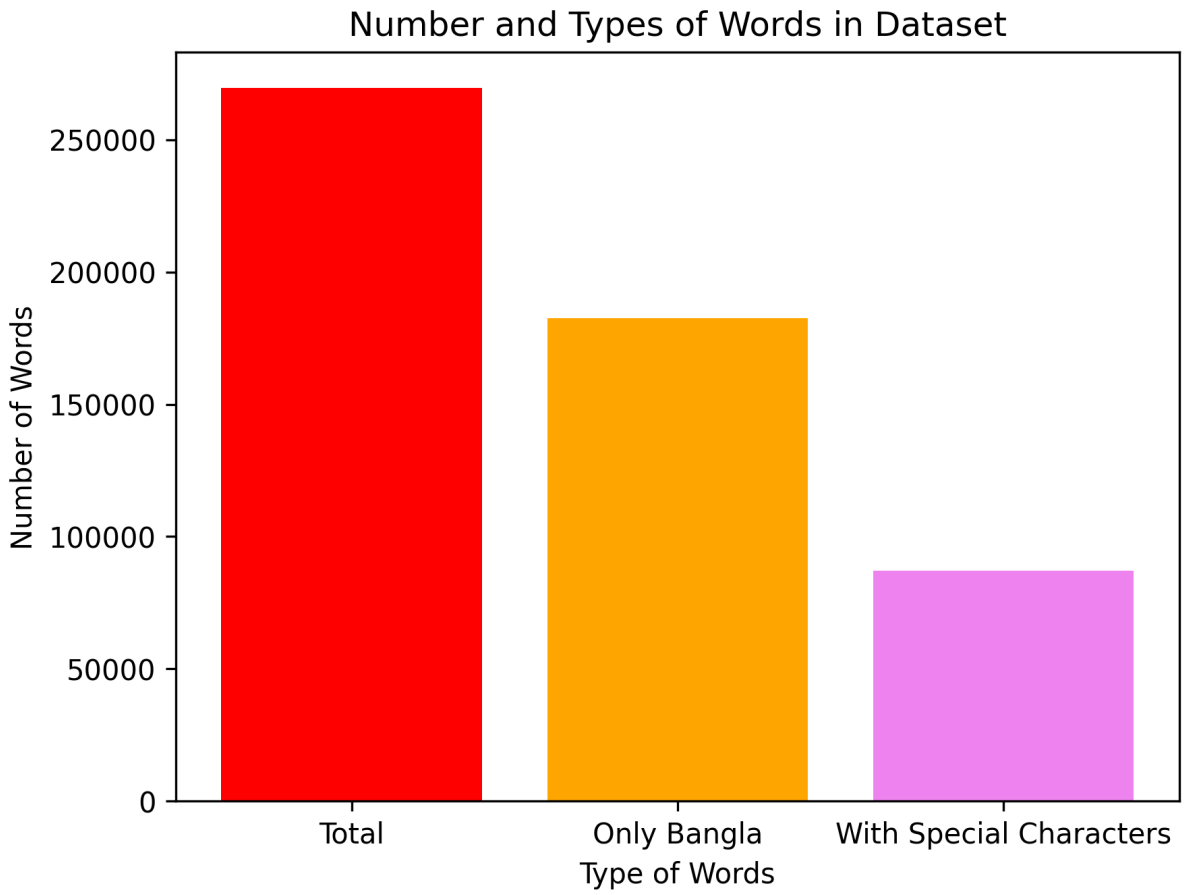

Figure 6.12: Number and Types of Words in Dataset

6.5 Ethical Considerations

6.5.1 Privacy:

No personally identifiable information (PII) was directly collected or stored within the dataset. All the data are collected from public sources only.

6.5.2 Bias:

As the dataset comprises text scraped from online sources, it may inadvertently reflect potential biases present in the original content. We acknowledge this inherent limitation and emphasize the importance of using the dataset responsibly and considering potential biases during analysis and interpretation of results. Additionally, we encourage further research into developing debiasing techniques for NLP applications using such datasets.

6.5.3 Transparency:

We have documented the data collection process and annotation guidelines with transparency, enabling users to understand the potential limitations and biases associated with the dataset. Sharing this information encourages responsible use and facilitates further discussion about ethical considerations in NLP research.

6.5.4 Respect:

Throughout the research process, we maintained respect for the cultural and linguistic nuances of the Bangla language. We consulted with language experts and adhered to established IPA transcription protocols to ensure the accuracy and cultural sensitivity of the dataset.

By acknowledging and addressing these ethical considerations, we strive to promote responsible development and utilization of the DUAL-IPA dataset, contributing to ethical advancements in Bangla NLP research.

Chapter 7

DUAL-IPA Dataset

7.1 The Heart of the Research: DUAL-IPA Dataset

- Meticulously designed DUAL-IPA dataset: 160,000+ Bangla sentences with linguist-validated IPA transcriptions.
- Construction required strategic data collection, rigorous annotation processes, and expert validation.
- The final dataset is **two CSV** file, one for newspaper sentence IPA transcription and one for Literature sentence IPA transcription, containing two columns which are: sentences, clean_validated_ipa sentences, IPA as shown in the figure 7.1

7.2 Expert Annotation: Harnessing Linguistic Prowess

- IPA transcription entrusted to four individuals with undergraduate and graduate degrees in linguistics in supervision of **Professor Dr. Syed Shahrier Rahman from Department of Linguistics, University of Dhaka.**
- Each sentence reviewed by all four experts following a stringent IPA transcription protocol.
- Independent evaluator cross-verified all annotated data to bolster consistency and ensure dataset's integrity.

