Dhaka University Institutional Repository

Bangla Language Processing: Trends, Challenges and Some Propositions

By

Most. Rumana Aktar
Registration No: 22
M.Phil. Session: 2013-2014

This Thesis submitted to the Department of Linguistics at the University of Dhaka in partial fulfillment of the award of M.Phil. in Linguistics.

Department of Linguistics
University of Dhaka

February 25, 2021

Dhaka University Institutional Repository

Declaration of the Examinee

I declare that all materials presented in this thesis paper are my work, thoroughly and specifically acknowledged wherever adapted from other sources. It has not been published or submitted for any other degree

……
Most. Rumana Aktar
Researcher
Registration No.: 22
M.Phil. Session 2013-2014
Department of Linguistics
University of Dhaka

Dhaka University Institutional Repository

Attestation by Thesis Supervisor

This examinee is a regular student of the Department of Linguistics, M.Phil. Session 2013-2014, University of Dhaka. The Thesis is conducted in fulfillment of the requirements of the M.Phil. (Thesis paper) under my direct supervision.

As far as I know, this thesis paper, as documented, has not been published in full or in part or submitted for any other degree.

Date:……

Supervisor:

…….……

Sikder Monoare Murshed
Professor, Department of Linguistics
University of Dhaka

Dhaka University Institutional Repository

Acknowledgment

First of all, my sincere gratitude goes to my academic supervisor, who counseled and guided me throughout to finish this work. His deliberation and advice will always be remembered. Special thanks go to Professor Daniul Huq, Professor Dr. Abdullah Al Mumin, Associate Professor Nabeel Mohammad, Associate Professor Dr. Tarik Manzoor; Associate Professor Dr. Mamun Or Rashid. They helped me with data collection and also advised me in different aspects.

I want to give my appreciation to my sister Shumi who assisted me in this regard. Their encouragement and support did accompany me through my four years of study.

Dhaka University Institutional Repository

Abstract

Bangla language processing (BLP) in Bangladesh is still at the amateur level. Computerization of Bangla Language is a must to forward and shape our country as a digital Bangladesh. Success up to the demand of users and experts in this field will open the opportunity to exchange our knowledge among many languages. In this paper, we tried to present Bangla Language processing's overall situations in terms of technical and linguistic challenges. Also, we proposed some recommendations to overcome significant barriers while developing tools of BLP. We collected data from five groups, including technologist, linguist, researcher, developer, and user. We tested some available software developed Android version and windows platform and found many anomalies or bugs that need to be fixed right away. Though Bangla is a complex language, Chinese, Arabic, and Japanese, it is far better in Natural Language Processing. The significant and big problem is not developing a well representative corpus of the Bangla Language. Thus, researchers cannot go further in Artificial intelligence. Moreover, lack of collaboration is another major issue that is not to develop the BLP tools successfully or solve compatibility issues in different platforms including, android, windows, Linux, and IOS. We create a well-representative Bangla corpus including speech and written to make user-friendly BLP tools. It is excellent news that for the first time, our government budgeted around Tk 160 crore to develop 16 different components of Bangla Language Processing tools. To step forward to a middle-income country, we must invest more in this sector.

Dhaka University Institutional Repository

Index

Dhaka University Institutional Repository

Dhaka University Institutional Repository

Dhaka University Institutional Repository

Dhaka University Institutional Repository

The lists of Figures, Diagrams and Tables

Dhaka University Institutional Repository

LIST OF ACRONYMS

NLP Natural Language Processing
BLP Bangla language Processing
AI Artificial Intelligence
OCR Optical Character Recognition
MT Machine Translation
T2S Text to Speech
S2T Speech to Text
POS Tagging Parts of Speech Tagging
BCC Bangladesh Computer Council
CRBLP Centre for Research on Bangla Language Processing
ICBLP International Conference on Bangla Language Processing

Dhaka University Institutional Repository

CHAPTER 1

INTRODUCTION

1.1. Background

Over 300 million people speak Bangla over the world (সিকদার, ২০১৪). Natural Language Processing (NLP) started in the 1950s when Alun Turing published an article named "Computing Machinery and Intelligence". Now, it is called the 'Turning test". All-natural Language Processing (NLP) is a subfield of both intelligence that is artificial and linguistics. There are two types of NLP. One is in the human brain, and the other is in the computer. NLP helps to computerized spoken form and written form of Human Language and give output like a human being. That is why Bangla Language Processing (BLP) is a must for using this language effectively in computers, and IT platforms, for that matter.

Nowadays, one can open a hidden file through software like voice recognition, and information in the English language can be retrieved in many other languages. Nonetheless, the English, Arabic, Chinese, Japanese, Korean, and other languages have been computerized up to a satisfactory level. In Bangladesh, we do not have a compatible keyboard Interface, and layout let alone made tools like a Machine Translator. The debate is going among experts & users' levels regarding the keyboard layout. Bangla is one of the modern language levels, including complex script, secured the 5th ranking amongst the world languages (হক, দানীউল ও সরকার, পবিত্র, ২০০৩). It should be computerized according to users' demands. We have to use English as an international language, but we should not compare BLP software with English. For example, someone might say that the English language's typing speed is far better than Bangla language typing. But we should keep in mind that the English typeface is linear, whereas Bangla is involved, including almost 300 ligatures. However, if we focus on the applied side, some limitations and challenges still need to be minimized. As we live in the era of technology, everybody expects to do any task faster than before. We appreciate that one Bangladesh government funding project (2016-2021) is going on at Bangladesh Computer Council, where 16 components have been selected to develop tools. By conversion, a mammoth task is necessary to reach a satisfactory level. We did not find any paper on the overall situation on

Dhaka University Institutional Repository

Bangla Language Processing in Bangladesh, including the problems of not interested much in this sector, technical challenges, linguistic challenges, the mentality or concept among experts, developers, researchers, policy-makers, and last, but not the least user. We had some papers where some anomalies are present on particular tools. In this paper, we tried to present an overall Bangla Language Processing scenario in Bangladesh, including challenges with some propositions.

The majority of the population lives in South Asia's eastern flank that surrounds the Bay of Bengal. They are geographically distributed as follows: over 95% of those living in Bangladesh, and from amongst the states that are Indian, 26% of those in Andaman and the Nicobar Islands, 28% of those in Assam, 67% of those in Tripura, and 85% of those in West Bengal. The user of the Bangla language is increasing day by day in many other countries. It is not necessarily as highly ranked in terms of the most read, the most wired, or the most archived, or the most used on the Internet, although it is the seventh most spoken language globally. The potential ICT benefit continues to elude many of the Bangla-speaking population who are not equipped with either English or the language of their diaspora despite significant progress in Information and Communication Technology (ICT) and the availability of a vast, enriched English knowledge database around the globe. This is complicated by the fact that Bangla is not without its peculiar nuances and issues that are structural. It has a relatively large alphabet that includes many compound letters, two acceptable forms with varying pronouns, verb conjugations, many regional dialects, and variant spellings for too many of its words.Additionally, at least two non-standard dialects of Bangla – Chittagonian, and Sylheti, respectively, with 47% and 30% lexical dissimilarity. Bangla Language Processing (BLP) is an evolving discipline that is aware of these language realities and seeks to create a robust digital platform for use by many of the Bangla-speaking population bypassed currently by the ICT revolution. The success of BLP is envisioned to have an effect that is good for many of the typical individuals and their socio-economic life (Karim et al., 2013).

1.2. Concern Area:

Dhaka University Institutional Repository

Our main concerned area is Bangla Language Processing tools in Bangladesh. We had to analyze the expert body's opinion, ongoing and completed research projects, the users' expectation or demand to focus the BLP tools. We try to find out the challenges in developing BLP tools and present some propositions to overcome them. We elaborately discussed the following BLP tools:

- Optical Character Recognition (OCR)
- Bangla Spell Checker
- Bangla Keyboard Interface and Layout
- Bangla Speech to Text Converter
- Bangla Text to Speech Converter
- Machine Translation (MT)
- Fonts
- Bangla Corpus Development
- Bijoy to Unicode, and vice-versa converter
- IPA Transcription Software
- Sentiment Analysis

1.3. Statement of the Problem:

Bangla is a wide-known language among other languages (almost 3,500 plus languages all over the world). Over 300 million people use the Bangla language worldwide, including Bangladesh and West Bengal of India. Day by day, the demand for learning the Bangla language is increasing. We are living in an era of technology. However, till now, we have no essential software like Bangla Spelling or Bangla Pronunciation Dictionary. We see the availability of the tools of other languages like English, Arabic, Chinese, etc. We see that many kinds of research on BLP tools are being conducted in Bangladesh and West Bengal. Some researchers are working on BLP tools abroad as well. We do not know the overall situation of Bangla Language processing in Bangladesh as there is no gathered knowledge. What kind of works are being conducted in this area, and tools available in the apps market or Windows or Linux platforms? Collaboration is a significant factor, especially in Bangla language processing, as the field demands two experts' bodies, including programmer and linguist. Most of the young generation uses NLP tools like google translation. We need to know what they are thinking about the Bangla Language processing tools like

Dhaka University Institutional Repository

Bangla fonts, keyboard interface, TTS (Text to Speech), STT (Speech to Text) software, Bangla OCR, Bangla Spell-checker, and many more.

1.4. Objectives of the Study:

Our primary concern is analyzing the trend of BLP in Bangladesh and challenges, including linguistic and technological, for developing tools and end with some propositions to step forward BLP. By observing these, we found some recommendations to overcome the obstacles. We focused on how much work has been done in Bangla language processing, what kind of software is available or under development, limitations of some software, finding the research gaps, reasons behind lack of collaborations among government, public, and private organizations, reasons behind lack of motivation to build a career in the field of BLP, what are the user's demand regarding Bangla software. Moreover, we also focused on the policymaker's planning and dedicated researcher in this particular field.

1.5. Justification of the Study:

This study presented the overall scenario of Bangla Language Processing in Bangladesh. It also revealed software tools already developed or under developed, the limitations of the developed tools, experts' opinions, challenges in developing tools, and solutions to these obstacles. The researcher quickly finds the overall research gap and the trend of Bangla Language Processing. We came to know the user's feedback. The researcher might interest in collaboration. More funding opportunities will develop if this study takes the policymaker's attention or universities' higher authority. Not only researchers but also students will be motivated in this field. The developer can update their tools according to users' feedback.

1.6. Research Question:

a) Main research Question: Our main research question is as follows:

What developments have been made in the field of Bangla language processing tools in Bangladesh?

This main query will motivate us to know the answer to the following supplementary questions:

Dhaka University Institutional Repository

b) Supplementary Questions:

I. What is the recent development of BLP tools in Bangladesh?

II. What are the limitations of developing the BLP tools in Bangladesh?

III. What measures to be taken for the further development of BLP tools in Bangladesh?

IV. What are the challenges in developing BLP tools?

IV.I. Are there any linguistic challenges?

IV.II. Are there any technical challenges?

V. Which sectors are involving in developing BLP tools?

V.I. Which software companies are actively working on BLP?

V.II. Is there any collaboration in developing BPL tools?

1.7. Methodology:

There are three types of Research methodologies like quantitative, qualitative, and mixed methods. This study has been conducted through the qualitative approach. The interview method, including the semi-structured question, has been applied for collecting data. Details of methodology have been presented in chapter 4.

1.8. Organization of the Study:

This paper is organized into a total of 8 chapters. The included chapters have been present in the following order:

Chapter 1: Introduction

Chapter 2: Theoretical Background

Chapter 3: Theoretical Discussion

Chapter 4: Methodology

Chapter 5: Data Presentation

Chapter 6: Result Analysis

Chapter 7. Propositions: Measures to be taken to overcome the challenges of Bangla Language Processing.

Chapter 8: Conclusion

Dhaka University Institutional Repository

CHAPTER 2
THEORETICAL BACKGROUND

In the 1980s, we see the first attempt was font development and Bangla's use in the Windows platform's computer. Especially software companies or personally dedicated persons like Shahid Munir Chowdhury, Saif-ud-Doha, Mustafa Jabbar (Honorable minister, Postal, Telecommunication, and ICT Ministry.), and others afterward. Bangla Language through 'Shahid Lipi' was first inserted in 1986. It was a breakthrough. The introduction of 'Bijoy' Bangla software made an excellent platform to write Bangla through the computer. The main problem at that period was the compatibility issue of the Bangla language in different media. Bangla was not usable as a general language on every system as there was no unique way to represent Bangla. In the late 1990s, Unicode shed new light on the issue, and the processing of Bangla computing began to take a new shape in the country. The open-source platform impacted Bangla Language Processing. Linux version in Bangla was introduced by J. Ahmed, a software developer, in 1988.Ankur, a software developer group, introduced Bangla Open Source Software like Linux, OpenOffice.org, Gaim, etc., in the late 1990s. Like Ekushey (ekushey.org), a voluntary organization worked on open-source Unicode fonts and Bangla input system. They determined how the Bangla fonts work under the existing keyboard interface.

From the government side in 2014, the BCC (Bangladesh Computer Council) proposed mapping for a national Bangla keyboard and a collation sequence; CRBLP at BRAC University also started conducting research projects on Bangla language processing, the researchers of CRBLP worked on Bangla information retrieval (i.e., Bangla spell-checking, and a Bangla search engine, etc.), morphological analysis, developing a digital lexicon, and an online dictionary, optical character recognition, and speech processing. The Bangladesh Open Source Network (BdOSN: bdosn.org) was launched with local open source volunteers in 2005. They considered Bangla Language processing as one of its main issues. That is why open-source in Bangla got successful (Islam, 2009). Recently, Apurba-DIU R&D Lab has launched the Bangla Character Annotated Project on the 10th of September, 2020, virtually in Daffodil

International University. The vendor software company is already working on a government project, which will end 2021 or extend, managed by Bangladesh Computer Council. Another software team, Rib, worked on this project. Besides, Therap, TigerIT, and other unknown groups are working on BLP in Bangladesh, but the main problem was the compatibility issue. Whether the Bangla can be seen in every platform like Windows XP, 2000, etc. Different operating systems like android, IOS, Windows, Linux do support fonts or not.

2.1 Bangla Language Processing Tools:

There are many Bangla Language Processing tools in Bangladesh as the android and windows or Linux platforms. We focused on the mechanisms of Bangla Language processing: Optical Character Recognition (OCR), Bangla Spell Checker, Bangla Keyboard Interface, Fonts, Machine Translation (MT), Text to Speech Converter (T2S), Speech to Text Converter(S2T), Bangla Search Engine, Bangla Corpus Development, IPA Transcription, Sentiment Analysis and Bangla POS tagging.

2.1.1. Bangladesh Computer Council:

Bangladesh Computer Council is working on 16 components of Bangla LanguageProcessing tools under the government's language project. The duration of the project is June 2016-June to 2021. The components are:

1. Development of complete Bangla Corpus following international standards
2. Further development of Bangla OCR (Bongo OCR) developed by ICDT and integrating handwriting recognition system
3. Development of Bangla speech to text & text to speech software
4. Improvement of the National Keyboard (Bangla)
5. Development of Bangla style guide
6. Development of the Bangla for interoperability engine
7. Development of Bangla CLDR (Unicode Common Locale Data Repository) resource, and submit to Unicode
8. Development of Bangla Spell & Grammar checker
9. Development of the Bangla Machine Translator
10. Development of Screen Reader Software
11. Development of software for disabled people

Dhaka University Institutional Repository

12. Development of sentiment analysis software in Bangla
13. Developing a service platform combining language processing with building processing pipelines for value-adding tasks in the multilingual content processing
14. Translation of most popular/used sites into an international language
15. Standard Keyboard for Tribal Languages (Other languages of Bangladesh)
16. Incorporating Bengali IPA fonts and software to world language linguistic list

2.1.1.1. Optical Character Recognition (OCR):

Once upon a time, OCR was available for the English language only. But many developers, including the Bangladesh computer council, tried and still trying to develop a user-friendly Bangla OCR. Now we will discuss some recent developments of Bangla OCR.

2.1.1.1.1. Bangla Optical Character Recognition (Bangla OCR):

Bangladesh Computer Council (BCC) developed an OCR named 'Bongo-OCR' earlier, but it was only for an online platform. They are now developing a Bangla OCR called Bangla Optical Character & Handwriting Recognition System (Version: 1.0). It supports both ASCII and Unicode. It will run in windows and Linux(Online and Offline). With the help of an optimized model, it will also run on an android phone.

Main features of the OCR developing by BCC:

- It will cover PDF/JPEG/PNG, and images
- Scanned documents, including photos, tables, numbers, bullets, some Arabic characters, Bangla simple to complicated orthographic name, consonant cluster, conjuncts, andsymbols of various forms continuous text (e.g., rotation angle minimum 30 degrees) and necessary typological features.
- It will detect paragraph, tables, photo caption, and the resolution will be 150-600 dpi
- It will support continuous and isolated handwriting written by the different scribe.

- The algorithm and Methods: Hence, the deep learning plus supervised machine learning (hybrid) method has been used. It will analyze both isolated and continuous characters: both a character and word-based detection will be enabled.
- Fonts support the most popular 20 fonts, including letterpress and typewriter font.
- This OCR will allow both online and offline texts. It supports dynamic hand-writing and touch-screen enabled. The off-line handwriting recognition module must work with scanned and captured handwriting recognition modules that need to be real-time and workable with live writing on touch-responsive digital screens by finger/stylus/digital pen or similar input tools. Cover all entities, including the following features with a gold standard dataset:
- Primary characters and consonant modifier
- The primary character and vowel modifier
- Compound characters
- Compound characters and consonant modifier
- Compound character and vowel modifier Some corrections made in BCC's OCR:
- A typical situation in the neural network method is being solved using context analysis, collocation, and occurrence frequency.
- Another problem, the number of black pixels in each column is calculated to construct a column histogram. When no black pixel is found in the vertical scan,it is considered the space between two words. In this way, one can separate the words. , but the problem occurs when the 'matraless' character is used in a word. The solution is that the system will handle this situation from context analysis and occurrence frequency.

After Md. Sanzidul Islam et al., 2019, LSTM (Long short-term memory) Recurrent Neural Network (RNN) generates sequence to sequence Bangla sentence. , but their dataset size was not very large because of hardware limitations. They used 917 days daily newspaper named online ProthomAlo. Per-day word counted around 15,5000. Afterward, other words have been used as a dataset, successfully generating Bangla

sentences, but the dataset was still limited. A paper was named ‘an efficient way to segmentation Bangla characters in printed documents using curved scanning.’This paper claimed that the curved scanning method used in the OCR software could solve a problem like:

- Different shapes of characters
- Connected characters
- Modifiers in the top and bottom
- The overlapped region between consecutive characters

And they claim that the proposed strategy has achieved a 99.8% accuracy rate in line segmentation, 99.5% accuracy rate in word segmentation, and 99% accuracy rate in character recognitions

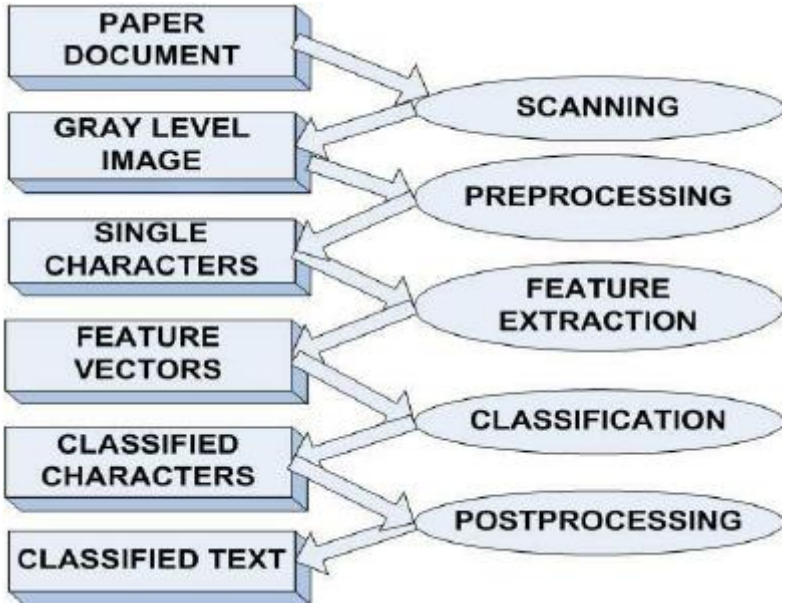

Figure 1: Basic Steps of an OCR

(Source:https://www.researchgate.net/publication/222109100_A_Complete_Workflow_for_Development_of_Bangla_OCR/figures?lo=1)

2.2. Computer Keyboard Layout:

QWERTY’ Layout had moved from typewriter to computer keyboard in 1874. For the first time, Herman Hollerith developed keypunch devices. By the 1930s, keys have been included for typing texts and numbers. There are various types of Keyboard, including desktop or full-size, laptop-size, Flexible Keyboard, hand held, Thumb size,

Dhaka University Institutional Repository

multifunctional, etc. Three types of mechanical layouts are available- ISO, ANSI, and JIS made worldwide, the United States, and Japanese standards, respectively. Total keys are 105. By conversion, 104 keys are available in US Conventions.

2.2.1. Bangla fonts, Keyboard Interfaceand Layouts Development:

After Rifat Hassan Jihan, in the 21st century, we need a keyboard driver for background processing, and the font is significant for displaying the character. Munir Chowdhury took the first attempt to create a Bangla layout for a typewriter's QWERTY keyboard. At that time, Bangla language was documented in the ASCII platform and machine dependable.