	A	B
1	**sentence**	**clean_validated_ipa**
2	থাইল্যান্ড বা শ্যামদেশের রাজধানী।তপন এক মাস হল ব্যাংককে এসে কাজে যোগ দিয়েছে।সিং ব্রাদার্সের মালিক গোবিন্দ সিং-এর বয়স প্রায় ষাট।	t̪heɪlend be ʃemd̪eʃer reɟd̪henɪ ltɔpon ek meʃ hɔl ben̪kke eʃe keɟe ɟog d̪ɪeche lʃɪŋ bred̪erser melɪk gobɪnd̪o ʃɪŋ-er bɔloʃ preɸ ʃet l
3	যদি বলেন, বাড়িতে নেইখুব হৈ-হল্লা করলে শেষ পর্যন্ত কি না বেরিয়ে পারবেন?অন্তত বকা-ঝকা করতে তো বেরুবেন একবার।	ɟod̪ɪ bolen, beɾɪt̪e neʃɪkʰub ho͡ɪ-holle korle ʃeʃ porʃont̪o kɪ ne berɪe perben?ɔn̪t̪ot̪o bɔke-ʃʰke kor̪t̪e t̪o beruben ekber l
4	বাঃ, বেশ নাম তো তোমার।কেমন ইতিহাসের গন্ধ আছে, তাই না?রতন পাশ থেকে টিপ্পনী কাটে।	be, beʃ nem t̪o t̪omer lkɛmon ɪt̪ɪheʃer gɔnd̪ʰo eche, t̪eɾ ne?rɔt̪ɔn peʃ t̪ʰeke tɪpponɪ kete l
5	টাকা চায় ভি-পি করে যেন পাঠায়।নারায়ণ ভট্টাচার্য গল্প পাঠালেন, দিন মজুর।একবার শরৎচন্দ্রের কাছে গেলে হয় না?	teke ceɸ bʰɪ-pɪ kore ɟeno pet̪ʰeɸ lnereɪon bʰɔttecerɟo gɔlpo pet̪ʰelen, d̪ɪn moɟur lɛkber ʃɔrot̪ɔnd̪rer kech̪e gele hɔɕ ne?
6	তার মানে ষোলো কলার এক কলা তখন বাকি।ধরে ফেললাম।	t̪er mene ʃolo kɔler ɛk kɔle t̪ɔkʰon bekɪ ld̪ʰore pʰellem l
7	যৌবন-নিপীড়িত দেহের প্রতিটি অণু-পরমাণু দিয়ে ভোগ করেছে।...তারপরতার পরের কথা পরে বলব।আপাতত শুনুন অলকানন্দার কাহিনি।	ɟo͡ubon-nɪpɪɾɪt̪o d̪eher proɾɪt̪ɪ ɔnu-pɔromenu d̪ɪe bʰog korech̪e l...t̪erpɔrɪt̪er pɔrer kɔt̪ʰe pore bolbo lepet̪ot̪o ʃunun ɔlokenɔnd̪er kehɪnɪ l
8	কথাটা মনে আছে তো?কেমন ভ্যাবাচেকা খাওয়া চোখে তাকাল অবনী, ঢাক গিলে বলল, স্ট্যান্ড করা মানে?-ফার্স্ট সেকেন্ড থার্ডের মধ্যে হওয়া।	kɔt̪ʰete mone ech̪e t̪o?kɛmon bʰɛbec͡ɕeke kʰe͡ɸʷe cokʰe t̪ekel ɔbonɪ, dʰek gɪle bollo, stend kɔre mene?-pʰest sekend t̪ʰerder mod̪d̪ʰe hʷe l
9	ঠোঁঠে শুষ্ক হাসি টেনে এনে বলল, বলুন।তুমি বলছি বলে রাগ করছ না তো?স্মার্ট মেয়ে বটে।	tʰɔ̈ʰe ʃuʃko heʃɪ tene ene bollo, bolun ltumɪ bolcʰɪ bole ɾeg korcʰo ne t̪o?serɪ meie bɔte l
10	কিন্তু ও জেগে থাকত, চিন্তা করত।বাচ্চারা চিন্তা করতে পারে না।ও পারে।	kɪnt̪u o ɟege t̪ʰekt̪o, cɪnt̪e kor̪t̪o lbeccere cɪnt̪e kor̪t̪e pere ne lo pere l
11	বাইরে থেকে ধরবার কোনো উপায় রইল না।বড় শান্তিতে ঘুমালাম সেরাত্রে-বেড়াল ছাড়া।	beɪre t̪ʰeke d̪ʰorber kono upeɸ roɪlo ne lbɔɾo ʃent̪t̪e gʰumelem ʃeret̪re-beɾel cʰeɾe l