There are two types of fonts available in Bangladesh- ASCII format and Unicode supported. For example,- popular ASCII format Bangla font is sutonnyMJ, and Vrinda. By conversion, Unicode supported fonts like Vrinda, ShiyamRupali, Nikosh, Kalpurush, Unibijoy, etc. At the change associated with century, aided by the arrival of modern-day Bangla Sriramapur Mission Press and its foundry (steel casting for movable types), took a task leading printing Bangla publications of different topics and disciplines. The Bengali department regarding the Fort William College (est. 1800 to teach the English civilians in neighborhood languages) fortified that trend while the School Book Society augmented Bangla printing circulation. The character is a celebrated scholar of Bengali P, anditIswar Chandra Vidyasagar happened to be Bangla Language Planning's historical designer. While besides he authored several books, including grammar along with other topics in 1841 he became the pinnacle Pandit associated with the Bengali and Sanskrit division of the Fort William College not just he composed the essential Bangla mastering book `Borno Poricoy' বর্ণপরিচয়. He introduced four letters being new connivance to Bangla orthography specifically: ড়, ঢ়, য়, ৎ. His contribution, which is certainly biggest into Bangla publishing's growth, ended up being which he took all the troubles (even torments) to work at the foundry to design and create needed Bangla movable types. Thus, it came into being theVidyasagar that is popular" for hand composing in old-fashioned letterpress publishing Bangla publications, and so forth. This authorized of the composing that is self-disciplined Bangla printing. Since said above, type creating foundry needed to

deal with the number that is huge for movable kinds to cater to Bangla printing. In 1984 ‘BangkimLipi was created by Goutam Shen and ShatiBandapaddhaya under Rahul Commerce in West Bengal. Jadavpur University in 1984 utilized this lipi by having a user interface showing on the screen. It had been successful but didn't make copies, which can be challenging. In 1986, Saif-Ud Doha Shahid created a lipi called ‘Shahid lipi.’ It was perhaps not supported in MS-word, however, in Mackintosh. Later on, it was found in a nationwide hit institute. In 1987, Mainul Hassan developed a lipi known as ‘MainulLipi.’ Bangla newspaper was initially published in Mainul fonts. It was fundamentally an updated type of Bangkim script relating to first and level. This is undoubtedly the second of Keyboard. In 1988, Ananda Computer created ‘Ananda Lipi’; afterward, Bijoy 2000 was developed by Mustafa Jabbar. It became the Bangla that is well-known Typing because of its better compatibility issue. However, one needed to put in this pc software for decoding the written text typed Bijoy that is using 2000. Bijoy 2000 keyboard introduced a design this is undoubtedly brand-new. Avro may be the first Unicode supported Keyboard manufactured by a genius student of Mymensingh university called medical Hassan Khan in 2003. It is undoubtedly no-cost. Open-source software, having graphical keyboard layout changer supporting Windows 2000, Vista, XP, Windows 7, Linux, and Ubuntu. Among the salient features of Avro that provided large popularities amongst lay/lame users had been its phonetic composting system with a standard Keyboard that is English. In place of ASCII, Avro adapted the Unicode that managed to make it a source computer software. This is undoubtedly open-source (Aktar, 2014).

There is an option to create a new keyboard layout in the Avro keyboard as well. Unicode, a universal character encoding standard, was used to overcome ASCII's problem over the internet. In 1987, it was introduced, and it used a unique symbol to represent any sound of any language. So, many developers developed a keyboard interface based on Unicode fonts. In the 19^{th} century, individuals, and developers, tried to build typewriters. ‘Sholes, and Glidden typewriter was the first commercially successful keyboard layout among many others. It was based on ‘QWERTY.’Afterward, according to IBM Electric typewriter, a computer keyboard layout has been developed. We see 104 or 105 keys in a standard keyboard layout. A

Dhaka University Institutional Repository

variety of keyboard layouts, including unique models, sizes, and functions, are available in the market. The layout means the pattern of keys in a keyboard. It was used in typewriters before used on the computer. Some designs were chosen to use in computers from those typewriters. Total 73 characters, including vowels, and digits, are used in a minimal Bangla keyboard layout. The primary characters used in Keyboard are as follows:

Vowel letters: অ আ ই ঈ উ ঊ ঋ এ ঐ ও ঔ

Consonant letters: ক খ গ ঘ ঙচ ছ জ ঝ ঞ ট ঠ ড ধ ণত থ দ ধ নপ ফ ব ভ ম য র ল শ ষ স হ ড় ঢ় য় ৎ ং ঃ ঁ

Signs, and digits: া ি ী ু ূ ৃ ে ৈ ো ৌ; ১ ২ ৩ ৪ ৫ ৬ ৭ ৮ ৯ ০৳

Six Phalas also can be seen in recent keyboards like ‘য’,‘র’,‘ল’,‘ম’,‘ন’, and ‘ল’ Some uncommon symbols like- ৯ ◌় ২ ◌ৃ ৌ ৯ ◌্ ◌্ ১ ১ ৯ ৯ । ৭ ০ ◌ঁ– are also available in Keyboard Layout. As ‘QWERTY’ layout is suitable so, all developers developed Bang keyboard layouts based on it. More characters in Bangla Characters are key mapped under regular layers, shift layers, AltGr layer, and Shift AltGr lasyer (Hassan, 2020).

Some significant features of the Bangla language are conjuncts. Almost 300 conjuncts are available in Bangla. Conjuncts may construct like:

ক্+ক=ক্ক **(**2 letter conjunct**)**

ক্+ষ্+ম=ক্ষ্ম (3 letter conjunct)

ন্+দ্+র-ফলা+য-ফলা=ন্দ্র্য (4 letter conjuncts)

Some popular Bangla keyboard layouts are as follows:

1. Shahid Lipi 2. Inscript Keyboard 3. Jatiyo keyboard 4. Probhat 5. Ekushey keyboard 6. Munier keyboard 7. Bijoy 8. Uni Gitanjali 9. Rupali keyboard 10. UniBijoy keyboard 11. Ridmik keyboard 12. Mayabi keyboard 13. Bangla word keyboard 14. Baishakhi keyboard 15. Avro keyboard

2.2.1.1. Manner of using Keyboard Layouts for Windows/Linux/, Android/IOS:

As we know that there are three kinds of keyboard interfaces available. So, the layouts are being used differently. The interfaces are Bijoy Classic, which does not support Unicode, Avro keyboard based on Unicode, and phonetically types to some extent, Unicode supported Keyboard is UniBijoy. Ekattor, Bangla Ankur, Shaon Bengali, Ekushe, etc., keyboards are also based on Unicode. There are two styles for typing Bangla using keyboard layouts in a standard keyboard. For example,- If we want to write 'খেলা' in oldest type keyboard, then we have to insert 'ে'first then for 'খ' next 'ল', and finally 'া'. Bijoy, Munier, Lekhoni, Gitanjali, Satyajit, etc., are such types of layouts. No keyboard is based phonetically entirely, yet some keyboards are known as phonetically typed Keyboard. Hence, 'খেলা' Unicode layouts (sometimes called Bangla Phonetic Based Layouts)The dependent vowels are ordered after the consonants it modifies, for example, should be typed like- first type 'খ' then 'ে' next 'ল' and finally 'া'. Such layouts are Unibijoy (modern style typing), Bangla Unicode, national, Munir Unicode, Unijoy, Baishakhi, etc. On the other hand, the English to Bangla transliteration scheme is used in transliteration-based layouts. Hence, It follows the English 'QWERT' format, if we want to write 'খেলা' we have to type first type 'kh' 'খ' then' for 'ে' next 'l' for 'ল'', and finally 'a' for 'া' ; That means we have to type 'khela'. Such types of layouts are Avro phonetic, Kickeys (Selim & Ismail, 2014).

2.2.1.2. Keyboard Interface for Android Version:

There are many Unicode supported keyboard layouts in the apps platform. For example, Ridmik, Mayabi, Bijoy Bayanno, Avro, Provat, Google Indic keyboard, G-board. Some of them are based phonetically, and others are developed traditionally.

Dhaka University Institutional Repository

2.2.1.2.1. Some Popular Bangla Keyboard Layouts:

In 1985, the Bangla language was first written on a Macintosh computer. We see that 180 keys were mapped under four layers. Later, developers from two Bangla successfully wrote Bangla Language both in Mackintosh and Windows.There are many types of keyboard layouts available for the Bangla language. “Most of the pc software firms design, develop Bengali Keyboard, and computer software this is undoubtedly associated commercial purpose. Some claim copyright throughout the specific layout and prohibit other individuals from distributing that layout, and pc software support is indeed associated. All the designs aren't medical or not designed following the personality, therefore typing the frequency. There is an inclination to keep unaspirated (অল্পপ্রাণ), and aspirated (মহাপ্রাণ) phoneme under one secret, such ক /k/ within the layer that is normal খ /kh/ into the shift layer (Hassan, 2020)

Some popular layouts have been presented below-

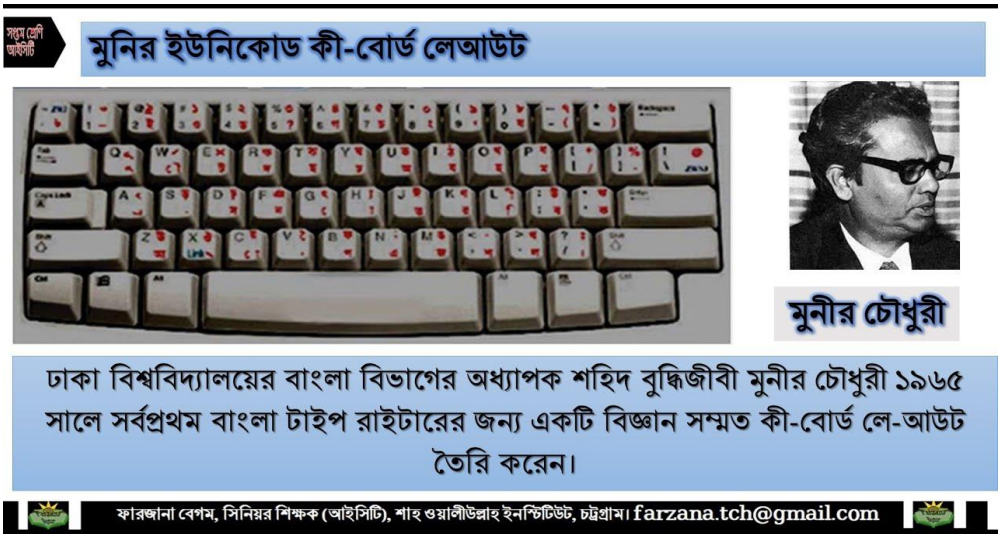

Figure 2:Munir keyboard layout

(source: https://i.ytimg.com/vi/k1squGjesUs/maxresdefault.jpg)

Figure 3: Shahid lipi keyboard layout (source: https://upload.wikimedia.org/wikipedia/commons/thumb/e/e7/KB-Bengali-Shahidlipi.svg/500px-KB-Bengali-Shahidlipi.svg.png)

Keyboard Layout

1st Layer)

(Default)

Esc	F1	F2	F3	F4	F5	F6	F7	F8	F9	10	F11	2	

	১	২	৩	৪	৫	৬	৭	৮	৯	০			Back space
	09E7	0	09	0	0	0	0	0	0	0			

Tab	ঙ	য	ড	ট	ঠ	চ	জ	হ	গ	ড়			
	0999	09AF	09A1	09AA	099F	099A	099C	09B9	0997	09DC			

Caps Lock	ৃ	ূ	ি	ব	্	া	ক	ত	দ			Enter
	09C3	09C1	09BF	09AC	09CD	09BE	0995	09A4	09A6			

Shift	ৎ	ো	ে	র	ন	স	ম					
	09B1	09	0	0	0	0	09					

Ctrl	Windows Key	Alt	Space Bar	Alt GR	Windows Key	Right button key

Figure 4: 1st keyboard layout of national Keyboard (National Keyboard; 2006, p-5)

Figure 5: Bijoy keyboard layout(source:https://www.wipo.int/export/sites/www/ipadvantage/images/article_0101_2_845.jpg)

Figure 6: Unibijoy keyboard layout (source:https://priozone.com/wp-content/uploads/2019/12/UniBijoy-Layout.jpg)

Figure 7: Avro Phonetic layout (source:https://bn.wikipedia.org/wiki/%E0%A6%9A%E0%A6%BF%E0%A6%A4%E0%A7%8D%E0%A6%B0:Avro_Phonetic_Keyboard_Layout.png)

Figure 8: Avro Mouse click layout

(source:https://sifatblog.blogspot.com/2012/10/a-full-unicode-supported-bangla-typing.html)

Figure 9:Probhat keyboard layout

(Source:https://bn.m.wikipedia.org/wiki/%E0%A6%9A%E0%A6%BF%E0%A6%A4%E0%A7%8D%E0%A6%B0:KB-Bengali-Probhat.svg)

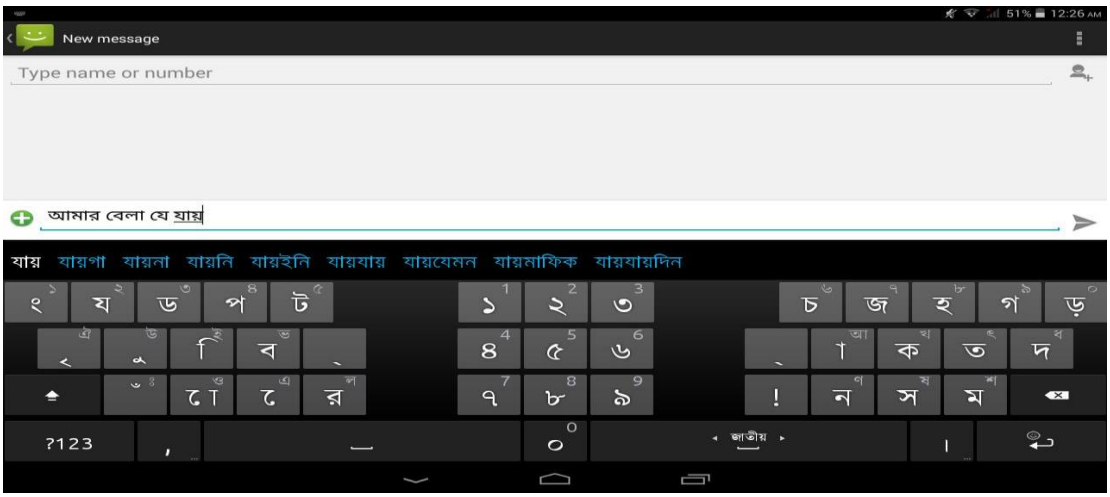

Figure 10: Ridmik keyboard layout for android

(Source: https://apkpure.com/ridmik-keyboard/ridmik.keyboard)

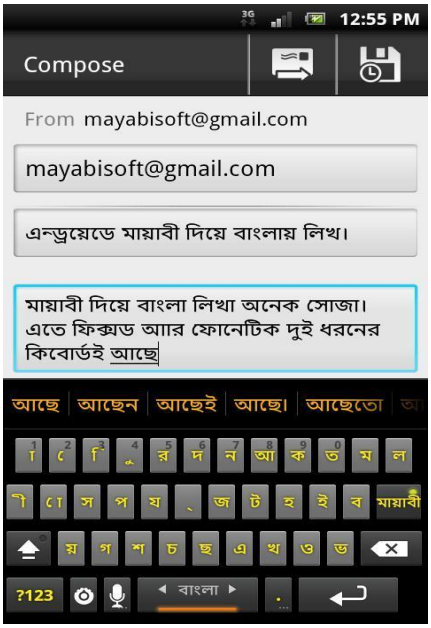

Figure 11: Mayabi keyboard layout for android

(Source:https://image.winudf.com/v2/image1/Y29tLm1heWFiaXNvZnQuaW5wdXRtZXRob2QubGF0aW5fc2NyZWVuXzRfMTU2Njk5NzA5Nl8wNzE/screen-4.jpg?fakeurl=1&type=.jpg)

2.3. Latest Unicode Chart Version 13.0 of Bangla Languages:

There are 128 spaces (96 codes are assigned, and 32 code points are reserved) in the Bangla language's latest Unicode chart. These codes cover Bangla Scripts coverlike Assamese, Bangla, Bishnupriya, Daphla, Garo, Hallam, Khasi, Manipuri, Mizo, Munda, Naga, Riang, and Santali languages.

The place of Bengali in Unicode is 0980 — O9FF. Except that the built-in language, this is undoubtedly complex. The alphabets and figures utilized in written kind reveal significant complexity. Some diacritics (shorthand sort of vowel indications whenever used in combination with consonants) are placed within the /eft associated with an associated consonant: some are found in the proper. Some others are identified below. And some other people have opportunities in both above, and left or appropriate, and above. A few of these vowel signs or diacritics have unique signals in Unicode. In contrast. Some diacritics have two components: one is put before the consonant and the various other rest. There exists a rule this is undoubtedly unique for every single among these indications. Nonetheless. These indications could be additionally represented by incorporating two other regulations and are formally allowed. For

Dhaka University Institutional Repository

example, The sign ‘ো’ is also written combining ‘ে’, ‘া’, and’. Some consonants also have two components: a person is the critical human anatomy, and the various other is just a unique sign ~Nukta’. This is undoubtedly the main character, on the other hand, a dot placed directly under your body. Every one of them features an exceptional location in Unicode as being a character. Your body, this is undoubtedly the primary of these characters, gets the model of other personality. Which has places that can be unique Unicode; additionally, the “Nukta” itself has a unique Unicode. Therefore, these figures may be represented often by their particular special rule or a mix of two Unicode’s. For instance, ‘ড়’ can be additionally compiled by combining ‘ড’, and ‘়’ without seen by the reader. Once More. Some character combinations are often represented with some special ligatures; the non-ligature types may also be acceptable. The form that is non-ligature easily be gotten by placing a zero-width non-joiner (ZWNJ) character (Unicode 200C). The ligature type and type that is non-ligature (for example, @, and *f) could be interchangeably utilized to portray the bit 0 and 1 (Khairullah&Ratul, 2018).

Recently, the use of a separate Nukta for ‘য় ড় ঢ়’ has not been accepted by the BCC (Bangladesh Computer Council), which is bargaining with the UnicodeConsortium for the atomic character code these. Bangladesh got membership in Unicode Consortium in 2014, but we still have many problems that need to be solved.

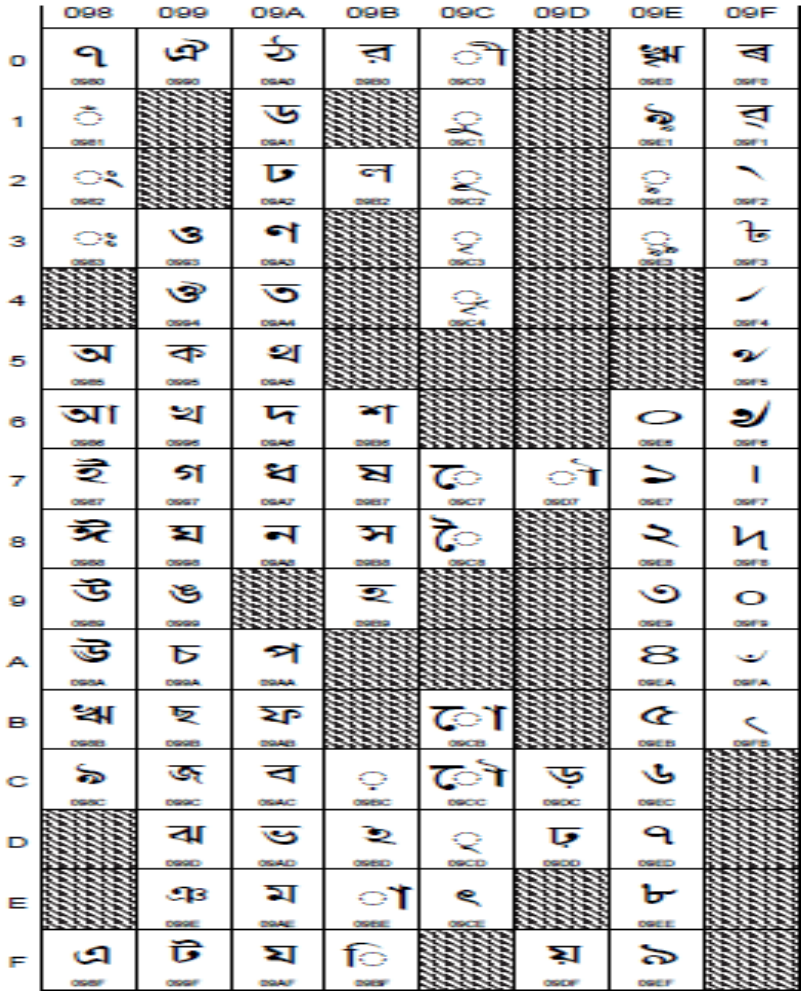

Figure 12: Official Bangla Unicode chart Version 13.0

(Source: https://unicode.org/charts/PDF/U0980.pdf)

2.4. Font Converter (ASCII to Unicode):

We inputted Bangla Language in the computer through ASCII format, but all platforms were not supported. We have already overcome this problem with the help of the Unicode system. “Most of the Bangla documents are stored in ASCII format. Nowadays, Unicode font is getting popular for supporting all devices and for web publication. Unicode font is the universally used coding system that provides a unique number of every character irrespective of the platform, program, and language(MA. Addison-Wesley, 2003).” On the other hand, we could not recognize the ASCII font if the system does not have a particular font. ASCII is a machine dependable encoding system. Some Unicode converters are- Avro converter developed by Omicron lab, Nikosh converter designed by Bangladesh Election Commission, CRBLP developed TTF to Unicode font converter, Intelligent Bengali Unicode Converter (IBUC)

Dhaka University Institutional Repository

designed by S. Sazzadur Rahman et al. Another one is Bangla Unicode, and ASCII Text converter that only supports Bijoy key mapping, Bijoy Ekushey converter is developed by Ananda computer, and some online converters are available.

2.5. Text to Speech Converter:

RajanShaha et al., 2019, proposed a Bangla Text to speech system using deepneural networks. Their database contains 40 hours of speech, including 12,500 utterances and a pronunciation dictionary containing 1,35,000 words for front-end text processing. Hence, a deep neural network extracts linguistic features and then converts them into acoustic features. K.M. Azharul Hassan et al., 2014, developed TTS tools based on text normalization and synthesis. Hence, text normalization and synthesis follow the Bangla pronunciation rule. ShammurAbsar Chowdhury et al., 2011, proposed a unit-selection-based speech synthesizer system for TTS. They developed a new tool using 'Katha' TTS developed earlier. The first step of building a parametric voice is collecting a large amount of speech data and the associated transcriptions. There is no useful quality public TTS data available for Bangla. Google has released their Bangla TTS data, but it contains only three hours of speech recorded with multiple speakers. A few more public datasets are available, all of which include merely a couple of hours of lecture. Those datasets were proposed for Unit-selection TTS and acoustical analysis of Bangla Speech (Rajon Shaha Raju et al., 2019). M. Masud Rashid et al., 2010, presented a framework based on text normalization and di-phone preparation. Abu Naser et al., 2009, developed the 'Shubachan' TTS system based on di-phone concatenation. Firoz Alam et al., 2007, created the 'Katha' TTS system based on the Festival toolkit.

2.6. Bangla Speech to Text Software:

Speech Recognition is a very challenging task in the field of BLP. Some software is available in, android version, IOS version, windows version. 'Speechtyping' is online software which only supports google chrome version 26 or upper. Hence, some sounds cannot be recognized, for example, 'ড়', 'ঢ়', 'ঞ, and 'চিত্রাংকন', 'জগনময়' these words recognized, but spelling is not correct. The correct spelling of the terms described aboveis 'চিত্রাঙ্কন', and 'জগন্ময়. In, android version,we did not get any

Dhaka University Institutional Repository

convenient tool. Most of the tools do not work or run correctly. “The difficulty of this task is due to the acoustic similarity of many of the letters. Accurate recognition requires the system to perform fine phonetic distinctions. English, and Bangla, two languages belonging to Indo-European Family, have spectacular similarities and differences in their phonemic systems (SAYEM, 2014).”

2.7. Bangla Search Engine:

There are many popular search engines like Google, Bing, Yahoo, Baidu, but before 2013, there was no Bangla search engine. For the first time, we got a Bangla search engine named ‘PIPILIKA’ in 2013. It was a joint project by SUST and GP IT Ltd. Some features added in February 2018 of the ‘PIPILIKA’:

1. Four search options are available- Product search, Job search, Library search & Latest news
2. One can search either in Bangla or English
3. This search engine can automatically analyze and preserve Bangla Blog, Bangla Wikipedia, the Latest news of Bangla, and English and Government information.

2.8. Bangla Pronunciation Software:

Thus far, this field, ‘Bangla Academy’, has come up with a publication of a "UccharonObhidhan"– ‘Pronunciation Dictionary’ that lacks linguistically recognized IPA (International Phonetic Alphabets) inscriptions for practicing pronunciations. NIMCO had published a pronunciation dictionary couple of years back, but that was not acclaimed by the scholars/P, and its, and users. Their activities are inadequate for general/free use and designed only for the electronic media performers - thus not available for extensive everyday use. Modern Language Institute of Dhaka University has no such materials except a CD produced jointly with the Ananda Computer. But that is only an abridged recording tailored for foreign learners/travelers in Bangladesh. BCC (BangladeshComputer Council) has started a project related to Bangla language processing but, there has no program of working on Bangla pronunciation within its 16 components master plan. They just collected some situation-based speeches for annotation. On the other h, and in google translate,

Dhaka University Institutional Repository

there is an option for the pronunciation of Bangla words. But that is not being authenticated and validated by the scholars or linguists. Also, their data is protected and will not be opened for developing any tools. The Bangladesh Government has also undertaken an initiative to define the Bangla language's practical usages through education and training (National Curriculum 2012, and forthcoming Curriculum). A project named ‘Standard Pronunciation of Selective Bangla Words: Digitalization, and Preservation’ is going on at North South University. In this project, almost 30,000 plus words categorically will be preserved through an audio file.

2.9. Bangla Word Clustering:

Bangla word clustering is significant for POS tagging, and it helps develop tools like spell checker, grammar checker, text classification, etc.Sabbir Ismail et al., 2014 developed Bangla word clustering based on the N-gram language model where 2,51,89,733 individual words facilitate effective employment of unsupervised machine learning-based clustering. Also, it was based on their semantic and contextual similarity.

2.10. Bangla POS Tagging Software:

Parts of Speech tagging is very significant in Bangla Language processing. It will help develop tools like speech to text, vice-versa, Machine Translation, Call Assistance, etc. In Bangla Language, this tool is not developed according to expert expectations, as Bangla is complex. There are two methods of POS tagging, including statistical and deep neural networks. Several works have been done in two ways. The top job has been done in a rule-based manner. In 2018, Md. Nazim Uddin et al. proposed a hybrid method combining statistical with neural network approach for Bangla POS tagging. They used a lot of tagged data and a converter for converting words to their root. They claimed their accuracy is 91.2%. In 2015, M. N. Haque and M. H. Seddiqui. proposed a Bangla POS tagger based on stemmer and rule-based analysis. In 2012, K. Sarkar and V. Gayen proposed a statistical method where used trigram and second-order Hidden Markov Model (HMM) for Bangla Parts of Speech Tagging.

Dhaka University Institutional Repository

2.11. Bangla Sentiment Analysis Software:

Sentiment analysis is very significant for machine translation. It is a challenging task to find out the exact emotion. In SUST, some researchers developed their dataset collected recorded full sentences of different contexts like angry mood, fund mood, everyday mood, loudly, and many more from other users. In BCC, a team developed a sentiment analysis software where the tool only recognizes 'ভালো' as positive, but if we put 'অত ভালো না' then the software also indicates positive. But the fact is the /meaning of the phrases is negative.

Sentiment evaluation (SA) involves removing the level. This is undoubtedly emotional of observation, evaluation, or viewpoint on various social aspects such as items, solutions, or individuals. This has developed into research this is certainly well-known in normal Language Processing (NLP) for enabling a machine to recognize just the right sentiments of texts. Recently, there have been works that can be numerous SA emphasizing English or other languages, but relatively fewer tasks are significant in Bangla. Many earlier deals with SA into the Bangla language have been completed integrating machine discovering algorithms. Still, an in-depth approach that is learn-ing is rarely found due to the scarcity of massive datasets. In this work, the researcher presentsBangla's performance by the deep is discovering, and device under, anding gets near making use of various term systems. This is undoubtedly embedding, then compare among all formulas using some analytical procedures to learn the overall performance methods that are best for Bangla binary sentiment from texts. Hence, the researchers also provide a pre-trained design that is fine-tuned Bn Sentiment to be used by the researchers and people who like to automate the detection of sentiment polarity on their review system. This module makes healthy the Bangla NLP neighborhood for additional analysis. The design has a reliability of 84.5%, which has been trained on a dataset that is sufficiently huge (Kowsar, 2020).

2.12. Bangla Spell Checker:

So far, we have no complete Bangla spell checker. For example, in English, we have an inbuilt spell checker responsible for spelling correction, sentence structure

Dhaka University Institutional Repository

correction, analyzing context, suggesting edits, and many more options in Microsoft word. English language and other language's like, Arabic, Chinese, and Hindi have their spell checker, which works well. In Bangla language first attempt was taken by 'Proshika'. They launched a spell checker named 'NIRVOOL'. But it is just primary level and slow. Sometimes, the computer gets hangs. A significant lacking is that there are no similar-sounding word suggestions. Many more researchers did their work on automatic spelling corrections of Bangla Language and showed their accuracy. We will discuss this in detail here. But the main problem is the corpus size.

2.12.1. Error Types of Spell Checker:

There are few error types which are described below in short-

Two kinds of typing errors were classified by Kukich. The errors- Real-word error, and another other is the non-word error. (K Kukich, 1992). 'আমি তোমাকে খাই' in this sentence 'খাই' is a valid word, real-word error, but grammatically incorrect. On the other hand, the example of non-word error, might be as following:

'আমি তোমাকে চাএ'- in this sentence'চাএ'is a non-word error because it is not a valid word.

2.12.2. Non-word Error Types:

'বাংলাদেশ' the word may be shown error like;

Insertion**:** 'বাংলালদেশ 'here **'**ল**'** is inserted

Deletion: 'বংলাদেশ' here **'**া**'** is deleted

Substitution: 'বাংনাদেশ 'here 'ন' is substituted

Transposition: 'বাংলাদশে' or **'**বাংলাশেদ' here **'**ে**'**, and **'**দ' are not in the right position.

These are typographical errors. Another error is a cognitive error. The mental mistake refers to that when a person lacks knowledge.

2.12.3. Cognitive Error Types:

Spelling mistake: ‘সাধীন’ (সঠিক- ‘স্বাধীন’) or ‘আমী’

Incomplete sentence: ‘তুমি আসলে...’

Lack of coherence: ‘নাম বাংলাদেশ দেশের আমার’

Invalid Sentence: ‘আকাশ মাটিতে অবস্থিত’

Punctuation Marks error: ‘আকাশ সবুজ নীল তিনবন্ধু’or ‘আকাশ! সবুজ! নীলতিনবন্ধু’

Redundant part: ‘সকল শিক্ষার্থীবৃন্দ’ (Double plural)

‘একত্রিত’(একত্র+ইত) (সঠিক-‘একত্র’) here ‘ইত’ is redundant suffix.

Lack of collocation: ‘গালি ছোঁড়ে’ (সঠিক- ‘গালিদেয়’)

or ‘বুদ্ধিমতী ছেলেটি এল’

The complexity of homophonous words (same pronunciation, but meaning, and spelling are different): ‘ধ্বনি’ ‘ধনী’

The complexity of homograph **(**same spelling, but different meaning**)**: ‘মাথা’

‘আমার মাথা ব্যথা করছে’ (‘ মাথা’-অঙ্গ অর্থে)

‘তার অংকে মাথা ভাল’ **(**‘মাথা’ ভাল-দক্ষতা অর্থে**)**

Tense error: ‘গতকাল, আমি যাব’

Subject-verb agreement error: ‘আমি কাজটি কর.’

Error in Proverbs: ‘আমড়া কাঠের ঢেঁকি’ meaning ‘অপদার্থ’

‘আম গাছের ঢেঁকি’ is not correct for ‘অপদার্থ’

Space among words: ‘অজ্ঞান’ (সঠিক- ‘অজ্ঞান’)

Write according to pronunciation: ‘আমায় খমা করে দাও’ (সঠিক-‘ক্ষমা’)

Almost similar ligature: ‘জ্ঞ’‘ঞ্জ’

Similarity between ligature and letter: ‘এ’ ‘ত্র’; ‘ও’ ‘ত্ত’

Sound change: ‘ক্ষমা’ (খমা), but if we analyze the ‘ক্ষ ’then we get ‘ক্ষ= ক্+ষ.’

Phoneme: Grapheme= 1:1 (Absent), but 1: many

for examples: ‘ন’ ‘ণ’ [n]; ‘শ’ ‘ষ’ [ʃ]; ‘জ’ ‘য’ [j]; ‘ঙ’ ‘ং’ [ŋ]

2.12.4. Spell Checker Developing by BCCand Others

At present (202), the Bangladesh computer council is developing a spell checker, launched soon. Correct the non-word error: Used technique/Algorithm (Edit distance). Correct the Real word-error: Used technique/ Algorithm (N-gram models). Mittra et al., 2019, proposed a hybrid approach by combining edit distance algorithm, and probability-based n-gram language model has been used. They focused on both structural and syntactic similarities—Corpus was used in this research containing 0.5 million words and 50,000 n-gram sentences to detect and correct errors. N-gram models are calculated using the principle of linear interpolation model. A grouped-based dictionary where each of the words equal length is grouped was used. That is why time and space complexity is minimized. They collected sources online. 97% accuracy is claimed. Md. Tarek Habib et al., 2018, proposed a context-sensitive approach where they had achieved a satisfactory and promising accuracy of 87.58%. This research's main features are musing edit distance, and stochastic, N-gram language model (Used six N-gram models, including forwarding bigram & trigram, backward bigram & trigram, and combined bigram & trigram). Under the proposed model researcher used both context-free and context-sensitive approaches. Edit distance is responsible for context-free assessment, and N-gram is accountable for context-sensitive approach. They used 3,734,596 words and 312,449 sentences as a test dataset. (Islam, M. I. K et al., 2018). After Tarek Habib et al., 2018, The following table represents some methods applied by a researcher with the type of error handled.

Developer/ Researcher	Algorithm/ Method Used	Size of the data	Accuracy %	Errors Handled
MuhammadIfte Khairul Islam et al., 2019	Edit distance	150,000 words	96.83%	Non-word error
Md. Tarek Habib et al., 2018	Context-sensitive technique based on N-gram, and edit distance	3734,596 words, and 312,449 sentences	87.58%	Non-word error
Prianka M, andal, B M Mainul Hossain, 2017	Clustering	2450	99.8%	phonetic, typographical
HossainKhan et al., 2014	N-gram model Character-based	50,000correct words , and 50,000 incorrect words	96.17%	Non-word error
Md. Zahurul Islam et al., 2007	Stemming algorithm and edit distance	13,000 words	90.8% for correcting single error misspellings, and 67% for correct multiple error misspellings	complex orthographic rules
M. K. Munshi et al., 2007	Finite-state automation	291	92% for correcting single character	substitution errors, insertion

			misspellings, and 70% for correcting multiple character misspellings	errors
N. Uzzaman, and M. Khan, 2006	2-edit distance and phonetic encoding	1607	91.67%	phonetic, typographical, OCR generated
Abdullah & Rahman, 2004	Direct Dictionary lookup method, and Recursive Simulation algorithm	NM	NM	typographical errors, and cognitive phonetic errors
N. Uzzaman, and M. Khan, 2004	phonetic encoding	*NM	More than 80%	phonetic, typographical
Bidyut Baran Chaudhuri, 2001)	String matching algorithm	25,000 words	high accuracy with 5% false positive detection	phonetic error

Table 1:Comparative study of Bangla Spell Checker Performance

2.13. Machine Translation (MT):

In the late 1950s, the researcher started work on Machine Translation with high hope and little realization of the difficulties. In the early 1960s, noticeable work was done. Still, it was perceived at the same time that without the necessary work on the text 'under, anding', it is not possible to present fully automatic high-quality translation. In the United States, very few projects on machine translation have been addressed in

computational linguistics. There are substantial projects in Europe and Japan (Slocum, 1984, 1985; Tucker, 1984).

In Bangladesh, the research on machine translation started in 1991. Earlier the statistical machine translation was used. Afterward, in the 21st century, the deep neural network was created to explore. Nowadays, we moved from SMT (Statistical Machine Translation) to NMT (Neural Machine Translation) for getting better output. We see most of the machine translations are being applied to Bangla to English, and vice-versa. According to FirojAlom et al., 2019, they proposed an NMT approach where they used publicly available corpora, including Indic Languages Multilingual Parallel Corpus approach (ILMPC), Six Indian Parallel Corpora (SIPC), Penn-Treebank Bangla-English parallel Corpora (PTB), SUPara Corpora, AmaderCat corpus. Professor Abdullah Al Mumin et al., 2014 developed a corpus consisting of 27 million words for machine translation. Bangladesh Computer Council is working on machine translation. Language modeling is a very significant task to apply many tools successfully. Shuvendu Roy, 2019, proposed a method named the convolution architecture approach to model the Bangla language two datasets used- English and Bangla.

2.14. Bangla Corpus:

In the 21st century, the Bangla language cannot be computerized up to the user's dem, and only for the lack of a Bangla corpus. The available English Language corpus is British National Corpus (BNC), American National Corpus (ANC).Md. Abdullah Al Mumin et al., 2014, developed a SuMono Bangla Corpus consisting of 27 million words. This study ran in 2010. SuMono corpus sources are into six categories: natural science, social science, computer, IT, Literature, Mass Media, and blogs. Bangladesh Computer council are developing a written based corpus. Md. Farukuzzaman et al., 2017, created a new Bangla corpus named BDNC01 with analysis. They spend nearly six years collecting around 12 million-word tokens, including literary canon. The sources were newspapers, includingProthomAlo, Daily Jugantor, Inqilab, Amar Desh, Daily Ittefaq, VhorerKagaj. The collection period was between 2004 to 2010. It is preserved in both ASCII and Unicode format.

Chapter 3
Theoretical Discussion

3.1.Language Processing:

There are two types of language processing -language processing in the human brain and, on the other hand, language processing in a computer. Shortly, we discussed the processing of language in the human brain, but our primary focus is the processing of speech on the computer.

3.1.1. Language Processing in the Human Brain:

Language processing in the brain is very complicated. From 1861, Researchers started to figure out the parts involving language processing. At that time, Paul Broca, a French neurosurgeon, proved that the brain's left hemisphere is responsible for speech production and was the first to identify the language center situated in the posterior portion of the frontal lobe of the left hemisphere. Now it is known as Broca's area. About ten years later, Carl Wernicke, a German neurologist, discovered another part of the human brain situated in the posterior portion of the left temporal lobe. This portion helps to understand the language. Broca's area is Broadman area 44/45, and Wernicke's area is Broadman area 42/22 [figure 13]

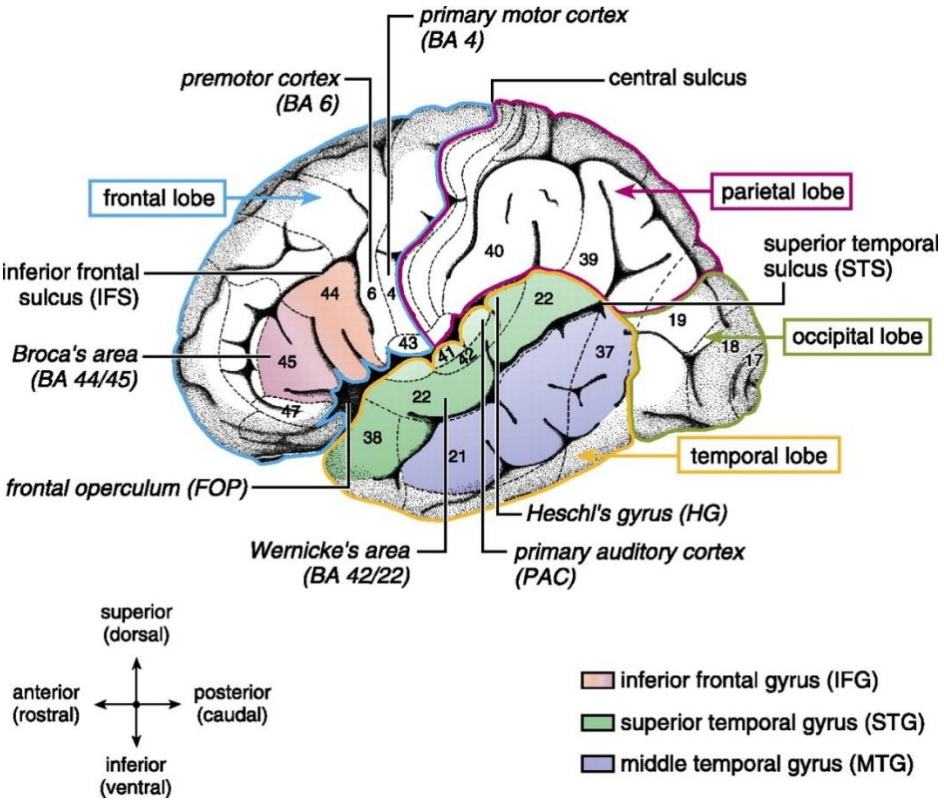

Figure 13:Anatomical, and cytoarchitectonic details of the left hemisphere (source:https://journals.physiology.org/doi/full/10.1152/physrev.00006.2011/)

3.1.1.1. The Functions of Broca Area:

Pierre Paul Broca names this area after an investigation, which reveals that a person lost the ability to speak for the reason of injury to the posterior inferior frontal gyrus of his brain. The main functions of Broca's area:

i) Language production

ii) Language comprehension to some extent (Devlin et al. 2003)

iii) It deals with semantic tasks, including understanding the meaning of the word.

iv) It focuses on phonological tasks, including knowing the sound of the word.

v) Speech production

vi) Facial expression

vii) Body Language

3.1.1.2. The Function of Wernicke Area:

Wernicke's location is known for the physician who first identified the region's tasks associated with the brain. Carl Wernicke, a 26-year-old physician-in-training in the early 1870s, saw several patients with the same language under and low. Their manufacturing had been disrupted as well. That they had difficulty finding words and produced many non-words (neologisms) and semantically associated term substitutions although they spoke fluently, their particular speech made small feelings(Saffran, 2002). After Alfredo Ardila et al., 2016, the main functions of the Wernicke area are-

i. Language understanding
ii. It recognizes phonemes and words (vocabulary).
iii. A way for linking the sensory patterns of words to the distributed associations that encode their meaning (Mesulam, 2001).
iv. It is included in language network as a third central hub that plays a critical role in language comprehension, particularly the understanding of words that denote concrete entities (Mesulam, 2013)

3.1.2. Brodman Area:

In the early 1900's German anatomist Korbinian Brodmann defined and numbered the Brodmann areas of the cerebral cortex. These areas initially were based on a cytoarchitectural organization of neurons in the cerebral cortex. Zone 4, known as the primary motor cortex, is responsible for executive motor movements, including finger movements. Area 5 is responsible for various functions, but language processing is one of the most crucial roles. By conversion, zone 6 is vital for language and memory functions, including speech motor programming, language processing, switching, speech perception, object naming, lip-reading, word retrieval, the lexical decision on words, and pseudowords, syntactical processing, etc.The essential functions of different areas are as follows:

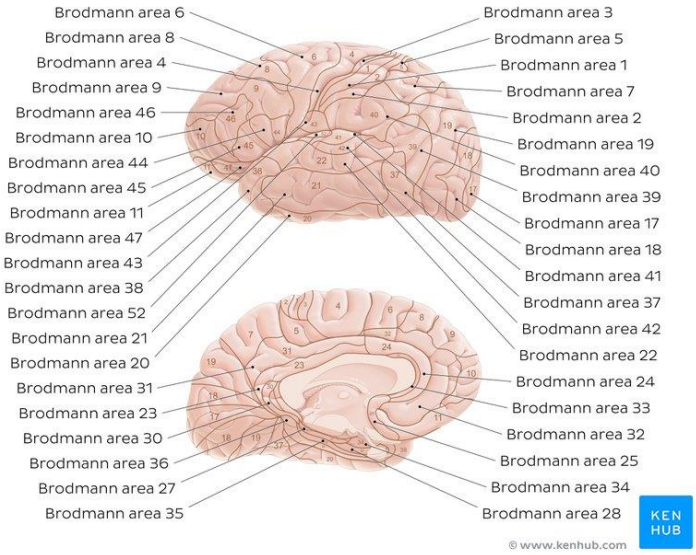

Figure 14: Brodman area overview

(Source: https://www.kenhub.com/en/library/anatomy/brodmann-areas)

critical facts about Brodmann areas (*Brodmann Areas: Anatomy and Functions | Kenhub*, n.d.)