12	গনগনে চোখে ম্যাডাম চেয়ে আছে তার দিকে।কিন্তু কীভাবে জানলি আমার সে ক্ষমতা আছে?আছে না কচু।	gɔngone cokʰe mɛdem ceie ech̪e t̪er d̪ɪke lkɪnt̪u kɪbʰebe ɟenlɪ emer ʃe kʰɔmot̪e ech̪e?ech̪e ne kocu l
13	বাবার কাছ থেকে উত্তরাধিকার সূত্রে পাওয়া।তার বদলে কি ছড়ি রেখে গেছে বলুন তো?দেখলে কোনও তফাত বুঝবেন না।	beber kech t̪ʰeke ut̪t̪oreɖʰɪker ʃut̪re peɡʷe lt̪er bod̪le kɪ cʰoɪ rekʰe gɛcʰe bolun t̪o?d̪ekʰle kono t̪ɔpʰeɾ buɲʰben ne l
14	অপরাধী রয়ে গেল অন্ধকারে।আর রইল আমার বৈধব্য।	ɔporeɖʰɪ roɪe gelo ɔnd̪ʰokere ler roɪlo emer boɪɖʰobbo l
15	কিন্তু লোকটা সত্যি দে দোষী না নির্দোষ আমি এখনো ঠিক বুঝছি না।সুনন্দ চুপ করে রইল কিছুক্ষণ।	kɪnt̪u lokte ʃot̪t̪ɪ de d̪oʃɪ ne nɪrd̪oʃ emɪ ekʰono tʰɪk buɲʰchɪ ne lʃunɔnd̪o cup kore roɪlo kɪcʰukkʰon l
16	আমার একটা কথা রাখুন।বিরাট নিশ্বাস ফেলে ইন্দ্রনাথ বললে, হায়রে বাংলার বধূ..গাছের সন্ধানটা বলে দিতে হবে, কেমন?আপনি পারবেন।	emer ekte kɔt̪ʰe rekʰun lbɪret nɪʃʃeʃ pʰele ɪnd̪ronet̪ʰ bolle, heɪre benler bod̪ʰu..gecʰer ʃond̪ʰente bole d̪ɪt̪e hɔbe, kɛmon?epnɪ perben l
17	আর গান হয়।কি বেজাত রে বাবা, ভাবা পারি না?তৃতীয় বছরে খালাসি জমি ভদুই ধানে হেসে উঠেছে।	er gen hɔɕ lkɪ beɟet̪ re bebe, bʰebe perɪ ne?ɾɪt̪ɪɾɪo bɔcʰore kʰeleʃɪ ɟomɪ bʰod̪ur d̪ʰene heʃe utʰeche l
18	ওটা আমি বিয়ের পর থেকে ওখানে দেখে আসছি।আমার চেহারা কি এলিসিয়ার মতো?অনেকটা তো বটেই, জবাব দিলেন মা।	ote emɪ bɪer pɔr t̪ʰeke okʰene ɖekʰe eʃcʰɪ lemer cehere kɪ elɪʃɪer mɔt̪o?ɔnekte t̪o bɔteɾ, ɟɔbeb d̪ɪlen me l
19	আশা করি হেনরিই ফিরে আসবে।মদের একটা প্রকার থেকে দারুনভাবে সতর্ক করা হয়েছে।তবে মদের কোনো প্রকারই ইসলামের দৃষ্টিতে রীতিসিদ্ধ নয়।	eʃe korɪ hen͡rɪɪ pʰɪre eʃbe lmɔd̪er ekte proker t̪ʰeke d̪erunbʰebe ʃɔt̪ɔrko kɔre holech̪e lt̪ɔbe mɔd̪er kono prokerɪ ɪslemer d̪ɾɪʃɪɾɪt̪e rɪɾɪʃɪd̪d̪ʰo nɔɕ l
20	একটা খবর আছে স্যার।''কি খবর?' আগ্রহ নিয়ে জিজ্ঞেস করল আহমদ মুসা।'আমি গিয়েছিলাম ওয়ে এমবাকে ডাকতে।	ekte kʰɔbor ech̪e seɾ l''kɪ kʰɔbor?' egroho nɪe ɾɪggeʃ korl ehmod̪ muse l'emɪ gɪlech̪ɪlem oie emobeke dekt̪e l
21	এরা গতকাল এসেছিল, পুলিশের লোক এরা নয়।''কারা হতে পারে বলে তোমার ধারণা?''জানি না স্যার।	ere gɔt̪okel eʃech̪ɪlo, pulɪʃer lok ere nɔɕ l''kere hot̪e pere bole t̪omer d̪ʰeronɐ?''ɪenɪ ne seɾ l