Areas 1, 2, 3	Primary somatosensory cortex (postcentral gyrus)
Area 4	Primary motor cortex (precentral gyrus)
Area 5	Somatosensory association cortex
Area 6	Premotor and supplementary motor cortex
Area 9	Dorsolateral/anterior prefrontal cortex (motor planning and organization)
Area 10	The anterior prefrontal cortex (memory retrieval)
Area 17	Primary visual cortex
Area 22	Primary auditory cortex
Area 37	Occipitotemporal (fusiform) gyrus
Areas 22, 39, 40	Wernicke's area (language comprehension)

Our primary focus is Bangla Language processing in Bangladesh. Now we have to shift towards natural language processing

3.1.3. Natural Language Processing in Computer:

Simply, Natural Language Processing means how a computer communicates with a human being like a human. Nowadays, it is also known as the branch of artificial intelligence (AI). Natural Language Processing is a technology to aid computers in understand and the human's natural language.It works like a bridge between human beings and computers. The final purpose of NLP is to recognize, decode, understand, and give output in a way that can be valuable for us (Michael, 2018).

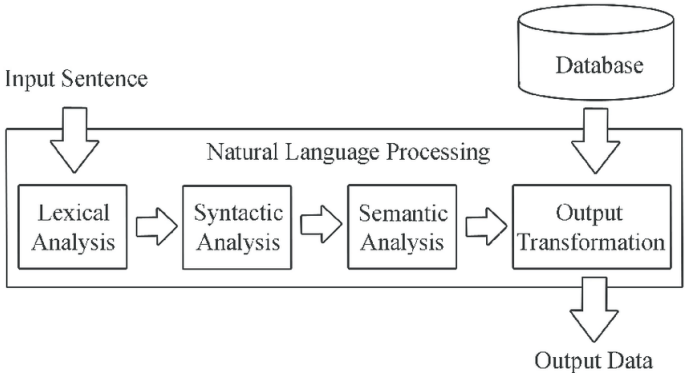

Figure 15: Natural Language Processing steps(source:https://www.researchgate.net/figure/Natural-Language-Processing-steps_fig1_311705165)

3.2. Language Processing Tools:

There are many Natural Language Processing software. For example, Optical Character Recognition, Machine translator, Speech to text software, Text to speech software, Spell, grammar checker, Font converter, IPA transcription, Sentiment analyzer, etc. POS tagging. Stemmer, Keyboard interface, fonts, etc. Some of the tools have been discussed below.

3.2.1. Machine Translation (MT):

By machine translation, one can translate a complete sentence from the source language (SL) to the target language (TL) as error-free. Nowadays, we see a vast improvement in its technology, including user data, information, and technology.

It can be considered a useful tool that can impact the global commercial market of cross-lingual information, interlingual communication, and information exchange. Moreover, it is now a cross-disciplinary sector directly impacting e-commerce, localization, and knowledge-based society (Winograd 1983). Representative Corpus is a must for a representative Machine Translation Tool. Conceptually, the device that is a corpus-based (CBMT) system is founded on a range of dilemmas developed from an empirical evaluation of BTC. The evaluation work requires both linguistic (morphological, semantic, and explanation. This is certainly practical of terms, expressions, phrases, sentences, etc.), extralinguistic analyses of data, and information present in a BTC. It uses different analytical ways to generate likelihood data from BTC to identify the nearest interpretation possible through the two languages (Altenberg and Ajmer, 2000). After Niladri Sekhar Dash, and L. Ramamoorthy, 2019, the primary issues relating to the development of a CBMT system are the followings:

(a) Generation of bilingual translation corpus,
(b) Alignment of bilingual translation corpus,
(c) Linguistic analysis of bilingual translation corpus,
(d) Extraction of translational equivalents,
(e) Generation of the terminology data bank,
(f) Building a bilingual dictionary,
(g) Algorithm for lexical selection,
(h) Dissolving lexical ambiguity,
(i) Formation of grammatical mapping rules,
(j) Formation of lexical mapping rules,
(k) Selection of translational equivalents,
(l) Addition of pragmatic information,
(m) Addition of sentential information.

3.2.2. Optical Character Recognition (OCR):

Optical Character Recognition (OCR) is a tool by which one can do the following things:Digitized an image, type, print text, scanned documents, screenshots, handwritten documents, etc., for searching, editing, or sorting from a digital version. It works with Artificial Intelligence (AI), Pattern Recognition, and Computer Vision.

OCR can recognize:

- An image
- Typed documents
- Printed texts
- Scanned documents
- Screenshots
- Handwritten documents
- A scene photo, including texts on signs, text on billboard-in a landscape photo

3.2.2.1. Basic Steps of an OCR:

The necessary steps of OCR are discussed shortly below:

1. Scanning/ Image Acquisition/ Document Input/ Optical Scanning
2. Location Segmentation
3. Segmentation
4. Preprocessing
5. Feature Extraction or Pattern Recognition
6. Training and Recognition/ Recognition using one or more classifiers
7. Contextual Verification or post Processing

3.2.3. Spell Checker:

The spell checker checks misspelled words. For the English language, the research on Spell Checker started in 1961. Based on the Mainframe computer, around the 1970s spell checker was developed. Common used methods are-clustering, edit-distance, phonetic encoding, string matching algorithm, stemming algorithm, etc. There are several types of spell checkers. For Example,-

1. Detect wrong words, and make suggestions
2. Search correct spelling through a search engine from an inbuilt electronic dictionary
3. Some spell checker is developed with an electronic dictionary, including grammatical definitions.

3.2.3.1. Some Features of Spell Checker:

A spell checker should include:

- A search engine
- An electronic dictionary
- Autocorrect/autosuggestions
- Detect all kind of errors, including non-word error, cognitive error, and typos

3.2.4. Fonts:

A specific form, including the weight, size, and style, is known as a font. The characteristics of fonts include-

- weight,
- slope,
- wide,
- length,
- metrics,
- optical size,
- sheriffs
- Character variants.

Some features of the fonts have been shown by the following figures 16, 17, 18, 19, and 20-

Figure16: Typeface nomenclature for the Bengali Script (Ross, 1999)

(Source: Karim et al., 2013, p-4)

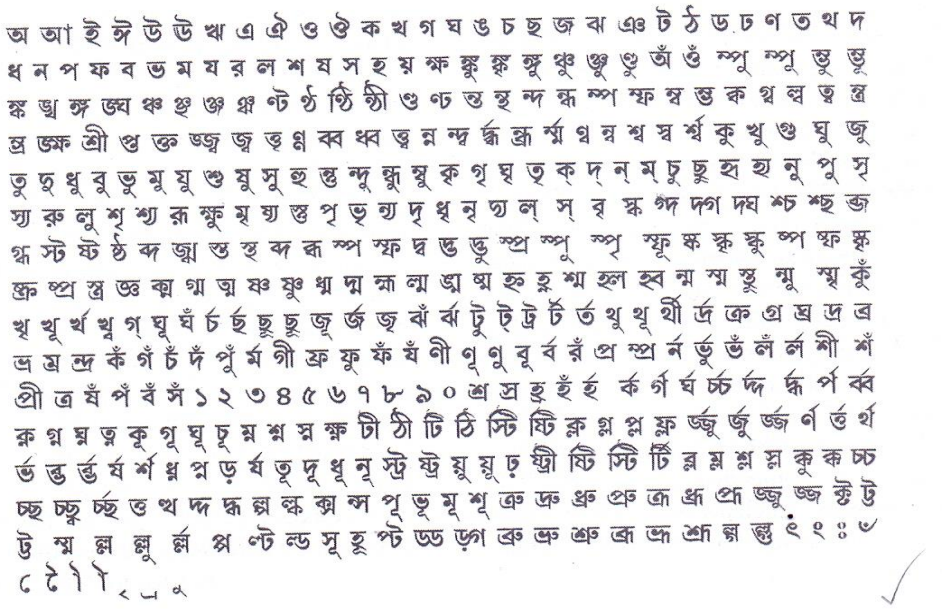

Figure 17: Foundry typeface no.6 used by Ananda Bazar Patrika Ltd before 1982 (Ross,2013)

(Source: Karim et al., 2013, p-6)

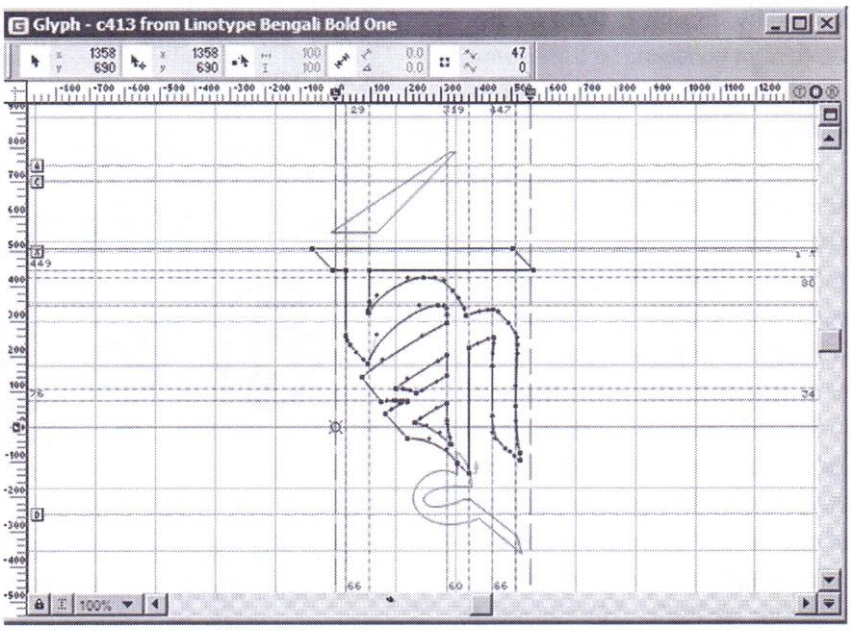

Figure 18: Establishing the dimensions using the resonant glyph with superscript and subscript (Ross, 2013) [Source: Karim et al., 2013, p-9]

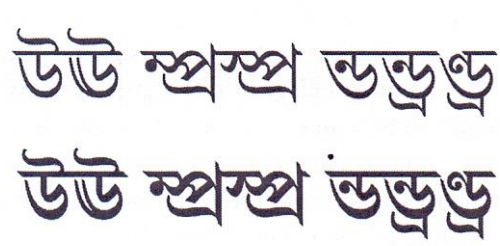

Figure 19: An example, of Bengali type forms in two weights that need to show affiliation, and differentiation (Ross, 2013) [Source: Karim et al., 2013, p-9]

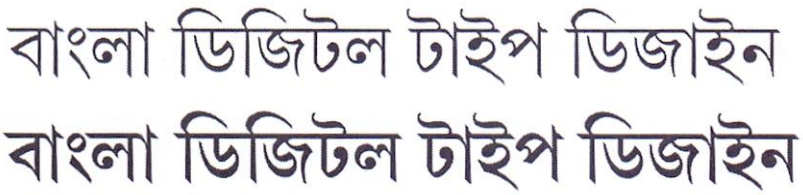

Figure 20: 'Bangla digital type design' set in Linotype Bengali light, and bold (Ross, 2013)

(Source: Karim et al., 2013, p-11)

3.2.4.1. Unicode:

Unicode Consortium manages this Unicode system. This uniform encoding system started in 1987. In 1991 the first version was released. There are some UTF (Unicode Transformation Formats)-UTF-8, UTF-16, and UTF-32. "Unicode is an 'international multi-byte character encoding that covers virtually all of the world's languages,' and assists in the portability of documents (Adobe, 2008, p: 2)." All languages of the world have come to a unique encoding scheme. Almost 1,40,000 plus characters are included in version 13.0. Not only characters but also 'emoji' has been included. All types of letters, characters, digits, punctuation marks, diacritic marks, technical symbols, trade symbols, and special characters have been included in this uniform encoding system.

3.2.4.2. ASCII:

ASCII's elaboration is **American Standard Code For Information Interchange**, an encoding system having 128 characters, including numbers, letters through, and punctuation marks through the computer. Earlier it was based on seven digits, now it moved to 8 digits, and the total number of characters is 256. Before, it was based on teletypewriters, but it is later used in modern applications until now. In 1981, this code was introduced by the International Business Machines Corporation (IBM).

3.2.5. Keyboard Interface and Keyboard Layout:

The console is a little PC of sorts. It is equipped for sending, getting orders, buffering information, and keeping up with wanted settings. This application note will make just arrangement with the console sending information.

Each key on a console is relegated to a sweeping code. This output code, a two-digit hexadecimal number for most tickets, is sent by the console whenever the key is squeezed. The code will be sent repeatedly on the off chance that the key is held down for more than its typematic defer. When the key is delivered, the console sends F0 in hex, and afterward, the sweep code of the key is provided. This works for all keys on the console. A fascinating note is about the "covers lock," "num lock," and "parchment lock" keys. These keys have a LED on most consoles that is flipped here and there as the key is squeezed. This LED is constrained by the gadget regulator, the console, and your PC much of the time. At the point when you press the "covers lock" key, for instance, the console tells your PC that the key has been squeezed, and the PC advises the console to turn on or off, by and large, the corresponding (Hassan, 2020)

3.2.6. Speech to Text Converter:

Speech annotation is obligatory for speech to text converter. Hence, we insert our voice throughthe microphone, and we get a digital waveform from the analog.The recognition system's primary function is to annotate the inserted waveform's necessary data to present the correct output. Training and recognition are the two phases of the system’s operation. In the training phase, data for known classes are fed

to the system. The system computes the pattern features for unknown input in the recognition phase and identifies the information with the reference pattern matching these features most closely (SAYEM, 2014). Then the annotated speech converts into text.

3.2.7. Text to Speech Converter:

Text to speech software is a process where text can be converted into sounds. It is a challenging task also. Sometimes, software got hang during the process. It is seen that some words can not be converted into speech in the android version. “Text to speech (TTS) is software or hardwaresystem for speech synthesis application that is used to create a spoken sound versionof the text from computer document. The purpose of TTS is to enable the reading of computer document or information for the visually challenged persons, or may be used to augment the reading of a text message(W. Ainsworth, 1973, pp-288-290).”

3.2.8. Corpus:

A corpus is a collection of large amounts of text where we can find any grammatical components and structure of the features.“A corpus is a collection of language text pieces in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of Data for linguistic data (John Sinclair, 2004).” There are two types of standard language corpora, including everyday speech and standard written corpus.

3.2.8.1. Speech Corpus:

It represents the large dataset of speeches from the various context of our daily life. For example,- discourse of classes, offices, markets, public transports, seminars, interviews, family, etc., with transliteration. There are two types of speech corpora, including Reading speech and spontaneous speech. Contexts above are considered as automatic speech. On the other h, and, Read speech like-

- Readout Books
- Broadcasts News
- Audio Books

- list of words
- list of numbers, and many more

3.2.8.2. Written Corpus:

According to Md. Farukuzzaman et al., 2017, any selection must be made on some criteria to collect a corpus. The first significant step in corpus building is determiningwhich text from the corpus will be selected. Common criteria include-

- The text; whether the language originates in speech or writing, or perhaps nowadays in electronic mode.
- For example, the text type is written, whether a book, a journal, a notice, or a letter.
- The domain of the text, for example, whether academic or popular.
- The language or languages or language varieties of the corpus.
- The location of the text, for example, (the English of) the UK or Australia.
- The date or period of the text.

Chapter 4

Methodology

4.1. Method:

Our research title and subject demand the interview of Experts as well as users in this field. We have conducted this study through the qualitative method; an interview exercise and document analysis have been chosen. We used some semi-structured questions for the interview session. As the pandemic situation is ongoing, so all interviews have been done virtually.There are three types of methodology available for example, Qualitative, quantitative, and mixed-method. This study followed the qualitative method. We set some semi-structured questions for the interview and also questionnaires for data collection.

4.2. Research Question:

a) Main Research Question:

What developments have been made in the field of Bangla Natural Language processing tools in Bangladesh?

This main query will motivate us to know the answer to the following supplementary questions:

b) Supplementary Questions

I. What is the recent development of BLP tools in Bangladesh?
II. What are the limitations of developing the BLP tools in Bangladesh?
III. How many BLP tools are available now?
IV. What are the challenges in developing BLP tools?
V. Are there any linguistic challenges?
VI. Are there any technical challenges?
VII. Which sectors are involving in developing BLP tools?
VIII. Which software companies are actively working on BLP?
IX. Is there any collaboration in developing BPL tools?
X. What measures to be taken for the further development of BLP tools in Bangladesh?

4.3. Justification of the Semi-structured Questions:

The main research question deals with the overall state of Bangla Language Processing, including Machine Translation (MT), Optical Character Recognition (OCR), Bangla Spell Checker, Bangla Keyboard Interface & Layout, Fonts, Speech to Text Converter, Text to Speech Converter, IPA Transcription, Sentiment Analysis, Bangla Corpus, etc. in -Bangladesh. We came to know the present status of Bangla Language processing in Bangladesh by the first question. The second one answered what kind of bugs need to be solved and BLP tools' compatibility for different platforms. The background history of our research has been answered by third questions. The third question discussed the overall challenges while developing Bangla Language Processing Tools. The fourth question presented specific challenges in terms of linguistic and technical in developing BLP tools. The next two sub-questions under the fifth questions informed the organization's activity or institution in this field. Our last question has discussed some recommendations.

4.4. Data collection:

Some free software running for both Android and Windows were taken, and paid software like Bijoy Bayyanno version: 2016, 2020 was used in windows 10, windows ten pro. The following conjuncts have been presented in different fonts, including SutonnyMJ, Kalpurush, Nikosh, AdorshaLipi, SolaimaniLipi, Akash Lipi, SagorLipi, UniBanglaOMJ. Only SutonnyMJ has been shown in three typefaces, including regular, italic, and bold, for understanding the aesthetic view or anomalies of displaying other fonts indeed.

ক্ত: ক + ত = ক্ত	ন্তু: ন + ত + উ = ন্তু	হূ: হ + ঊ = হূ
ক্র: ক + র = ক্র	ন্ধ: ন + ধ = ন্ধ	হূ: হ + ঊ = হূ
ক্রু: ক + র + উ = ক্রু	ন্ম: ন + ম = ন্ম	হৃ: হ + ঋ = হৃ
ক্ষ: ক + ষ = ক্ষ	ন্মু: ন + ম + উ = ন্মু	হ্ণ: হ + ণ = হ্ণ
ক্স: ক + স = ক্স	ন্স: ন + স = ন্স	হ্ন: হ + ন = হ্ন
গু: গ + উ = গু	প্স: প + স = প্স	হ্ম: হ + ম = হ্ম
গ্ধ: গ + ধ = গ্ধ	ব্ধ: ব + ধ = ব্ধ	ড়: ড় + ় = ড়
ঙ্ক: ঙ + ক = ঙ্ক	ব্লু: ব + ল + উ = ব্লু	ড়ু: ড় + উ = ড়ু
ঙ্গ: ঙ + গ = ঙ্গ	ভ্র: ভ + র = ভ্র	ঢ়ু: ঢ় + উ = ঢ়ু
জ্ঞ: জ + ঞ = জ্ঞ	ভ্রু: ভ + র + উ = ভ্রু	
ঞ্চ: ঞ + চ = ঞ্চ	ভ্রূ: ভ + র + ঊ = ভ্রূশ্চ	
ঞ্জ: ঞ + জ = ঞ্জ	ম্ন: ম + ন = ম্ন	
ট্ট: ট + ট = ট্ট	ম্ফ: ম + ফ = ম্ফ	
ণ্ঠ: ণ + ঠ = ণ্ঠ	ম্ম: ম + ম = ম্ম	
ণ্ড: ণ + ড = ণ্ড	রু: র + উ = রু	
ত্ত: ত + ত = ত্ত	রূ: র + ঊ = রূ	
ত্থ: ত + থ = থ	ল্ড: ল + ড = ল্ড	
ত্ন: ত + ন = ত্ন	শু: শ + উ = শু	
ত্ম: ত + ম = ত্ম	শ্চ: শ + চ = শ্চ	
ত্র: ত + র = ত্র	শ্রু: শ + র + উ = শ্রু	
ত্রু: ত + র + উ = ত্রু	শ্রূ: শ + র + ঊ = শ্রূ	
দ্ধ: দ + ধ = দ্ধ	ষ্ণ: ষ + ণ = ষ্ণ	
দ্রু: দ + র্ + উ = দ্রু	ষ্ম: ষ + ম = ষ্ম	
দ্রূ: দ + র্ + ঊ = দ্রূ	স্ট: স + ট = স্ট	
ন্ঠ: ন + ঠ = ন্ঠ	স্তু: স + ত + উ = স্তু	
ন্ড: ন + ড = ন্ড	স্থ: স + থ = স্থ	

Our selected, android apps were Bangla OCR, Text to Speech, and vice-versa, Keyboard interface. By conversion, Bijoy Bayanno 2016, Bijoy Bayanno 2020, Avro keyboard 5.5.0, Avro Spell-checker, Sentiment Analyzer were selected for the Windows platform. Online Bangla OCR, Spell-checker, Bijoy to Unicode converter, Unicode to Bijoy converter, Google translator, Speech to Text, Text to speech, IPA transcription were selected as well. Some letters, words, and sentences have been input for Text to Speech, Spell-checker, google translator, and IPA transcription. Some sounds, pronunciation of words, and sentences have been input for Speech text, Sentiment analysis. Some Bangla documents as images and pdf had been selected for scanning by Bangla OCR. The details of the text, utterance, and tools tested have been presented in detail in data presentation chapter five.

4.5. Tools used for Typing the Fonts:

Two keyboard drivers and some fonts are used for analyzing the typeface of the data described above. Bijoy Bayyanno 2020 and Avro 5.5.0.0 keyboard drivers were used. Both the Bijoy Classic and Bijoy Unicode were used. Fonts like SutonnyMJ, Kalpurush, Shonar Bangla, Nikosh, Adorsho Lipi, Solaimani Lipi, Akash Lipi, Sagor Lipi, UnibanglaOMJ have been used.

4.6. Devices used for Software Testing:

We used Redmi 9C (MIUI version-MIUI Global 12.0.1 stable; android version-10QP1A. 190711.020)and OPPO A371 (Version-5.11, processor QualcommMSM8916Quadcore) for testing, android apps. On the other hand, Desktop (Windows 10) and Laptop (Window s 10 Pro; DESKTOP-35F3HM7; Processor-AMD A4-9125 RADEON R3 COMPUTER CORE 2C+2G 2.3GHz, 64 bit)

4.7. Interviewee Selection:

We have collected data from five interviewee groups, including policymaker, language expert, technology expert, developer& user group. The number of policymakers was three to four, language experts were five, technology experts were four and researchers were four to five, and finally, users were four to five.

a) Language expert:

1. Professor (Retd.) Daniul Haq, Bangla Department, Jahangirnagar university
2. Professor Dr. Shakhawat Ansari, Department of Linguistics, University of Dhaka
3. Professor Md. Sajjadul Islam, Bangla Department, Jahangirnagar university
4. Associate Professor, Tarik Manzoor, Bangla Department, University of Dhaka
5. Associate Professor. Mohammad Azam, Bangla Department, University of Dhaka
6. Associate Professor Dr. Mamun Or Rashid, Jahangirnagar university

b) Technological Expert:

1. Prof. Dr. Md. Zafar Ikbal, CSE, SUST
2. Prof. Dr. Abdullah Al Mumin, CSE, SUST
3. Associate Professor Dr. Nabil Mohammad, CSE, NSU

c) Researcher Group:

1. Prof. Dr. Niladri Shekhor Dash, ISI, India
2. Assistant Professor Mashrur Imtiaz, Department of Linguistics, DU
3. Rifat Hasan Jihad, Mechanical Engineering, KHUET
4. Arif Ahmed, Information Science, Leading University, Sylhet

d) User Group:

1. A.S.M Shakil Haider, Ph.D.candidate, Texas Tech University, USA
2. Sabbir Mollah, CSE, North South University
3. Sayada Jahan, M.Ed (Continuing), IER, University of Dhaka
4. SanjidAlom, Arabic language, University of Rajshahi

4.8. The summary of Our Interview:

Professor Dr. DaniulHaq, JU, said,

1) We need a compatible keyboard layout
2) We have spaces in Unicode where we can insert our necessary ligatures. Recently, we found almost 300 ligatures in a workshop. By conversion, the current keyboard support nearly 250 ligatures.

Professor Dr. Muhammad Zafar Ikbal, SUST, said,

1) We are working much better in Bangla Language processing than before

2) Initially, six students are conducting their Ph.D. on Bangla language processing in SUST.

Prof. Abdullah Al Mumin, SUST, said,

1) Collaboration is a must for developing BLP software.
2) Lack of motivation researcher seems less interested in this field rather than other areas.

Associate Professor Dr. Tarik Manjoor, DU, said,

1) Bijoy or Avro keyboard should develop an inbuilt speech to text converter like Google Board.
2) We need a spell and grammar checker.

Associate Professor Dr. Mamun Or Rashid, JU, said,

1) We are developing some BLP tools like Bangla Spell Checker, Bangla OCR, Machine Translator, and some other software in the Bangladesh Computer Council funded by the government.
2) If we want to increase acceptability, we have to survey on languages.

Associate Professor Nabil Mohammad, NSU, said

1) Lack of Funding opportunity is a great barrier in developing BLP software
2) Though we have written corpus to some extent, we should develop a speech corpus as well.

Assistant Professor Mashrur Imtiaz, DU, said,

1) We need collaboration in developing BLP tools.
2) We should build a national keyboard immediately.

Lecturer, Linta Islam, Jagannath University said that-

The major problem was there was no dataset for the Bangla Spell Checker. No literature reviews are available. Then, we had to start it from scratch.

Rifat Hassan Jihad, Mechanical Engineering, KHUET, said,

1) The main challenge I faced while developing a Bengali layout is to map the মহাপ্রাণ (Aspirated) & অল্পপ্রাণ (Unaspirated) phoneme according to keystroke frequency. On most of the existing designs, these types of characters are mapped under one key.
2) Again, all vowel letters have to be mapped under the Alt-Gr state to keep all keys in a uniform style. It is logical, but it is not possible to maintain the keystroke frequency efficiency in this style. All vowels can be generated with

the Hasanta key, but it is the Dead Key feature, which is not available in Linux-based OSs.

3) Mapping Bengali ligatures in a layout is kind of problematic. As the current input method system (XKB) in Linux-based OSs does not support this kind of key-mapping.

Arif Ahmed, Information science, Leading University, Sylhet, said,

1.The 1st challenge is a good dataset/corpus. Since most of the researches is now being conducted in deep learning methods, a suitable dataset is essential. We didn't have any, so we had to prepare our own, which took a lot of time.

2. 2nd challenge: computing resources. Again, deep learning algorithms need to be run on powerful machines. Luckily, our department provided us a GPU-based server for our experiments.

3. Lack of previous researches: There are not much researches done in Bangla with state-of-the-art techniques. So, we couldn't take much from the literature review.Expert linguists: Although we didn't need in-depth linguistic knowledge for our tasks, I would appreciate some help from linguists at the earlier stage of my studies.

A.S.M. Shakil Haider, the Ph.D. applicant at Texas Tech University, USA, said that

1) It would be better if spelling suggestions are available in, android keyboard software.
2) I used the ‘SHARCHAKRA’ keyboard, an android app for Bangla Typing, but the problem is that there are no spelling suggestions.

Sayada Jahan, M.Ed in Education (continuing), DU, said,

1) In Avro, while typing automatically conjuncts formed, it is time-consuming to space the target word and press backspace.
2) Bijoy Software should be open source for better acceptability.

The interview has been attached in Appendix-I

CHAPTER 5
DATA PRESENTATION

5.1. Linguistic Features of Bangla Writing:

In 1454,the printing machine was invented in Germany, and Gutenberg first printed Bible. After a century, around 1554, a printing machine was available in Goa, India. Then about 1778, it comes to Hugly, a border of Bangladesh at that time. A typewriter keyboard layout was first designed by Shahid Munir Chowdhury in 1969 under a project that started in 1965.Bangla Writing System is left to right. In Bangla, we have a total of 50 letters.

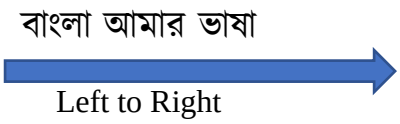

Figure 21: Bangla Writing system

Regular faces of 11 vowel alphabets: অ আ ই ঈ উ ঊ ঋ এ ঐ ও ঔ

Modified vowel sign: া ি ী ু ূ ৃ ে ৈ ো ৌ

Regular faces of 39 consonant Alphabets:

ক খ গ ঘ ঙ

চ ছ জ ঝ ঞ

ট ঠ ড ধ ণ

ত থ দ ধ ন

প ফ ব ভ ম

য র ল

শ ষ স হ

ড় ঢ় য়

ৎ ং ঃ ঁ

য-ফলা (্য)- ক্যাসেট

র-ফলা (্র) ট্র

ন-ফলা- চিহ্ন/অপরাহ্ণ

ব-ফলা- আহ্বান

ম-ফলা- সূক্ষ্ম

ল-ফলা- আহ্লাদ

5.1.1. Complexities of Bangla Letter:

Modified Consonants: Bangla is one of the most complicated scripts. that is why user-friendly software needs quality research. Hence, some complexities have been presented-

Single sound, but two or more letters- vowel letters**,** modified vowels & consonant letters

Mixed letters**:** ি/ী/ই/ঈ; ু/ূ/উ/ঊ; ঙ/ং; জ/য; শ/ষ; শ/ষ/স; শ/স; স/শ; ত্/ৎ; হ্/ঃ; ড়/ঢ়, ঋ/রি, ন/ণ; আ/য়া, য়ু/য়ূ/ঊ; য়ি/য়ী/ই

Mattra: Vowel Letter- অ আ ই ঈ উ ঊ

Consonant Letter- ক ঘ চ ছ জ ঝ ট ঠ ড ত দ নফ ব ভ ম য র ল ষ স হ ড় ঢ় য়

Half Mattra: Vowel Letter- ঋ

Consonant Letter**–** খ গ ণ থ ধ স প শ

Mattraless: Vowel Letter- এ ঐ ও ঔ

Consonant Letter**–** ঙ ঞ ৎ ং ঃ ঁ

5.1.2. Consonants and their Conjuncts of Bangla Language:

Recently, Language experts found almost 300 possible conjuncts of the Bangla language in a workshop conducted by the Bangladesh Computer Council. Most of them are presented in the table below-

ক	ক্র	খ	ঘ্র	চ	ছ্
কু	ক্র	খ্র	ঙ	চু	ছ্র
কৃ	ক্ল	খ্ল	ঙ্	চৃ	জ্র
ক্	ক্ল	খ্র	ঙ্	চ্	জ্জ
ক্ক	ক্ল্	গ	ঙ্	চ্চ	জ্ঝ
ক্কু	ক্ল্	গু	ঙ্ক	চ্চু	জ্ঝু
ক্কৃ	ক্ষ	গ্দ	ঙ্ক্ষ	চ্চৃ	জ্ঝৃ
ক্ক্	ক্ষ্	গ্ধ	ঙ্ক্	চ্চ্	জ্ঝ্
ক্ত	ক্ষ্	গ্ধু	ঙ্ক্	চ্ছ	ঞ্ঞ
ক্ট	ক্ষ্	গ্ধৃ	ঙ্খ	চ্ছু	ঞ্ঞু
ক্টু	ক্ষ্র	গ্ধ্	ঙ্গ	চ্ছৃ	ঞ্ঞৃ
ক্টৃ	ক্ত	গ্ন	ঙ্ঘ	চ্ছ্	ঞ্ঞ্
ক্ট্	ক্ত্ত	গ্ন	ঙ্ম	চ্ছ্র	ঞ্ব
ক্ন	ক্ত্র	গ্র	ঙ্ক্ষ	চ্ঞ	জ্ঝ্ব
ক্নু	ক্ত্র	গ্রু	ঙ্ক্ষু	চ্ঞু	ঝ
ক্নৃ	ক্ষ	গ্র্	ঙ্ক্ষূ	চ্ঞৃ	ঝু
ক্ন্	ক্ষ্ণ	গ্ব	ঙ্ক্ষ্	চ্ঞ্	ঝৃ
ক্ব	ক্ষ্ণ	ঘ	ঙ্ঘ্র	চ্ছ্ব	ঝ্
ক্ষ	ক্ষ্ণ	ঘ্ন	ঙ্ক্ষ	ছ	ঞ
ক্র	ক্ষ্ম	ঘ্র	ঙ্ক্ষু	ছু	ঞু
ক্রু	ক্ষ্ম্য	ঘ্রু	ঙ্ক্ষূ	ছূ	ঞৃ
ঞ্চ	ক্ষ্ম্র	ণ	ত্ত		
ঞ্চু	ক্ষ্ম্র্য	ণ্ট	ত্ত্		
ঞ্চৃ	ক্ষ্ম্র্য্য	ণ্ট্য	ত্ত্র		

ঞ্চৃ	ঠ	ণ্টূ	খ
ঞ্ছ	ঠু	ণ্টৃ	ত্ন
ঞ্ছু	ঠূ	ণ্ঠ	ত্ব
ঞ্ছূ	ঠৃ	ণ্ঠু	ত্ম
ঞ্ছৃ	ড	ণ্ঠূ	ত্ত্ব
ঞ্জ	ডু	ণ্ঠৃ	ত্র
ঞ্ঝ	ডূ	ণ্ড	ত্রু
ঞ্ঝু	ডৃ	ণ্ডু	থ
ঞ্ঝূ	ড্ড	ণ্ডূ	থ্ব
ঞ্ঝৃ	ড্ডু	ণ্ডৃ	দ
ট	ড্ডূ	ণ্ন	দ্গা
টু	ড্ডৃ	ণ্ব	দ্ঘ
টূ	ড্র	ণ্ড্র	দ্দ
টৃ	ঢ	ত	দ্ধ
ট্ট	ঢু	তু	দ্ধু
ট্টু	ঢূ	তূ	দ্ধূ
ট্টূ	ঢৃ	তৃ	দ্ধৃ
ট্টৃ	ঞ্ছ	ত্ত	দ্ভ
প্ল	ঞ্জ	ম্ব	দ্র
প্ন	ঞ্জু	ম্ভ	ত্র
প্র	ন্ধূ	ম্ম	ল্ব
প্রু	ন্ধৃ	ম্র	ল্ম
প্রূ	ব্ব	ম্প্ল	ল্ল
পূ	ব্রু	ম্ভ্র	ল্ক্র
ফু	ব্রূ	ম্নু	মৃ
ফূ	ঋ	ম্লু	দ্দ

ক্ষ্	ঞ্ছ	ব্য	শ্র
ক্ষ্র	ঞ্ছ্ব	ম্য	ক্ল
ক্ষ্ণ	ঞ্জ্ব	ম্য	ও
ক্ষ্ণ	ঞ্ঝ	র	ন্ড
ক্ষ্ণ	এ	রু	ক্ল
ক্ষ্ণ	এ্ন	রূ	শ্ব
ক্ষ্ণ	ণ্ম	ল	শ্ল
ক্ষ্ণ	স্প	ল্ক	শ্ব
ক্ষ্ণ	ড্ড	ল্ক্ষ	শ্র
ড	ড্ড্ব	ক্ষ্ণ	শ্
ঞ্	ক্ষ্ণ	ক্ষ্ণ	ব্ল
ঞ্চ	স্ক্ষ্ণ	গ্ন	ব্দ

Table 2: Bangla Conjuncts**(**After, Mustafa, Jabbar**)**

5.2. Our Selective Android Apps:

5.2.1. Subachan- Bangla Text to Speech:

It works only online. There are options like text can be inserted manuallyand .txt format document can be upload for text to speech conversion. Some features of these tools are as follows:

Version: 1.0

Updated on: 5 January 2019

offered by: Technext Limited

Developer Contact: technext.limited@gmail.com

Letter insert: ঙ, ঞ, ষ, র, ড়, ঢ়, ৎ, ং, ঁ, ঃ, জ্ঞ, ঞ্চ, ক্ষ

output: We found only (র)for this three letters- র, ড়, ঢ়

'ক্ষ' is pronounced like 'খ'

We did not find any pronunciation of ঙ, ঞৎ, ং, ঁ, ঃ, জ্ঞ, ঞ্চ

Word insert: ছত্রিশ, একাত্তর, একক, গারো, গাড়, গাঢ়, বাহ্য, বাহ্যেন্দ্রিয়, বাহ্যদৃষ্টি, দৃষ্টি, সৃষ্টি

Output (Pronunciation): (ছোত্রিশ্), (একাত্তোর্), (অ্যাকক্), (গারো), (গারো), (গারো), (বাজ্জো), (অঙ্) (বাজ্জোদৃশি), (দৃশি), (সৃশি)

Sentence input: 'বঙ্গবন্ধু আমাদের জাতির পিতা'

Output: The sentences mentioned above pronounced with echo

Some Reviews from comment sections:

1) Cannot detect text from the image (Sarkar Kumar Mondal)
2) Did not work. Just sound like GaGaGa... (Md. Delwar Hosen)

5.2.2. Bangla Speech to Text Software:

Software Name: Bangla Voice to Text

Word Pronunciation:গার, গাড়, গাঢ়, বারি, শব

We pronounced /garo/, /gar^{h}o/, /ga$\mathfrak{t}^{h}$o/, but we get output only /garo/

5.3. Software tools for Windows 10/10 Pro:

There are many online tools for BLP. The following tools are used for windows 10/10 Pro-

5.3.1. Google Translation:

In Google translate software, there is an option for pronunciation. Pronunciation or letter name is not available of Some Bangla Letters like-ঞ, ং, ঁ, ঃ, and the faint sound of some conjuncts likeষ্ণ, ট্ট, ত্ত, ক্ত, দ্ধ, etc.

There is similar pronunciation available like /ra/ for the different letter, including র, ড়, ঢ়

Bangla to English Translation: ‘গরু/গোরু’ or ‘গড়ু/গোড়ু’ has been translated ‘cows’ in English

5.3.2. Fonts/Typeface (Not Unicode Supported):

We presented only SutonnyMJ 3 types, including Normal, Italic, and Bold typeface, to understand how the typeface looks. So, the rest of the fonts are presented only in regular typeface. All fonts are typed with 14 points size.

5.3.2.1. Normal Typeface of SutonnyMJ

ক্ত: ক + ত = ক্ত	ন্তু: ন + ত + উ = ন্তু	হু: হ + উ = হু
ক্র: ক + র = ক্র	ন্ধ: ন + ধ = ন্ধ	হূ: হ + ঊ = হূ
ক্রু: ক + র + উ = ক্রু	ন্ন: ন + ম = ন্ন	হৃ: হ + ঋ = হৃ
ক্ষ: ক + ষ = ক্ষ	ন্নু: ন + ম + উ = ন্নু	হ্ণ: হ + ণ = হ্ণ
ক্স: ক + স = ক্স	ন্স: ন + স = ন্স	হ্ন: হ + ন = হ্ন
গু: গ + উ = গু	প্স: প + স = প্স	হ্ম: হ + ম = হ্ম
গ্ধ: গ + ধ = গ্ধ	ব্ধ: ব + ধ = ব্ধ	ড়: ড় + ় = ড়
ঙ্ক: ঙ + ক = ঙ্ক	ব্লু: ব + ল + উ = ব্লু	ড়ু: ড় + উ = ড়ু
ঙ্গ: ঙ + গ = ঙ্গ	ভ্র: ভ + র = ভ্র	ঢ়ু: ঢ় + উ = ঢ়ু
জ্ঞ: জ + ঞ = জ্ঞ	ভ্রু: ভ + র + উ = ভ্রু	ত্রু: ত + র + উ = ত্রু
ঞ্চ: ঞ + চ = ঞ্চ	ভ্রূ: ভ + র + ঊ = ভ্রূশ্চ	দ্ধ: দ + ধ = দ্ধ
ঞ্জ: ঞ + জ = ঞ্জ	ম্ন: ম + ন = ম্ন	দ্রু: দ + র্ + উ = দ্রু
ট্ট: ট + ট = ট্ট	ম্ফ: ম + ফ = ম্ফ	দ্রূ: দ + র্ + ঊ = দ্রূ
ণ্ঠ: ণ + ঠ = ণ্ঠ	ম্ম: ম + ম = ম্ম	ন্ঠ: ন + ঠ = ন্ঠ
ণ্ড: ণ + ড = ণ্ড	রু: র + উ = রু	ন্ড: ন + ড = ন্ড
ত্ত: ত + ত = ত্ত	রূ: র + ঊ = রূ	শ্রূ: শ + র + ঊ = শ্রূ
ত্থ: ত + থ = ত্থ	ল্ড: ল + ড = ল্ড	ষ্ণ: ষ + ণ = ষ্ণ
ত্ন: ত + ন = ত্ন	শু: শ + উ = শু	ষ্ম: ষ + ম = ষ্ম
ত্ম: ত + ম = ত্ম	শ্চ: শ + চ = শ্চ	স্ট: স + ট = স্ট
ত্র: ত + র = ত্র	শ্রু: শ + র + উ = শ্রু	স্তু: স + ত + উ = স্তু
		স্থ: স + থ = স্থ

5.3.2.2. Italic Typeface of SutonnyMJ:

হু: হ + উ = হু	ক্ত: ক + ত = ক্ত	দ্রু: দ + র্ + উ = দ্রু
হূ: হ + ঊ = হূ	ক্র: ক + র = ক্র	দ্রূ: দ + র্ + ঊ = দ্রূ
হৃ: হ + ঋ = হৃ	ক্রু: ক + র + উ = ক্রু	ন্ঠ: ন + ঠ = ন্ঠ