Figure 7.1: Sentence Level Dataset

7.3 Expediting the Process: Embracing Efficiency

- Meticulousness remained paramount, but efficiency was acknowledged.
- Multi-pronged approach to expedite annotation process:
  - **Pre-annotation:** Combination of rule-based and early-stage model-based techniques for preliminary annotation.
  - **Word-Level Correction:** Experts refined transcription at the word level, correcting whitespace-separated tokens.
  - **Sentence-Level Mapping:** Mapping word-level annotations back to corresponding sentences for integration and consistency.
  - **Homograph Resolution:** Final validation step addressing homograph cases, numerical representation, and remaining alignment errors.
- Combined strategies significantly streamlined annotation process, enabling dataset curation within a one-month timeframe.

7.4 DUAL-IPA Dataset: Rigor and Efficiency

- Dataset stands as a testament to unwavering commitment to rigor and efficiency.
- Each sentence embodies linguists' expertise, independent evaluation scrutiny, and innovative pre-annotation and manual refinement blend.
- Diverse sources, comprehensive coverage, and meticulous validation pave the way for groundbreaking advancements in Bangla NLP research and applications.

7.5 Dataset statistics (EDA)

Unveiling the intricacies of the DUAL-IPA dataset necessitates a thorough examination of its statistical composition. This section embarks on a journey of discovery, delving into the numerical landscape of the dataset and illuminating its key characteristics.

7.5.1 Quantifying the Corpus:

At the heart of the dataset lies a vast collection of 150,000 sentences, each capturing the essence of the Bangla language. To facilitate efficient training and evaluation, the dataset has been strategically split into two segments:

- **Training Split (100,000 sentences):** This portion serves as the training ground for models, providing them with rich linguistic material to learn from and hone their skills.

- **Test Split (50,000 sentences):** This segment acts as the ultimate challenge, where models demonstrate their acquired knowledge by attempting to accurately transcribe unseen data.

7.5.2 Glimpses from the Dataset:

The phoneme distribution of the sentences is shown in Figure 7.2. Also, the length distribution of the Dual-IPA dataset is shown in Figure 7.3.