হ্ণ: হ + ণ = হ্ণ	*ক্ষ: ক + ষ = ক্ষ*	ন্ড: ন + ড = ন্ড
হ্ন: হ + ন = হ্ন	*ক্স: ক + স = ক্স*	ন্তু: ন + ত + উ = ন্তু
হ্ম: হ + ম = হ্ম	*গু: গ + উ = গু*	ন্ধ: ন + ধ = ন্ধ
ড়ঁ: ড় + ঁ = ড়ঁ	*গ্ধ: গ + ধ = গ্ধ*	ন্ম: ন + ম = ন্ম
ড়ু: ড় + উ = ড়ু	*ঙ্ক: ঙ + ক = ঙ্ক*	ন্মু: ন + ম + উ = ন্মু
ঢ়ু: ঢ় + উ = ঢ়ু	*ঙ্গ: ঙ + গ = ঙ্গ*	ন্স: ন + স = ন্স
স্ট: স + ট = স্ট	*জ্ঞ: জ + ঞ = জ্ঞ*	প্স: প + স = প্স
স্তু: স + ত + উ = স্তু	*ঞ্চ: ঞ + চ = ঞ্চ*	ব্ধ: ব + ধ = ব্ধ
স্থ: স + থ = স্থ	*ঞ্জ: ঞ + জ = ঞ্জ*	ব্লু: ব + ল + উ = ব্লু
ড়ু: ড় + উ = ড়ু	*ট্ট: ট + ট = ট্ট*	ভ্র: ভ + র = ভ্র
ঢ়ু: ঢ় + উ = ঢ়ু	*ণ্ঠ: ণ + ঠ = ণ্ঠ*	ভ্রু: ভ + র + উ = ভ্রু
প্স: প + স = প্স	*ণ্ড: ণ + ড = ণ্ড*	ভ্রূ: ভ + র + ঊ = ভ্রূশ্চ
ব্ধ: ব + ধ = ব্ধ	*ত্ত: ত + ত = ত্ত*	ম্ন: ম + ন = ম্ন
ব্লু: ব + ল + উ = ব্লু	*ত্থ: ত + থ = থ*	ম্ফ: ম + ফ = ম্ফ
ভ্র: ভ + র = ভ্র	*ত্ন: ত + ন = ত্ন*	ম্ম: ম + ম = ম্ম
ভ্রু: ভ + র + উ = ভ্রু	*ত্ম: ত + ম = ত্ম*	রু: র + উ = রু
ভ্রূ: ভ + র + ঊ = ভ্রূশ্চ	*ত্র: ত + র = ত্র*	রূ: র + ঊ = রূ
ম্ন: ম + ন = ম্ন	*ত্রু: ত + র + উ = ত্রু*	ল্ড: ল + ড = ল্ড
ম্ফ: ম + ফ = ম্ফ	*দ্ধ: দ + ধ = দ্ধ*	শু: শ + উ = শু
ম্ম: ম + ম = ম্ম	*দ্রু: দ + র্ + উ = দ্রু*	শ্চ: শ + চ = শ্চ
রু: র + উ = রু	*দ্রূ: দ + র্ + ঊ = দ্রূ*	শ্রু: শ + র + উ = শ্রু
রূ: র + ঊ = রূ	*ণ্ঠ: ন + ঠ = ণ্ঠ*	শ্রূ: শ + র + ঊ = শ্রূ
ল্ড: ল + ড = ল্ড	*হু: হ + উ = হু*	ষ্ণ: ষ + ণ = ষ্ণ
শু: শ + উ = শু	*হূ: হ + ঊ = হূ*	ষ্ম: ষ + ম = ষ্ম
শ্চ: শ + চ = শ্চ	*হৃ: হ + ঋ = হৃ*	*ন্স: ন + স = ন্স*
শ্রু: শ + র + উ = শ্রু	*হ্ণ: হ + ণ = হ্ণ*	*স্ট: স + ট = স্ট*
শ্রূ: শ + র + ঊ = শ্রূ	*হ্ন: হ + ন = হ্ন*	*স্তু: স + ত + উ = স্তু*
ষ্ণ: ষ + ণ = ষ্ণ	*হ্ম: হ + ম = হ্ম*	
ষ্ম: ষ + ম = ষ্ম	*ড়ঁ: ড় + ঁ = ড়ঁ*	*স্থ: স + থ = স্থ*

5.3.2.3. Bold Typeface of SutonnyMJ:

ক্তঃ ক + ত = ক্ত	অঃ ত + ম = ত্ম	শ্চঃ শ + চ = শ্চ
ক্রঃ ক + র = ক্র	ত্রঃ ত + র = ত্র	শ্রুঃ শ + র + উ = শ্রু
ক্রুঃ ক + র + উ = ক্রু	ত্রুঃ ত + র + উ = ত্রু	শ্রূঃ শ + র + ঊ = শ্রূ
ক্ষঃ ক + ষ = ক্ষ	দ্ধঃ দ + ধ = দ্ধ	ষ্ণঃ ষ + ণ = ষ্ণ
ক্সঃ ক + স = ক্স	দ্রুঃ দ + র্ + উ = দ্রু	ষ্মঃ ষ + ম = ষ্ম
গুঃ গ + উ = গু	দ্রূঃ দ + র্ + ঊ = দ্রূ	স্টঃ স + ট = স্ট
গ্ধঃ গ + ধ = গ্ধ	ষ্ঠঃ ন + ঠ = ষ্ঠ	স্তুঃ স + ত + উ = স্তু
ঙ্কঃ ঙ + ক = ঙ্ক	ন্ডঃ ন + ড = ন্ড	স্থঃ স + থ = স্থ
ঙ্গাঃ ঙ + গ = ঙ্গ	স্তুঃ ন + ত + উ = স্তু	হুঃ হ + উ = হু
জ্ঞঃ জ + ঞ = জ্ঞ	ন্ধঃ ন + ধ = ন্ধ	হূঃ হ + ঊ = হূ
ঞ্চঃ ঞ + চ = ঞ্চ	ন্মঃ ন + ম = ন্ম	হৃঃ হ + ঋ = হৃ
ঞ্জঃ ঞ + জ = ঞ্জ	ন্মুঃ ন + ম + উ = ন্মু	হ্নঃ হ + ণ = হ্ন
ট্টঃ ট + ট = ট্ট	ন্সঃ ন + স = ন্স	হ্নঃ হ + ন = হ্ন
ণ্ঠঃ ণ + ঠ = ণ্ঠ	প্সঃ প + স = প্স	হ্মঃ হ + ম = হ্ম
ণ্ডঃ ণ + ড = ণ্ড	ব্ধঃ ব + ধ = ব্ধ	ড়ুঃ ড় + ু = ড়ু
ত্তঃ ত + ত = ত্ত	ব্লুঃ ব + ল + উ = ব্লু	ড়ুঃ ড় + উ = ড়ু
ত্থঃ ত + থ = ত্থ	ভ্রঃ ভ + র = ভ্র	ঢ়ুঃ ঢ় + উ = ঢ়ু
ত্নঃ ত + ন = ত্ন	ভ্রুঃ ভ + র + উ = ভ্রু	
রুঃ র + উ = রু	ভ্রূঃ ভ + র + ঊ = ভ্রূশ্চ	
রূঃ র + ঊ = রূ	ম্নঃ ম + ন = ম্ন	
ল্ডঃ ল + ড = ল্ড	ম্ফঃ ম + ফ = ম্ফ	
শুঃ শ + উ = শু	ম্মঃ ম + ম = ম্ম	

5.3.3.Typeface of different fonts (Unicode Supported):

There are many Unicode fonts available, but frequently used fonts are selected. Some conjuncts have been shown in different Unicode fonts below.

5.3.3.1. Normal Typeface of Kalpurush:

ক্র: ক + র = ক্র	ত্র: ত + র = ত্র	ভ্রু: ভ + র + ু = ভ্রু
ক্রু: ক + র + ু = ক্রু	ক্র: ত + র + ু = ক্র	ভ্রূ: ভ + র + ূ = ভ্রূ
ক্ষ: ক + ষ = ক্ষ	দ্ধ: দ + ধ = দ্ধ	ম্ন: ম + ন = ম্ন
ক্স: ক + স = ক্স	দ্রু: দ +র্ + ু = দ্রু	ম্ফ: ম + ফ = ম্ফ
গু: গ + ু = গু	দ্রূ: দ +র্ + ূ = দ্রূ	ম্ম: ম + ম = ম্ম
গ্ধ: গ + ধ = গ্ধ	ন্ঠ: ন + ঠ = ন্ঠ	রু: র + ু = রু
ঙ্ক: ঙ + ক = ঙ্ক	ন্ড: ন + ড = ন্ড	রূ: র + ূ = রূ
ঙ্গ: ঙ + গ = ঙ্গ	ন্ত: ন + ত + ু = ন্তু	ল্ড: ল + ড = ল্ড
জ্ঞ: জ + ঞ = জ্ঞ	ন্ধ: ন + ধ = ন্ধ	শু: শ + ু = শু
ঞ্চ: ঞ + চ = ঞ্চ	ন্ম: ন + ম = ন্ম	শ্চ: শ + চ = শ্চ
ঞ্জ: ঞ + জ = ঞ্জ	ন্মু: ন + ম + ু = ন্মু	শ্রু: শ + র + ু = শ্রু
ট্ট: ট + ট = ট্ট	ন্স: ন + স = ন্স	শ্রূ: শ + র + ূ = শ্রূ
ণ্ঠ: ণ + ঠ = ণ্ঠ	প্স: প + স = প্স	ষ্ণ: ষ + ণ = ষ্ণ
ণ্ড: ণ + ড =ণ্ড	ব্ধ: ব + ধ = ব্ধ	ষ্ম: ষ + ম = ষ্ম
ত্ত: ত + ত = ত্ত	ব্লু: ব + ল + ু = ব্লু	স্ট: স + ট = স্ট
থ্থ: ত + থ = থ্থ	ভ্র: ভ + র = ভ্র	
ত্ন: ত + ন = ত্ন		

5.3.3.2. Normal Typeface of SolaimaniLipi:

স্তু: স + ত + ঊ = স্তু	ক্ত: ক + ত = ক্ত	ঞ্জ: ঞ + জ = ঞ্জ
স্থ: স + থ = স্থ	ক্র: ক + র = ক্র	ট্ট: ট + ট = ট্ট
হু: হ + উ = হু	ক্রু: ক + র + উ = ক্রু	ণ্ঠ: ণ + ঠ = ণ্ঠ
হূ: হ + ঊ = হূ	ক্ষ: ক + ষ = ক্ষ	ণ্ড -: ণ + ড = ণ্ড
হৃ: হ + ঋ = হৃ	ক্স: ক + স = ক্স	ত্ত: ত + ত = ত্ত
হ্ণ: হ + ণ = হ্ণ	গু: গ + উ = গু	থ: ত + থ = থ
হ্ন: হ + ন = হ্ন	গ্ধ: গ + ধ = গ্ধ	তন: ত + ন = ত্ন
হ্ম: হ + ম = হ্ম	ঙ্ক: ঙ + ক = ঙ্ক	তম: ত + ম = ত্ম
ড়: ড় + ◌ = ড়	ঙ্গ: ঙ + গ = ঙ্গ	ত্র: ত + র = ত্র
ড়–: ড় + উ = ড়ু	জ্ঞ: জ + ঞ = জ্ঞ	ক্র: ত + র + উ = ক্র
ঢ় ◌ু: ঢ় + উ = ঢ়	ঞ্চ: ঞ + চ = ঞ্চ	দ্ধ: দ + ধ = দ্ধ
ন্ঠ: ন + ঠ = ন্ঠ	ভ্র: ভ + র + ঊ = ভ্র	দ্রু: দ +র্ + উ = দ্রু
ন্ড: ন + ড = ন্ড	ম্ন: ম + ন = ম্ন	দ্র: দ +র্ + ঊ = দ্র
ন্তু: ন + ত + উ = ন্তু	ম্ফ: ম + ফ = ম্ফ	স্ট: স + ট = স্ট
ন্ধ: ন + ধ = ন্ধ	মম: ম + ম = ম্ম	স্তু: স + ত + উ = স্তু
ন্ম: ন + ম = ন্ম	রু: র + উ = রু	স্থ: স + থ = স্থ
ন্ম ◌ু: ন + ম + উ = ন্মু		হু: হ + উ = হু
ন্স: ন + স = ন্স		

প্স: প + স = প্স ব্ধ: ব + ধ = ব্ধ ব্লু: ব + ল + ঊ = ব্লু ভ্র: ভ + র = ভ্র ভ্রু: ভ + র + ঊ = ভ্রু	রূ: র + ঊ = রূ ল্ড: ল + ড = ল্ড শু: শ + উ = শু শ্চ: শ + চ = শ্চ শ্রু: শ + র + ঊ = শ্রু শ্রূ: শ + র + ঊ = শ্রূ ষণ: ষ + ণ = ষ্ণ ষ্ম: ষ + ম = ষ্ম	হূ: হ + ঊ = হূ হৃ: হ + ঋ = হৃ হ্ণ: হ + ণ = হ্ণ হ্ন: হ + ন = হ্ন হ্ম: হ + ম = হ্ম ড়্: ড় + ্ = ড়্ ড়–: ড় + ঊ = ড়ু ঢ়ু: ঢ় + ঊ = ঢ়ু

5.3.3.3. Normal Typeface of AdorshoLipi:

ক্ত: ক + ত = ক্ত ক্র: ক + র = ক্র ক্রু: ক + র + ঊ = ক্রু ক্ষ: ক + ষ = ক্ষ ক্স: ক + স = ক্স গু: গ + উ = গু গ্ধ: গ + ধ = গ্ধ ঙ্ক: ঙ + ক = ঙ্ক ঙ্গ: ঙ + গ = ঙ্গ ভ্র: ভ + র + ঊ = ভ্র	জ্ঞ: জ + ঞ = জ্ঞ ঞ্চ: ঞ + চ = ঞ্চ ঞ্জ: ঞ + জ = ঞ্জ ট্ট: ট + ট = ট্ট ণ্ঠ: ণ + ঠ = ণ্ঠ ণ্ড: ণ + ড =ণ্ড ত্ত: ত + ত = ত্ত ত্থ: ত + থ = ত্থ ত্ন: ত + ন =ত্ন শ্রু: শ + র + উ = শ্রু শ্রূ: শ + র + ঊ	ত্ম: ত + ম = ত্ম ত্র: ত + র = ত্র ত্রু: ত + র + উ = ত্রু দ্ধ: দ + ধ = দ্ধ দ্রু: দ +র্ + উ = দ্রু দ্রূ: দ +র্ + ঊ = দ্রূ ন্ঠ: ন + ঠ = ন্ঠ ন্ড: ন + ড = ন্ড ন্ত: ন + ত + উ

ম্ন: ম + ন =ম্ন	= শ্ৃ	= ষ্ত্ত
স্ফ: ম + ফ = স্ফ	ষ্ণ: ষ + ণ = ষ্ণ	
মম: ম + ম = ম্ম	ষ্ম: ষ + ম = ষ্ম	
রু: র + উ = রু	স্ট: স + ট = স্ট	
রূ: র + ঊ = রূ	স্ত্ত: স + ত + উ	
ল্ড: ল + ড = ল্ড	= স্ত্ত	
শু: শ + উ = শু	স্থ: স + থ = স্থ	
শ্চ: শ + চ = শ্চ	হু: হ + উ = হু	
	হূ: হ + ঊ = হূ	

5.3.3.4. Normal Typeface of NikoshLipi:

হৃ: হ + ঋ = হৃ	ক্ত: ক + ত = ক্ত	দ্ধ: দ + ধ = দ্ধ
হ্ণ: হ + ণ = হ্ণ	ক্র: ক + র = ক্র	দ্রু: দ +র্ + উ = দ্রু
হ্ন: হ + ন = হ্ন	ক্রু: ক + র + উ = ক্রু	দ্রূ: দ +র্ + ঊ = দ্রূ
হ্ম: হ + ম = হ্ম	ক্ষ: ক + ষ = ক্ষ	ষ্ঠ: ন + ঠ = ষ্ঠ
ড়: ড় + ◌ = ড়	ক্স: ক + স = ক্স	ন্ড: ন + ড = ন্ড
ড়—: ড় + উ = ড়ু	গু: গ + উ = গু	ন্ত্ত: ন + ত + উ = ন্ত্ত
ঢ় ◌ু: ঢ় + উ = ঢ়ু	গ্ধ: গ + ধ = গ্ধ	ন্ধ: ন + ধ = ন্ধ
ন্ম: ন + ম = ন্ম	ঙ্ক: ঙ + ক = ঙ্ক	ম্ম: ম + ম = ম্ম
ন্ম ◌ু: ন + ম + উ = ন্মু	ভ্রু: ভ + র + উ = ভ্রু	রু: র + উ = রু
ন্স: ন + স = ন্স	ভ্রূ: ভ + র + ঊ = ভ্রূ	রূ: র + ঊ = রূ
প্স: প + স = প্স	ম্ন: ম + ন = ম্ন	
ব্ধ: ব + ধ = ব্ধ	স্ফ: ম + ফ = স্ফ	

ব্ল ু: ব + ল + উ = ব্লু ভ্র: ভ + র = ভ্র		

5.3.3.5. NormalTypeface of Sagar Lipi:

ল্ড: ল + ড = ল্ড	ক্রু: ক + র + উ = ক্রু	দ্ধ: দ + ধ = দ্ধ
শু: শ + উ = শু	ক্ষ: ক + ষ = ক্ষ	দ্রু: দ +র্ + উ = দ্রু
শ্চ: শ + চ = শ্চ	ক্স: ক + স = ক্স	দ্রূ: দ +র্ + ঊ = দ্রূ
শ্রু: শ + র + উ = শ্রু	গু: গ + উ = গু	ন্ঠ: ন + ঠ = ন্ঠ
শ্রূ: শ + র + ঊ = শ্রূ	গ্ধ: গ + ধ = গ্ধ	ন্ড: ন + ড = ন্ড
ষ্ণ: ষ + ণ = ষ্ণ	ঙ্ক: ঙ + ক = ঙ্ক	ন্তু: ন + ত + উ = ন্তু
ষ্ম: ষ + ম = ষ্ম	ঙ্গ: ঙ + গ = ঙ্গ	ন্ধ: ন + ধ = ন্ধ
স্ট: স + ট = স্ট	জ্ঞ: জ + ঞ = জ্ঞ	ন্ম: ন +ম = ন্ম
স্তু: স + ত + উ = স্তু	ঞ্চ: ঞ + চ = ঞ্চ	ন্ম ু: ন+ ম + উ = ন্মু
স্থ: স + থ = স্থ	ঞ্জ: ঞ + জ = ঞ্জ	ন্স: ন + স = ন্স
হু: হ + উ = হু	ট্ট: ট + ট = ট্ট	প্স: প + স = প্স
হূ: হ + ঊ = হূ	ণ্ঠ: ণ + ঠ = ণ্ঠ	ব্ধ: ব + ধ = ব্ধ
হৃ: হ + ঋ = হৃ	ণ্ড: ণ + ড = ণ্ড	ব্ল ু: ব + ল + উ = ব্লু
হ্ন: হ + ণ = হ্ন	ৎত: ত + ত = ৎত	ভ্র: ভ + র = ভ্র
হ্ন: হ + ন = হ্ন	ৎথ: ত + থ = ৎথ	ভ্রু: ভ + র + উ = ভ্রু
হ্ম: হ + ম = হ্ম	ত্ন: ত + ন = ৎন	ভ্রূ: ভ + র + ঊ = ভ্রূ
ড়: ড় + ি = ড়ি	ত্ম: ত + ম = ৎম	ম্ন: ম + ন =ম্ন
ড়ু: ড় + উ = ড়ু	ক্ত: ক + ত = ক্ত	ত্র: ত + র = ত্র
ঢ় ু: ঢ় + উ = ঢ়ু	ক্র: ক + র = ক্র	ক্র: ত + র + উ = ক্র

5.3.3.6. Normal Typeface of SutonnyUniBanglaOMJ:

শু: শ + উ = শু	হু: হ + উ = হু	স্ফ: ম + ফ = স্ফ
শ্চ: শ + চ = শ্চ	হূ: হ + ঊ = হূ	ম্ম: ম + ম = ম্ম
শ্রু: শ + র + উ = শ্রু	হৃ: হ + ঋ = হৃ	রু: র + উ = রু
শ্রূ: শ + র + ঊ = শ্রূ	হ্ণ: হ + ণ = হ্ণ	রূ: র + ঊ = রূ
ষ্ণ: ষ + ণ = ষ্ণ	হ্ন: হ + ন = হ্ন	ল্ড: ল + ড = ল্ড
ষ্ম: ষ + ম = ষ্ম	হ্ম: হ + ম = হ্ম	শু: শ + উ = শু
স্ট: স + ট = স্ট	ড়্: ড় + ্ = ড়্	শ্চ: শ + চ = শ্চ
স্তু: স + ত + উ = স্তু	ড়ু: ড় + উ = ড়ু	শ্রু: শ + র + উ = শ্রু
স্থ: স + থ = স্থ	ঢ় ু: ঢ় + উ = ঢ়ু	শ্রূ: শ + র + ঊ = শ্রূ
হ্ম: হ + ম = হ্ম	ঢ় ূ: ঢ় + ঊ =	ষ্ণ: ষ + ণ = ষ্ণ
ড়্: ড় + ্ = ড়্	ক্ত: ক + ত = ক্ত	ষ্ম: ষ + ম = ষ্ম
ড়ু: ড় + উ = ড়ু	ক্র: ক + র = ক্র	স্ট: স + ট = স্ট
ক্রু: ক + র + উ = ক্রু	ৎথ: ত + থ = ৎথ	স্তু: স + ত + উ = স্তু
ক্ষ: ক + ষ = ক্ষ	ৎন: ত + ন =ৎন	স্থ: স + থ = স্থ
ক্স: ক + স = ক্স	ৎম: ত + ম = ৎম	ন্ড: ন + ড = ন্ড
গু: গ + উ = গু	ত্র: ত + র = ত্র	ন্তু: ন + ত + উ = ন্তু
গ্ধ: গ + ধ = গ্ধ	ক্রু: ত + র + উ = ক্রু	ন্ধ: ন + ধ = ন্ধ
হু: হ + উ = হু	দ্ধ: দ + ধ = দ্ধ	ন্ম: ন + ম = ন্ম
হূ: হ + ঊ = হূ	দ্রু: দ +র্ + উ = দ্রু	ন্ম ু: ন + ম + উ = ন্মু
হৃ: হ + ঋ = হৃ	দ্রূ: দ +র্ + ঊ = দ্রূ	ন্স: ন + স = ন্স
হ্ণ: হ + ণ = হ্ণ	ন্ঠ: ন + ঠ = ন্ঠ	
প্স: প + স = প্স	ভ্র: ভ + র = ভ্র	
ঙ্ক: ঙ + ক = ঙ্ক	ভ্রু: ভ + র + উ = ভ্রু	
ঙ্গ: ঙ + গ = ঙ্গ	ভ্রূ: ভ + র + ঊ =ভ্রূ	
জ্ঞ: জ + ঞ = জ্ঞ	ম্ন: ম + ন =ম্ন	
ঞ্চ: ঞ + চ = ঞ্চ	স্ফ: ম + ফ = স্ফ	
ঞ্জ: ঞ + জ = ঞ্জ	মম: ম + ম = ম্ম	