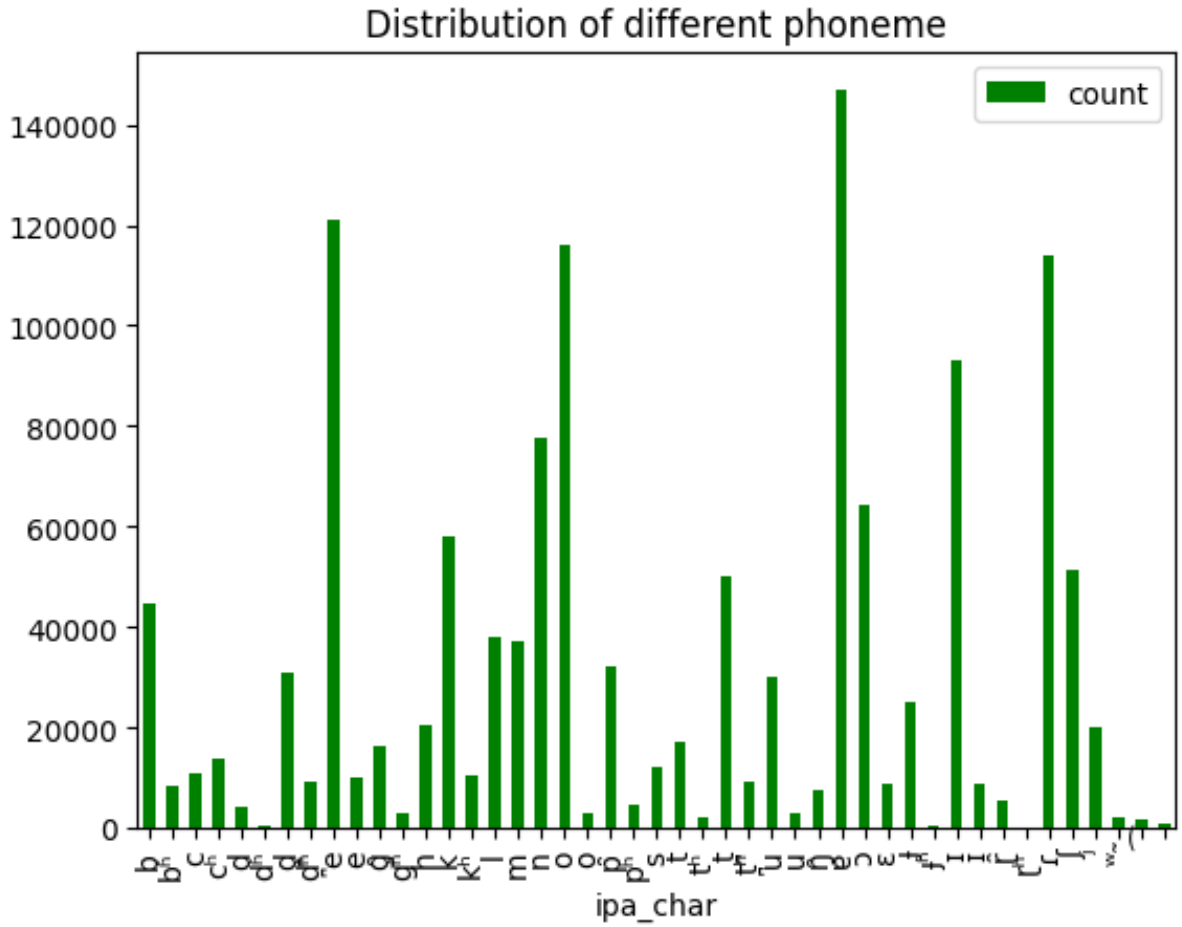

Figure 7.2: Phoneme Distribution

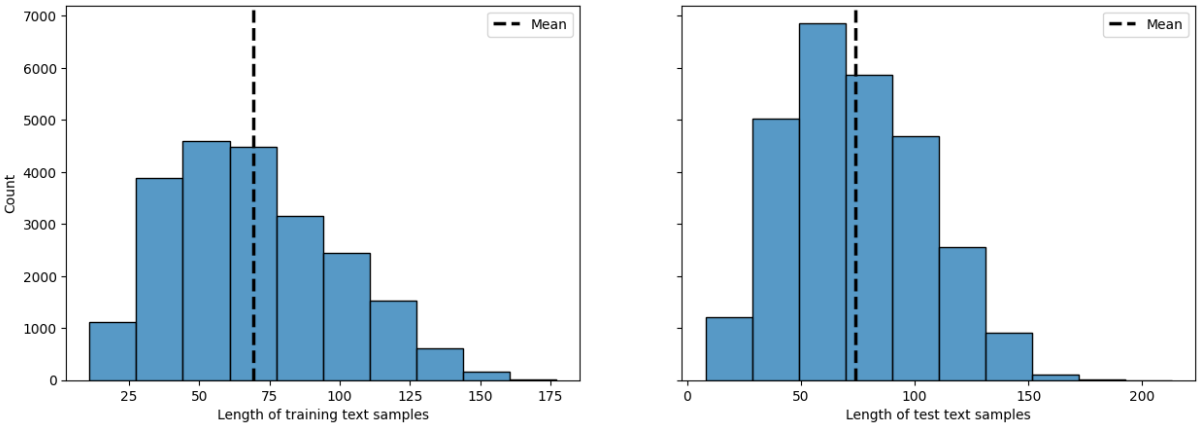

Figure 7.3: Length of Training Text and Test Text Samples

We also present a few samples from the dataset in the table 7.1.

Subset	Phoneme characteristic pairs	Count
Rangpur	(h,ʰ), (h, ɦ), (k, t), (t, n), (n, p)	5
Kishoreganj	(ʰ, h), (ɟ, z)	2
Narail	(h, ʰ), (h, ɦ), (k, t), (t, n), (n, p), (ɟ, z)	6
Chittagong	(c, ç), (cc, çç), (ɟ, z), (p, f), (pʰ, f), (pʰ, ff)	6
Narshingdi	(ʰ, h), (ɟ, z)	2

Table 7.1: Regional Phoneme Characteristics Pairs (Standard, Region)

By embarking on this statistical exploration, we have gained a deeper understanding of the DUAL-IPA dataset, uncovering its numerical facets and revealing its potential for enriching research and applications in Bangla NLP. The diverse vocabulary, sentence length distribution, phoneme landscape, and provided samples paint a vivid picture of this invaluable resource, laying the foundation for further analysis and innovation.

Chapter 8

Modelling and Benchmarking

8.1 Model Selection and Training

For this benchmarking exercise, we opted for a sequence-to-sequence (seq2seq) model due to its inherent capability to handle variable-length inputs and outputs, characteristics essential for tackling IPA transcription tasks. We specifically chose the MT5 model, a powerful multilingual pre-trained transformer architecture developed by Google [10]. The ”small” variant of the MT5 model was selected, offering a balance between computational efficiency and performance.

To leverage the multilingual capabilities of MT5, we utilized the model pre-trained on a massive dataset encompassing text and code from the Common Crawl project, covering a staggering 101 languages. This pre-training provided the model with a solid foundation for understanding linguistic structures and handling variations across languages, including Bangla.

In terms of training specifics, we employed a moderate training regime encompassing 10 epochs and a learning rate of 3e-4. This configuration proved to be effective in balancing learning speed with model stability, yielding optimal performance on our designated IPA transcription task.

In this chapter, we present a comprehensive benchmarking exercise to evaluate the performance of IPA transcription for Bengali utilizing our newly proposed Dual-IPA dataset. This evaluation serves as a crucial step in assessing the effectiveness of our approach and demonstrating its potential value for the advancement of NLP tasks in the Bangla language.

8.2 Benchmarking Dual-IPA Dataset

8.2.1 Evaluation Metric: Word Error Rate (WER)

To assess the model's performance objectively, we opted for the widely used Word Error Rate (WER) metric. WER calculates the number of errors, including substitutions, insertions, and deletions, between the predicted IPA sequence and the reference ground truth. This metric provides a sentence-level evaluation, offering a comprehensive understanding of the overall transcription accuracy achieved by the model.

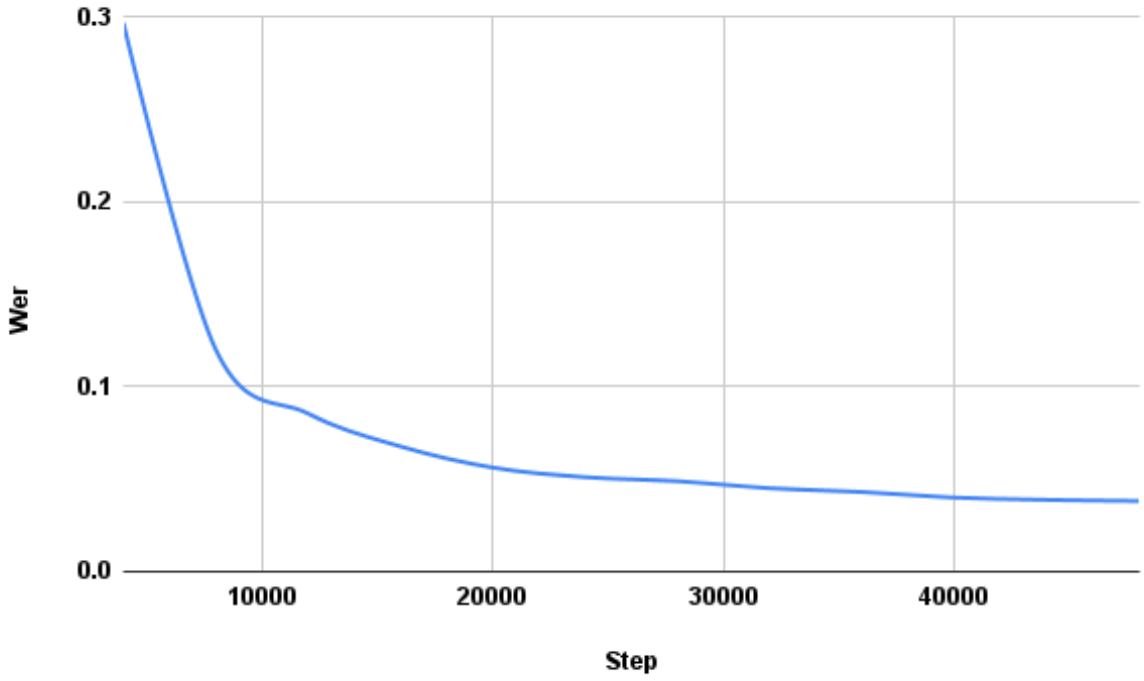

Figure 8.1: Step vs Wer

8.2.2 Benchmarking Results and Analysis

The conducted benchmarking yielded a highly encouraging WER of 0.1 on the held-out test dataset. This impressive result suggests that our chosen model, the small MT5 variant, effectively leverages the DUAL-IPA dataset to accurately transcribe Bengali text into its corresponding IPA representation.

While a detailed investigation into the exact causes of this success awaits further analysis, we hypothesize that several factors might be contributing to the model's strong performance.

Firstly, the Bangla language possesses a relatively smaller number of homographs compared to other languages. This characteristic reduces ambiguity faced by the model during transcription, as each word typically maps to a unique IPA sequence.

Secondly, the Dual-IPA dataset incorporates rich contextual information through the inclusion of both source and target languages. This additional information likely aids the model in handling out-of-vocabulary (OOV) instances more effectively, as the source language provides clues about the intended pronunciation even if the specific word has not been encountered during training.

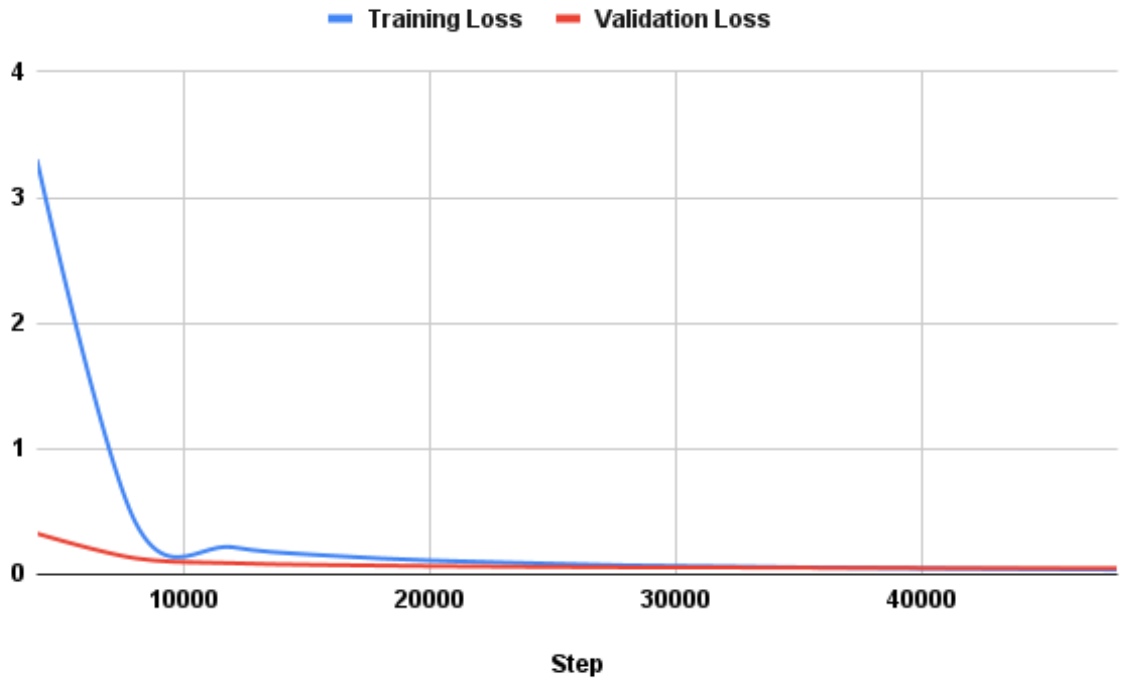

Figure 8.2: Step vs Training Loss Validation Loss

These preliminary findings showcase the immense potential of the Dual-IPA dataset and the chosen MT5 model for tackling the challenging task of Bengali IPA transcription. Further research exploring different model architectures, training regimens, and evaluation metrics would provide valuable insights into further optimizing performance and pushing the boundaries of this application.