ট্ট: ট + ট = ট্ট ণ্ঠ: ণ + ঠ = ণ্ঠ ঙ্গ: ণ + ড = ঙ্গ ত্ত: ত + ত = ত্ত হ্ন: হ + ন = হ্ন ব্ধ: ব + ধ = ব্ধ ব্ল ু: ব + ল + ঊ = ব্লু	রু: র + উ = রু রূ: র + ঊ = রূ ল্ড: ল + ড = ল্ড	

5.4. Fonts size Comparision:

There are many Bangla Fonts, including ASCII and UNICODE, but they are different sizes in 14 points as we can see in the following table

Font Name	Text in 14 point
SutonnyMJ (ASCII)	ক্ত: ক + ত = ক্ত
Kalpurush (Unicode)	ক্ত: ক + ত = ক্ত
Nikosh (Unicode)	ক্ত: ক + ত = ক্ত
AdorshoLipi: (Unicode)	ক্ত: ক + ত = ক্ত
Sagar Lipi (Unicode)	ক্ত: ক + ত = ক্ত

Table 3: Fonts comparisons in 14 points

5.5. Bijoy to Unicode, and Unicode to Bijoy Converter:

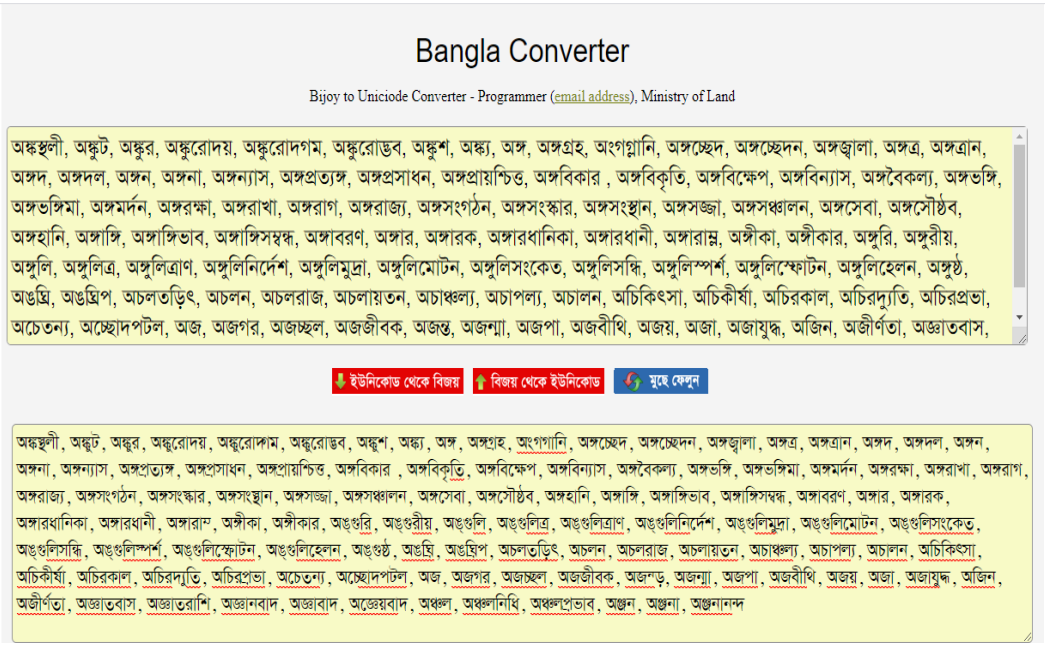

Figure 22: Bijoy to Unicode Converter

Figure 23: Unicode to Bijoy Converter

5.6. IPA Transcription Website:

Some words transcription in IPA have been shown in figure 24, 25, and 26, respectively. This tool has been developed by the Bangladesh computer council.

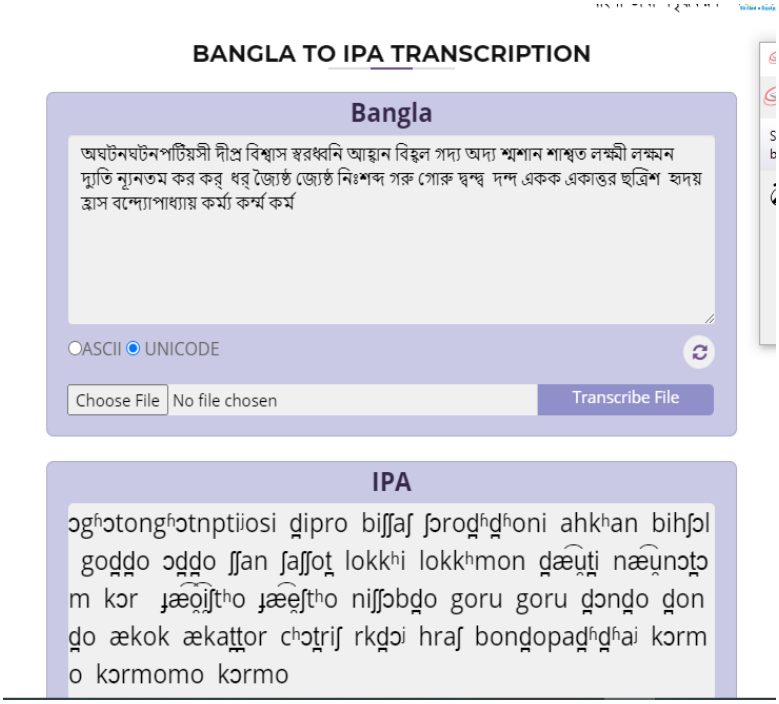

Figure 24: IPA transcription of some Bangla words

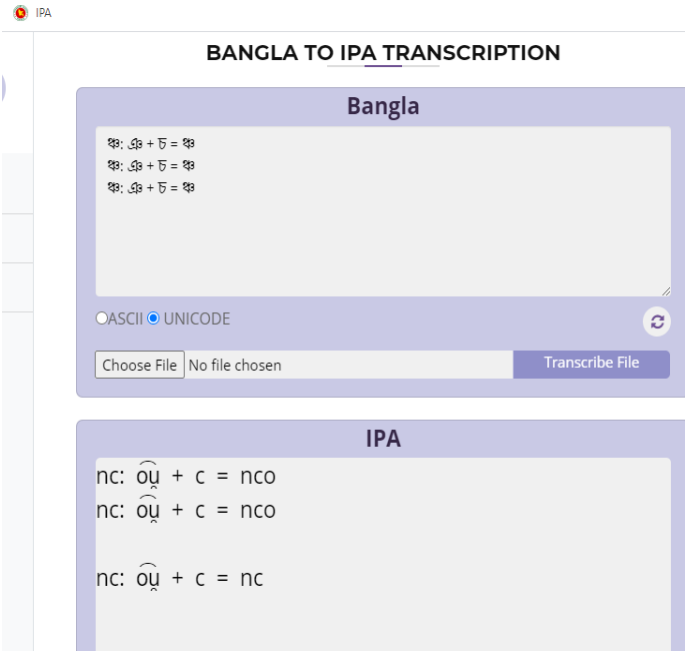

Figure 25: IPA transcription of some Bangla words in different fonts

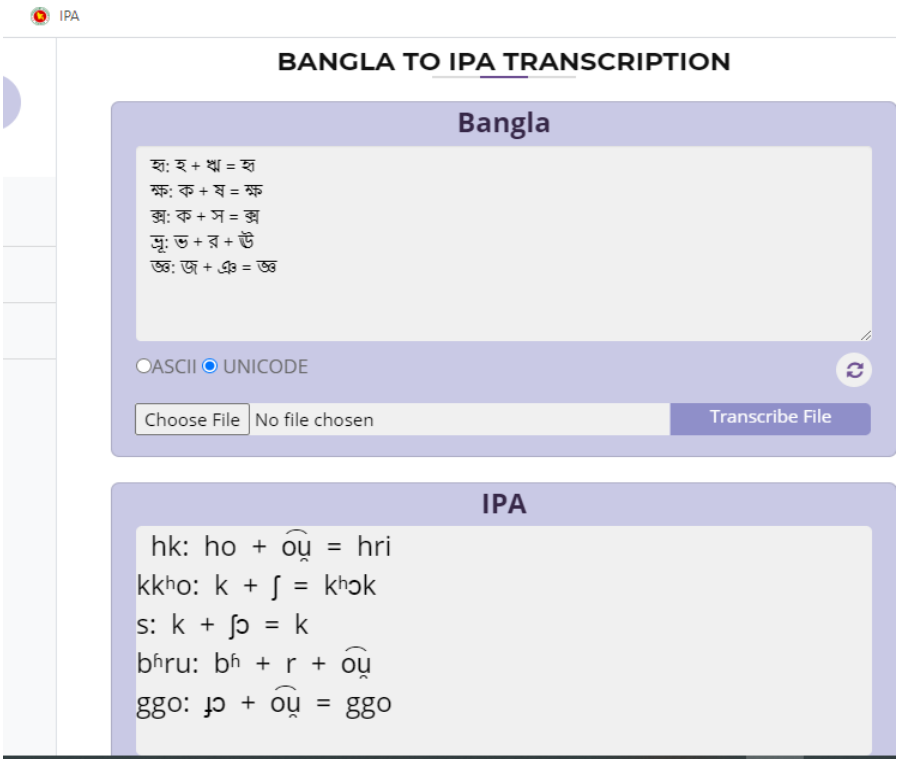

Figure 26: IPA transcription of some conjuncts

5.7. Online Speech to Text Converter:

Speech Input(https://speechtyping.com/voice-typing/speech-to-text-bengali**)**	
Input	**output**
বাস	বাস
বাঁশ	বাস
নিঃক্ষত্রিয়	ক্ষত্রিয়
খড়ক	খরক
একাত্তর	71

গার	গার
গাড়	গার
গাঢ়	গার
বৃষ্টির বারি	বৃষ্টির বাড়ি
বারি পড়ছে	বাড়ি পড়ছে
লোকটি বিষ খেল	লোক t২০ খেল

Table 4: Pronunciation of Some words and phrases input into Speech to Text converter apps

	Google Translate (Speech to text)
Input	**Output**
কলরোল	কলরব
গার	গার
গাড়	গার
গাঢ়	গার
বাঁশ	বাস
বৃষ্টির বারি	বৃষ্টির বাড়ি
বারি পড়ছে	বাড়ি পড়ছে

লোকটি বিষ খেল	লোকটি ব্রিজ খেলা
খড়ক	খরক
একাত্তর	61

Table 5: Pronunciation of Some words and phrases input into Google translate

5.8. Machine Translator:

Google translate	
Input	**Output**
বল বীর বল উন্নত মম শীর	Ball hero ball advanced mum shir

Table 6: Sentences input into Google Translate

5.9. Avro Spell Checker:

All words without 'একাডেমি 'have been misspelled in the software, for example, 'একাডেমি, ইতপুরবে, মহিয়সি, দূতি, অকস্মাত, প্রতিযোগীটা, দন্দ, সাধিনতা' had been input in Avro spell checker. The correct spelling of the mentioned words is- 'একাডেমি, মহীয়সী, দূতী, অকস্মাৎ, প্রতিযোগিতা, দ্বন্দ্ব, স্বাধীনতা'. The results are shown in figure 27, 28, 29, 30, and 31, respectively.

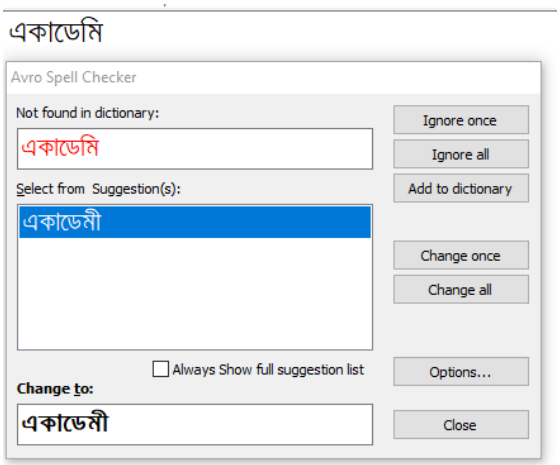

Figure 27: ‘একাডেমি’ word had been input in Avro Spell Checker

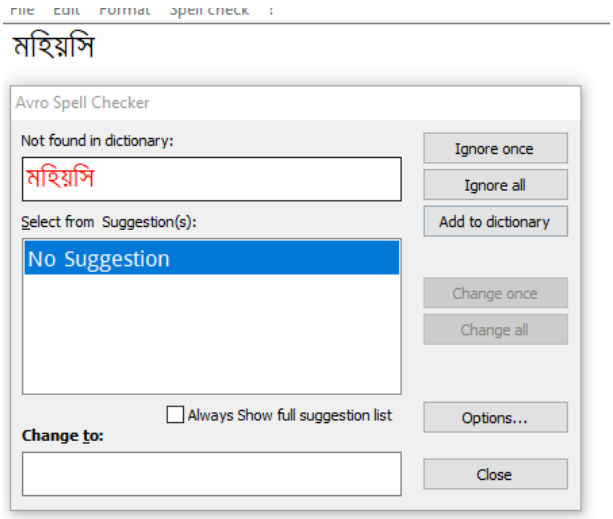

Figure 28: ‘মহিয়সি’ word had been input in Avro Spell Checker

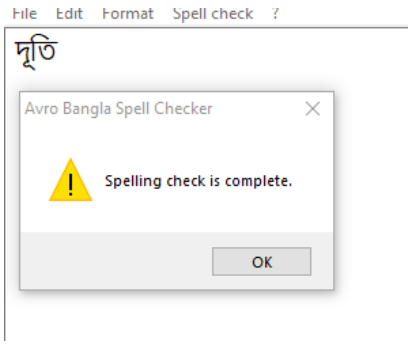

Figure 29: ‘দূতি’ word had been input in Avro Spell Checker

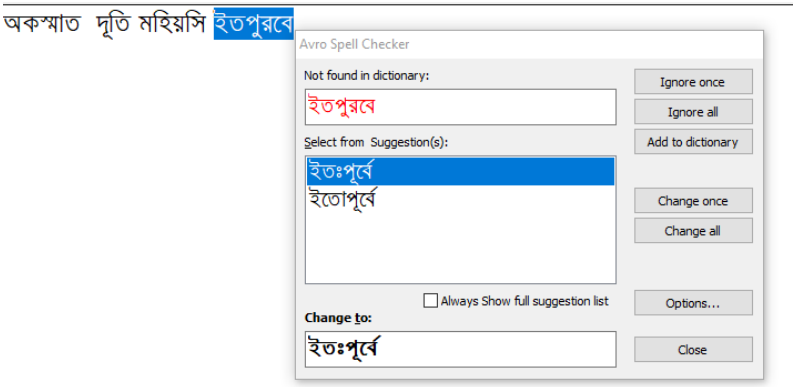

Figure 30: 'ইতপুরবে' word had been input in Avro Spell Checker

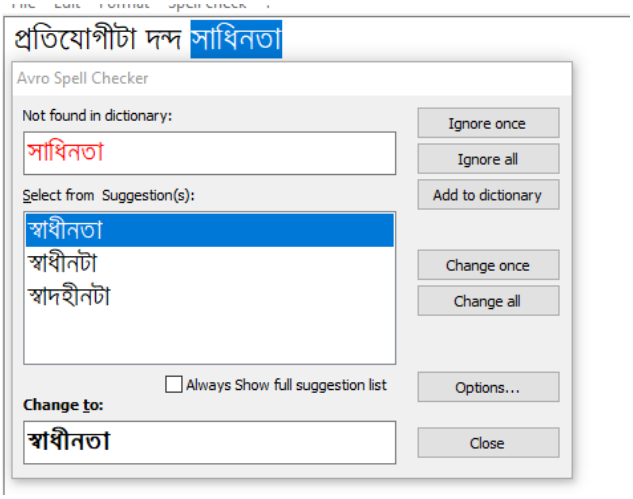

Figure 31: 'স্বাধীনতা' word had been input in Avro Spell Checker

5.10. Online Bangla Spelling Software:

Some misspelled words have been input for spell checking in Online Bangla Software. The misspelled words are: 'বন্দোপাধ্যায়, নুন্যতম, জৈষ্ঠ্য, জেষ্ঠ্য, চত্তর, অপরাহ্ন, অংক, গিতাঞ্জলি'

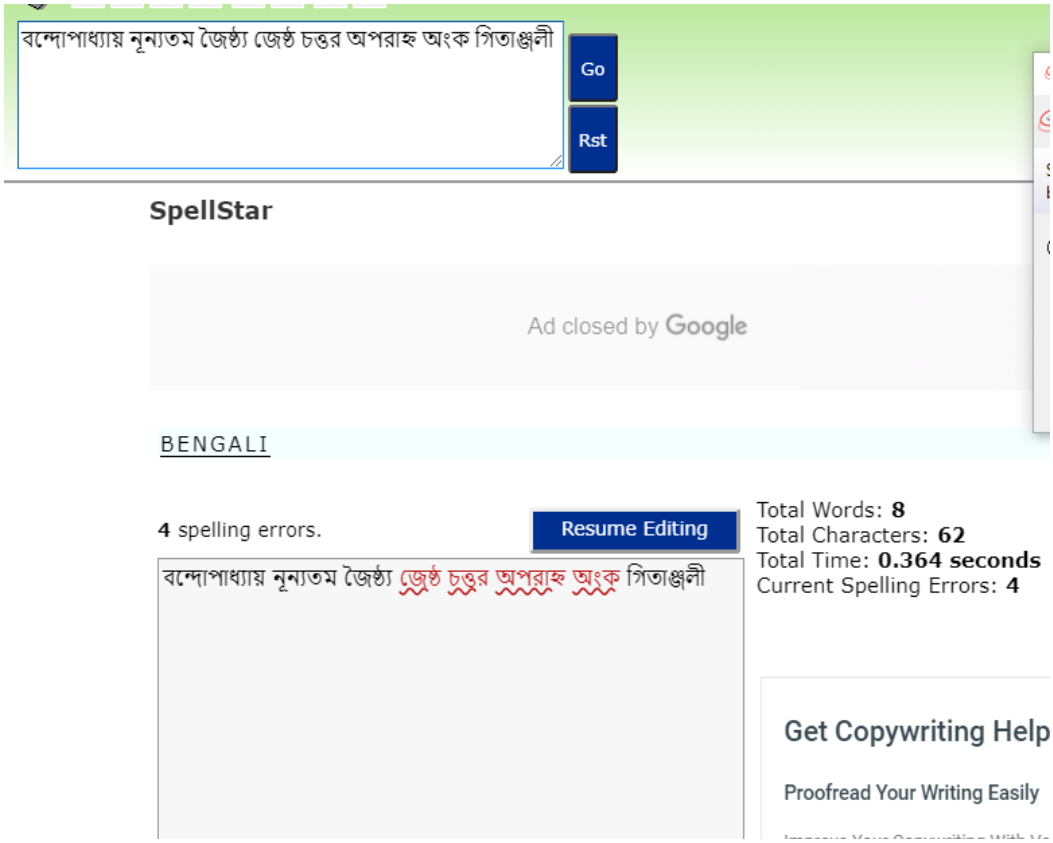

Figure 32: Some misspelled words had been inserted for spell checking

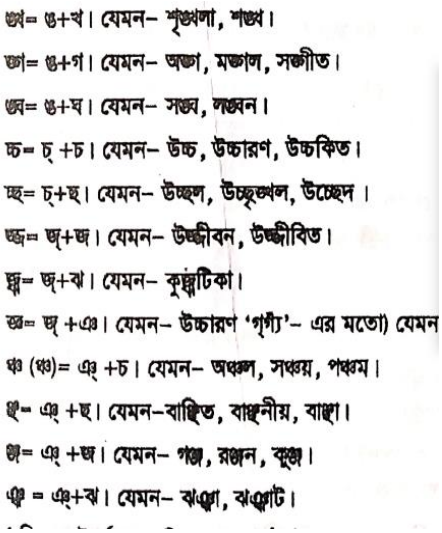
ঙ্খ= ঙ+খ। যেমন– শৃঙ্খলা, শঙ্খ।
ঙ্গ= ঙ+গ। যেমন– অঙ্গ, মঙ্গল, সঙ্গীত।
ঙ্ঘ= ঙ+ঘ। যেমন– সঙ্ঘ, লঙ্ঘন।
চ্চ= চ্ +চ। যেমন– উচ্চ, উচ্চারণ, উচ্চকিত।
চ্ছ= চ্+ছ। যেমন– উচ্ছল, উচ্ছৃঙ্খল, উচ্ছেদ।
জ্জ= জ্+জ। যেমন– উজ্জীবন, উজ্জীবিত।
জ্ঝ= জ্+ঝ। যেমন– কুজ্ঝটিকা।
ঞ্জ= জ্ +ঞ। যেমন– উচ্চারণ 'গৃহ্য'– এর মতো) যেমন
ঞ্চ (ঞ্চ)= ঞ্ +চ। যেমন– অঞ্চল, সঞ্চয়, পঞ্চম।
ঞ্ছ= ঞ্ +ছ। যেমন–বাঞ্ছিত, বাঞ্ছনীয়, বাঞ্ছা।
ঞ্জ= ঞ্ +জ। যেমন– গঞ্জ, রঞ্জন, কুঞ্জ।
ঞ্ঝ = ঞ্+ঝ। যেমন– ঝঞ্ঝা, ঝঞ্ঝাট।