Chapter 9

National Competition on DUAL-IPA

9.1 Overview of the Competition

Bengali Text to IPA (International Phonetic Alphabet) Transcription is an area that has seen relatively limited development compared to other languages, despite Bengali being one of the world's most widely spoken native languages. There is a growing need for automated systems that can accurately convert Bengali text into IPA notation due to the vast audience and various applications in linguistics, language learning, and phonetic research. Having this in mind, we welcome participants to participate in DataVerse, a part of ITVerse 2023 organized by IIT Software Engineers' Community (IITSEC) as we partner with Bengali.AI to advance research in Bengali text to IPA domain.

9.2 Competition Schedule

Initial Round: Was held on Kaggle from October 9th to November 1st. Participants are expected to build and train models during this phase. The top 15 registered teams from this round was invited to join us in the final round.
Final Round: Was held onsite at the Institute of Information Technology, University of Dhaka on 5th November. Only invited teams got a chance to present their work in front of the judge's panel, where they had to submit an IEEE/ACM (2 column) paper. Furthermore, their inference notebooks were evaluated on a hidden dataset.

9.3 NLP Wrokshop

A Workshop titled **Hands-on NLP Workshop: Bengali Transcription Modeling** was arranged in collaboration with Bengali.AI and the organizers where Asif Shahriyar(Coordinator, Bengali.AI) was the host and Dr. Ahmedul Kabir(Associate Professor, IIT, University of Dhaka) was a guest speaker.

Figure 9.1: Hands-on NLP Workshop: Bengali Transcription Modeling

9.4 Scoring

The scoring weights are divided in the following way -

Round 1 Public Standing : 12%

Round 1 Private Standing : 18%

Hidden Evaluation : 50%

Paper and Presentation : 20%

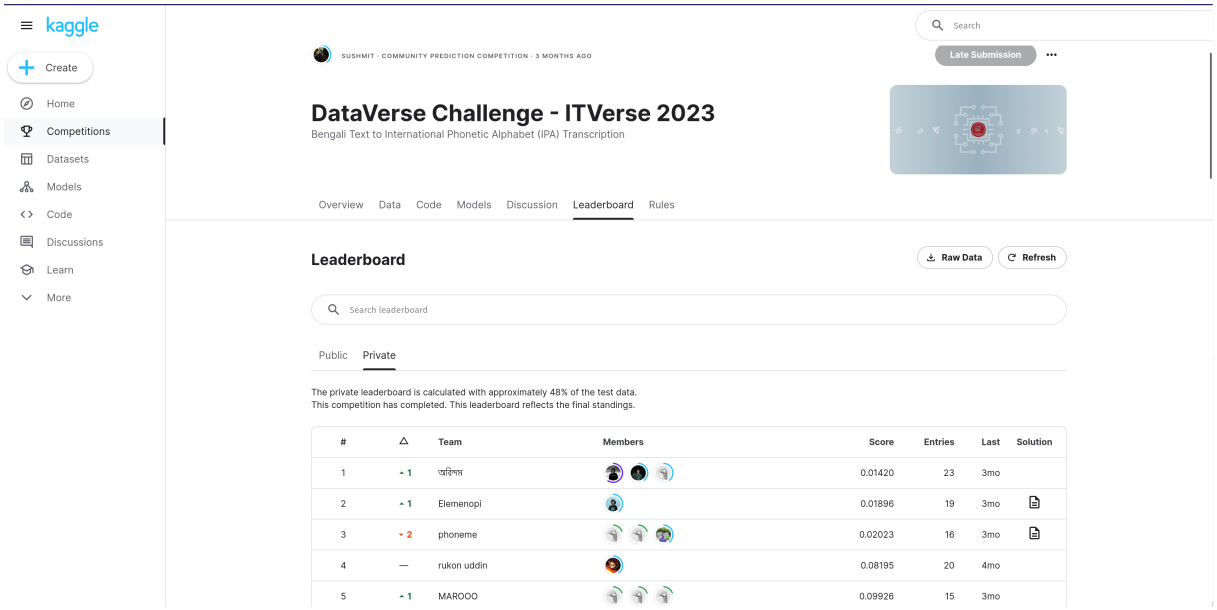

Figure 9.2: ITVersre 2023, DataVerese segment leaderboard from Kaggle

9.5 Goal of the Competition

The goal of this competition is to recognize model IPA transcription from Bengali texts(Remember the Greek characters in the dictionary, to help us find out the accurate pronunciation of words? That was International Phonetic Alphabet (IPA) transcription! They had to build models trained on a linguist validated dataset containing Bengali text from different domains. The test set contains numbers, loan-words and domain-specific words to add to the challenge.

Their efforts could improve Bengali computational linguistics and NLP research using the first Bengali sentence level IPA transcription dataset from Bengali.AI. In addition, Their submissions have been among the first open-source IPA transcription methods for Bengali.

9.6 Evaluation

Submissions were evaluated by a mean Word Error Rate, proceeding as follows:

- The WER is computed for each instance in the test set.
- The WERs are averaged within domains, weighted by the number of words in the sentence.
- The (unweighted) mean of the domain averages is the final score.

9.7 Competition Statistics

- **Total Participants:** 104
- **Teams:** 64
- **Final Model Submission:** 11

Figure 9.3: Glimpse from ITVersre 2023, DataVerese segment

Figure 9.4: Prize Giving Ceremony ITVersre 2023, DataVerese segment

Chapter 10

Future Plans and Perspectives

The creation of this dataset and initial model for Bengali IPA transcription marks a significant step forward in the field of Bengali NLP. However, it also opens up exciting avenues for future exploration and further development. Here, we outline some potential directions for expanding upon this work:

10.1 Expanding the Dataset

Size: Increasing the size and diversity of the dataset will enhance the model's generalizability and robustness. This could involve collecting data from different dialects, age groups, and speaking styles.

Speech Recordings: Incorporating speech recordings alongside IPA annotations would enable research on tasks like automatic speech recognition and text-to-speech synthesis for Bengali.

Multilingualism: Including translations of the Bengali text could facilitate cross-lingual research and applications.

10.2 Model Development

Advanced Architectures: Experimenting with more sophisticated model architectures like Transformer-based approaches could further improve performance on various NLP tasks.

Task-Specific Models: Developing specialized models for different tasks like machine translation, sentiment analysis, or summarization could offer practical applications.

Multimodal Learning: Integrating audio and text data within a single model could lead to advancements in speech-related NLP tasks.

10.3 Applications and Impact

Language Learning: The dataset and model could be used to develop language learning tools and resources for Bengali.

Accessibility: The technology could be adapted to create speech-to-text applications for individuals with disabilities.

Cultural Preservation: The project can contribute to the preservation and documentation of Bengali dialects and regional variations.

10.4 Collaboration and Community Building

Sharing the Dataset: Making the dataset publicly available would encourage further research and collaboration within the NLP community.

Open-Sourcing the Model: Open-sourcing the model would allow others to build upon it and create new applications.

Community Engagement: Building a community around Bengali NLP could attract researchers, developers, and enthusiasts to contribute to the field's growth.

10.5 LREC Coling - 2024

In addition to the aforementioned future directions for research, I am thrilled to announce that this work has been submitted for presentation at the prestigious **LREC-COLING 2024 – The 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation Lingotto Conference**, scheduled for 20-25 May 2024 in Torino (Italia). This esteemed international forum provides an invaluable platform to engage with leading researchers and practitioners in computational linguistics, language resources, and evaluation. Presenting our findings at LREC-COLING 2024 offers an exciting opportunity to gather expert feedback, foster further collaboration, and contribute to the advancement of the field. We anticipate fruitful discussions and a chance to learn from the diverse perspectives of the conference attendees, ultimately enriching our research and propelling it towards even greater impact.

Chapter 11

Conclusion

This thesis transcends the mere examination of Bangla IPA complexities. It delves into the language's nuances, leaving an indelible mark on both linguistic theory and practical NLP applications.

Our journey began with a thorough exploration of existing literature, dissecting the intricate web of conflicting viewpoints on Bangla IPA. This analysis meticulously constructed a robust and comprehensive IPA transcription framework specifically tailored for Bangla texts. This framework not only addresses previously identified issues but also sheds light on the previously obscured subtleties of the language.

Our contribution extends far beyond theoretical constructs. Recognizing the dearth of large-scale resources for Bangla NLP, we built a groundbreaking dataset consisting of 150,000 sentences. This monumental undertaking marks the first of its kind, offering an invaluable tool for future research and development. Its impact transcends linguistics, holding immense potential to influence and enrich the field of NLP dataset creation.

Furthermore, this dataset paves the way for advancements in Language Model (LLM) downstream tasks. By providing LLM systems with rich, nuanced Bangla language data, we unlock unparalleled opportunities for future exploration and innovation. This opens doors to an array of potential applications, propelling research forward and expanding the boundaries of language technology.

In conclusion, this thesis is not merely a culmination of efforts; it's a springboard for the future. The developed framework and pioneering dataset stand as testaments to the profound impact our work can have on linguistics and NLP. Looking ahead, we anticipate witnessing this research blossom into practical applications that enrich lives and redefine our understanding of language processing. This is not the end, but rather the beginning of a new era for Bangla language technology, and we are proud to have played a pivotal role in shaping its trajectory.

Bibliography

[1] I. P. Association, *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*. Cambridge University Press, 1999.

[2] S. K. Chatterji, "Bengali phonetics," *Bulletin of the School of Oriental and African Studies*, vol. 2, no. 1, pp. 1–25, 1921.

[3] D. Jones, *An outline of English phonetics*. BG Teubner, 1922.

[4] A. K. M. Morshed, *Adhunik Bhashatatwa*, 2nd ed. Noya Udyog, 1997.

[5] Z. I. Ali, "Dhanibijnaner bhumika (introduction to linguistics)," 2001.

[6] A. Hai, *Dhwonibijnan O Bangla Dhwonitottwo*, 3rd ed. Bornomichil, 1964.

[7] D. Huq, "Bhasha bigganer katha (facts about linguistics)," *Dhaka□ Mowla Brothers*, 2002.

[8] S. Sen, *Bhasar Itibritta*. Ananda Publishers Private Limited, 1993.

[9] C. A. Ferguson and M. Chowdhury, "The phonemes of bengali," *Language*, vol. 36, no. 1, pp. 22–59, 1960.

[10] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, and C. Raffel, "mt5: A massively multilingual pre-trained text-to-text transformer," *arXiv preprint arXiv:2010.11934*, 2020.