Figure 33: Text for scanning by Bango OCR

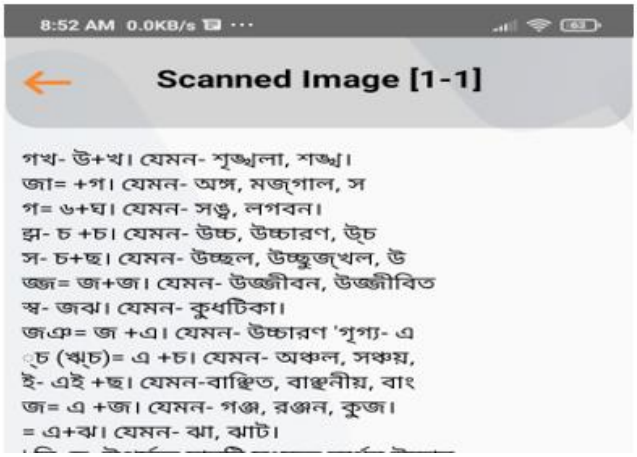

Figure 34: Editable text from jpg by Bango OCR

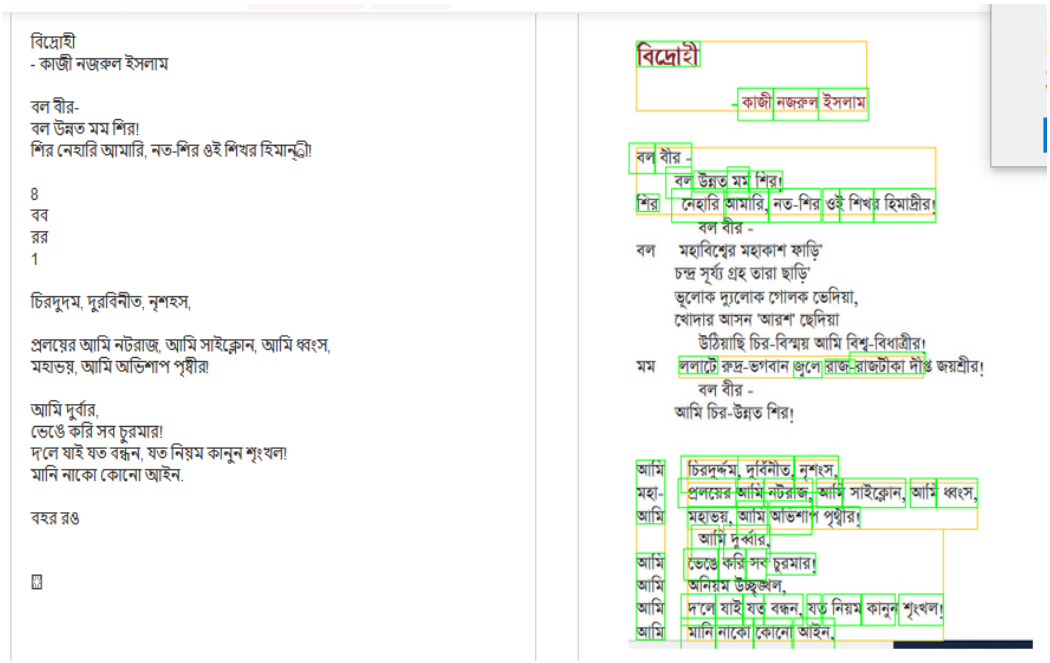

Figure 35: Text extract from Pdf (Left) and scanned by i2c online OCR

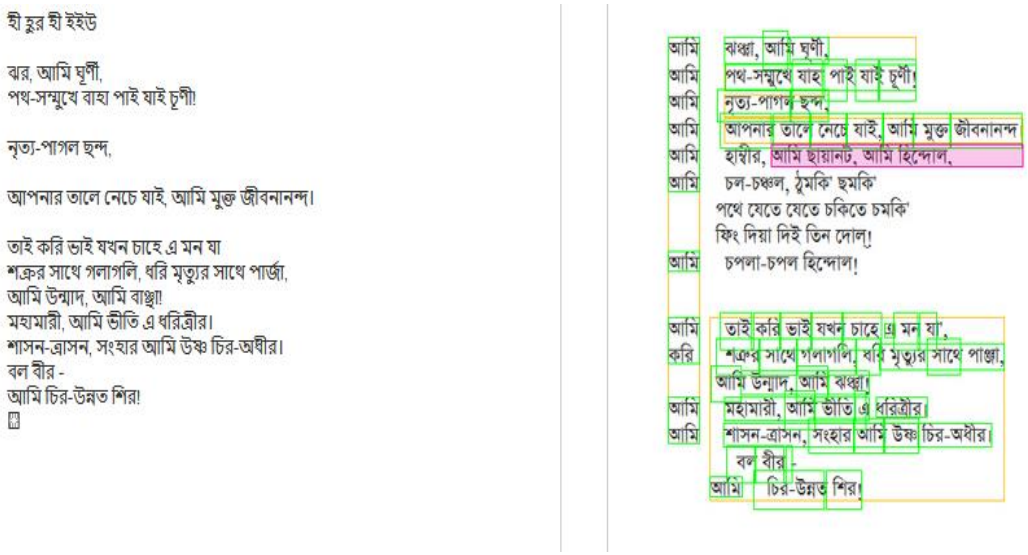

Figure 36: Text extract from pdf by i2c online OCR

CHAPTER 6
RESULT ANALYSIS

6.1. Obscurity and some Limitation of Typefaces, including ASCII and UNICODE format:

We presented fonts like SutonnyMJregular, Italic, and Bold typeface. By conversion, other fonts like Kalpurush, Solaimanilipi, AdorshoLipi, SagorLipi, NikoshLipi, SutonnyUniBanglaOMJ only in regular typeface. Presenting the three styles of SutonnyMJisonly to percept the aesthetic view or the limitations. If given all the fonts in three typefaces, then it will be lengthier. We see that in SutonnyMJtypeface, obscure typefaces like-গু ঋ গু ত্ন ত্া দ্রু ন্তু ন্মু ব্লু ভ্রু ষ্ম ষ্ট্র ড় ঢ়ু ঢু. Some points to be noted-

- '.' of ড় ঢ়ু ঢু cannot be recognized
    - 'ন', 'ম', 'ল' positioned under the 'ত্ন ত্া ন্মু ব্লু' conjuncts are not clear.
    - 'প্লু' of 'আপ্লুত' is not clear
    - The aesthetic view of the conjuncts-'ক্ত গু ঋ গু দ্রু ভ্রু ষ্ম ষ্ট্র'are problematic.
    - Bended form like-ঁ is not in proper position in- 'দ্রু ভ্রু শ্রঁ রু.'
    - Inappropriate writing form of some conjuncts like- 'হ্ন ম্ন.'
    - When we see italic of the form of '*দ্রু',* then the typeface is challenging to identify.
    - If we look at the Boldtypeface of '**ন্তু**' then, it is tough to understand the conjuncts than the regular or italic typeface.

6.1.1. Obscurity of Unicode Typefaces:

Problematic typefaces of Kalpurush are- ল্ড ন্মু দ্রু ষ্ঠ ড় ঢ়ু ঢু ম্ন. Some points to be noted-

- '.' of ড় ঢ়ু ঢু cannot be recognized
    - The aesthetic view of the conjuncts 'ল্ড দ্রু ম্ন' is problematic.
    - The typefaces of 'ন্মু ষ্ঠ' are challenging to identify.

Problematic typefaces of Solaimani Lipi are- ক্রু শ্ন ম্ন ধ্ন দ্র দ্রু ক্র ষ্ম ড়ু ঢ়ু ভ্র ভ্রু হ্রু; the typefaces of conjuncts of হ+ ণ = হ্ন, and হ্রু: হ + ন = হ্রু are similar.

- '.' of 'ড়ু ঢ়ু'cannot be recognized
- Aesthetic view of the conjuncts- 'ক্রু ম্ন ধ্ন শ্র ভ্র দ্র দ্রু ভ্রু ক্র'are problematic
- 'হ + ণ = হ্ন' which is wrong. The correct typeface is হ + ন = হ্রু

Problematic typefaces of Adorsho Lipi are- 'ক্রু ধ্ন ত্ম ন্ম স্প প্স ত্ন দ্র ভ্র ম্ন শ্র ষ্ম হ্ল'

- The typefaces of 'ত্ম ন্ম স্প প্স ত্ন' are really difficult to identify
  - Aesthetic view of the conjuncts- 'ক্রু ধ্ন দ্র ভ্র ম্ন শ্র ষ্ম' are problematic
- 14 points size of these typefaces- 'ক্রু ধ্ন ত্ম ন্ম স্প প্স ত্ন দ্র ভ্র ম্ন শ্র ষ্ম হ্ল' are a bit large

Problematic typefaces of Nikosh Lipi are-ক্ত ক্র ক্রু ধ্ন দ্ধ হ্ল হ্ন শু শূ স্তু ল্ড ত্ত খ ত্ন ত্ম ত্রু দ্রু হ্ল ড়ু ড়ূ ঢ়ু ম্ন দ্রূ ভ্রূ ব্লু স্তু

- '.' of 'ড়ু ড়ূ ঢ়ু'cannot be recognized
- The typefaces of 'শু শূ স্তু ল্ড ত্ত খ ত্ন ত্ম ত্রু দ্রু হ্ল' are really difficult to identify
- Aesthetic view of the conjuncts- 'ব্লু স্তু ম্ন দ্রূ ভ্রূ' are problematic
- 14 pointssize of these typefaces- 'ক্ত ক্র ক্রু ধ্ন দ্ধ হ্ল হ্ন শু শূ স্তু ল্ড ত্ত খ ত্ন ত্ম ত্রু দ্রু হ্ল ড়ু ড়ূ ঢ়ু ম্ন দ্রূ ভ্রূ ব্লু স্তু' are very small to read

Problematic typefaces of Sagor Lipi are- ক্র ক্ত ণ্ড রূ রু শ্রু ভ্রু শ্র দ্র স্থ ষ্ম হ্ম ষ্ম শু ড়ু ড়ূ ঢ়ু স্থ ম্ন

- '.' of 'ড়ু ড়ূ ঢ়ু cannot be recognized
- The typefaces of 'স্থ ষ্ম হ্ম ষ্ম শু ড়ু ড়ূ ঢ়ু ম্ন'are really difficult to identify
- Aesthetic view of the conjuncts- 'ক্র ক্ত ণ্ড রূ রু শ্রু ভ্রু শ্র দ্র' are problematic

The typefaces of conjuncts of হ্ণ: হ + ণ = হ্ণ, and হ্ন: হ + ন = হ্ন are similar.

The conjunct of ত্ন cannot be formed. For example- তন: ত + ন = ৎন

Problematic typefaces of Sutonny UniBanglaOMJ Lipi are-ক্ত ক্র গ্ধ ট্ট ণ্ঠ ণ্ড থ্থ ত্ন ত্ম দ্রু ভ্রু ভ্র দ্র ন্ধ ন্মু ন্ন শ্চ স্প স্প ব্লু হ্ণ হ্ন ড়ু ঢ়ু

- Aesthetic view of the conjuncts- 'ক্ত ক্র গ্ধ ট্ট ণ্ঠ ভ্র দ্র ন্ধ শ্চ' are problematic
- The typefaces of 'ণ্ড থ্থ ত্ন ত্ম দ্রু ন্মু ন্ন স্প স্প ব্লু হ্ণ হ্ন' are really difficult to identify
- 14 points size of these typefaces- 'ক্ত ক্র গ্ধ ট্ট ণ্ঠ ণ্ড থ্থ ত্ন ত্ম দ্রু ভ্রু ভ্র দ্র ন্ধ ন্মু ন্ন শ্চ স্প স্প ব্লু হ্ণ হ্ন ড়্ঢ়ু'are very small to read So, the common problem is opacity, the obscurity of typefaces, aesthetic view, and font size. We see that in table 3, the font size of SutonnyMJ and Kalpurush are acceptable, but Nikosh andSagorLipi are too small. On the other hand, AdorshoLipi is too large in 14 points size.

6.1.2. Bijoy to Unicodeand Unicode to Bijoy Converter:

Some conjuncts have been detached like অঙ্গুরি>অঙ্‌গুরি, অংগুরিয়>অঙ্‌গুরীয়, অঙ্গুলি>অঙ্‌গুলি and অজন্ত>অজনড় etc.

Some points to be noted-ল্ড

- Here 'ঙ্' & 'গ' positioned separately in অঙ্‌গুরি, অঙ্‌গুরীয়, and অঙ্‌গুলি

'ত' letter deleted from 'অজন্ত'; we see 'ড়'instead of 'ত'.

On the other h, and, from Bijoy to Unicode, some words are converted wrong like- কনিকা>কনকিা, বিন্দু>বন্‌দি, বিনিময়>বনিময়, বিরুদ্ধপক্ষ>বরিদ্ধপক্ষ, etc. Some points to be noted- 'ি' of 'কনিকা' deleted, and 'া' positionedin 'কনকিা'.

- 'ি' &'ু' graphemes deleted from 'বিন্দু', and converted into 'বন্‌দি';
- 'বিনিময়' transformed into 'বনিময়'. These are some examples from many anomalies.

- There is an option in the software like -‘ভাঙ্গা ফন্ট ঠিক করুন’ but it does not work.

So, displacements of graphemes, conjuncts in the incorrect form are the main problem of Bijoy to Unicode converter, and vice-versa.

6.1.3. Bangla Text to Speech Apps:

We pronounced three words like- গার, গাড়, গাঢ় inBangla to Text apps, but the output was only /garo/, but we know that there are three different pronunciations of those words like- /garo/ /gar^{h}o/, and /ga$\downarrow^{h}$o/

Analysis of output from Subachan Software (Bangla Text to Speech): We inserted- ‘ঙঞ ষ র ড় ঢ় ৎ ং ঁ ঃ জ্ঞ ঞ ক্ষ’- these words to see the pronunciation, but we got accent only ‘র’ of ‘র ড় ঢ়’. Phonetically the Pronunciation of ‘র, ড়, ঢ়’, is different. The standard pronunciation of the letter above is- ‘র’/r/, ড়/ɽ/, ঢ়/ɽ^{h}/’. The software was unable to pronounce the letters - ‘ঙ ঞ ৎ ং ঁ ঃ জ্ঞ ঞ. We can say that- The pronunciation of ‘ড়, ঢ়’ is not inserted. Only the pronunciation of ‘র’ is inserted in the software program

- The pronunciation of ‘ঙ ঞ ৎ ং ঁ ঃ জ্ঞ ঞ’ is not inserted in the software program

The words- ‘ছত্রিশ, একাত্তর, একক, গারো, গাড়, গাঢ়, বাহ্য, বাহ্যেন্দ্রিয়, বাহ্যদৃষ্টি, দৃষ্টি, সৃষ্টি’ were inserted to figure out the pronunciation. The following table will present the output and standard pronunciation of those words

Inserted words	output	Standard pronunciation
গারো	/garo/	/garo/
গাড়	/garo/	/ga$\downarrow$o/

গাঢ়	/garo/	/gaɽʰo/
বাহ্যেন্দ্রিয়	/ɔng/	/bajjʰondriyo/
বাহ্যদৃষ্টি	/bajjʰodriʃi/	/bajjʰodriʃti/
দৃষ্টি	/driʃi/	/driʃti/
সৃষ্টি	/sriʃi/	/sriʃti/

Table 7: software output and standard pronunciation of 'গারো, গাড়, গাঢ়, বাহ্যেন্দ্রিয়, বাহ্যদৃষ্টি, দৃষ্টি, সৃষ্টি'

We can say that-

- As the pronunciation of 'র'only inserted that is why we got the incorrect pronunciation of the following words- 'গাড়, গাঢ়'
- The pronunciation of 'বাহ্যেন্দ্রিয়, বাহ্যদৃষ্টি, দৃষ্টি, সৃষ্টি' are not processed according to the standard pronunciation.The pronunciation of the sentence like 'বঙ্গবন্ধু আমাদের জাতির পিতা' -was not clear because of echo. Hence, the problem is the recognition of speech

6.1.4. IPA Transcription:

Here some of the mentioned words have been transcribed incorrectly by the software in table 8, for example–

Word	IPA Transcription	IPA Transcription by software
অঘটনঘটনপটিয়সী	/ɔgʰɔtongʰɔtonpotiʲoʃi/ or /ɔgʰɔtongʰɔtonpotiʲoši/	/ɔgʰɔtongʰɔtnptiʲosi/
অদ্য	/oɖɖo/	/ɔɖɖo/
শ্মশান	/ʃɔʃan/	/ʃʃan/
শাশ্বত	/ʃaʃʃoṭo/	/ʃaʃʃoṭ/

হৃদয়	/hriɖoJ/	/rkɖɔJ/
ন্যূনতম	/nunɔtɔmo/	/nœ͡u̯nɔtɔm/
লক্ষ্মণ	/lɔkkʰon/	/lokkʰomon/

Table 8: IPA transcription of some words by manually and software

IPA transcription of 'অঘটনঘটনপটিয়সী' should be /ƆgʰƆtongʰƆtonpotiʲoʃi/ or /ƆgʰƆtongʰƆtonpotiʲoši/, but the software output is /ƆgʰƆtongʰƆtnptiʲosi/. If we analyze, there is no 'o' sound in-between 't' and 'n' sound for pronunciation of **'ঘটন'.**

- We know the pronunciation rule that if initial 'অ' follows the 'i', and 'u' sound, then the standard pronunciation of 'অ' will be /o/, but the software output is /Ɔ/.
- The pronunciation of the 'শ্মশান' will be /ʃƆʃan/, but the output is /ʃʃan/. Here, there is no /Ɔ/ sound afterthe/ʃ/ sound. Similarly, there is no /Ɔ/ sound after /ʃ/ sound for 'শাশ্বত'
- IPA transcription of **'**হৃদয়' should be /hriɖoJ/, but the software output is /rkɖɔJ/. Here, there is no /i/ sound precedes /k/ sound.
- IPA transcription of **'**ন্যূনতম'/nunɔtɔmo/, but the software output is /nœ͡u̯nɔtɔm/.
- As we know that there are few silent letters in Bangla Language. Here 'm' sound under the **'**ক্ষ**'** is a silent letter, but the output of **'**লক্ষ্মণ**'** in IPA transcription is /lokkʰomon/.
- An interesting matter is in figure 25. When fonts had changed from Kalpurush to Nikosh, the transcription was also converted.

In figure 26, some conjuncts are split into letters during transcription. For example, হৃinto 'hk' and letter ঋ into o͡u̯ in IPA transcription instead of /hri/, /ri/ respectively.

6.1.5. Avro Spell Checker:

If more than one word is inserted in the Avro Spell-checker, it detects one misspelled word, and then we are to move the next word manually. In figure 27, we see the spelling of 'একাডেমি' is shown wrong, but the spelling of the foreign word is validated by Bangla Academy. Similarly, in figure 29, 'দূতি' is shown correct, but the correct spelling of the word is 'দূতী'. By conversion, in figure 28, 'মহিয়সি' spelling was detected as incorrect, but there were no suggestions for the wrong word. In figure31, the software detects 'সাধিনতা' as incorrect, but showed one direction like 'স্বাধীনটা' which can confuse the user that which one is the correct one.

6.1.6. Online Bangla Spelling Checker:

Total words were eight, but the software detected only four words misspelled in figure32. We can see the results by following table 9-

Mis-spelled words input	Standard Spelling	output by software
বন্দোপাধ্যায়	বন্দ্যোপাধ্যায়	correct
নুন্যতম	ন্যূনতম	correct
জৈষ্ঠ্য	জ্যৈষ্ঠ	incorrect
জেষ্ঠ্য	জ্যেষ্ঠ	incorrect
অপরাহ্ন	অপরাহ্ণ	incorrect
অংক	অঙ্ক	incorrect
গিতাঞ্জলি	গীতাঞ্জলি	correct
চত্তর	চত্বর	correct

Table 9: Mis-spelled words' output by online Bangla Spelling software

Here-

- The spelling of 'বন্দোপাধ্যায়', 'নুন্যতম', 'গিতাঞ্জলি', 'চত্তর' is incorrect, but the software shows correct

So the correct spelling of those words should have reprogrammed.

6.1.7. Bango OCR Apps:

As we can see in table 10, maximum conjuncts and their examples in jpg format are split as well:

Original text	output by Bongo OCR apps
ঙ্গ= ঙ+গ। যেমন-অঙ্গ, মঙ্গল, সঙ্গীত	জা= +গ। অঙ্গ, মজগাল, স
ঙ্ঘ = ঙ+ঘ। যেমন- সঙ্ঘ, লঙ্ঘন	গ= ঙ+ঘ। সঙব, লগবন
চ্চ, উচ্চকিত	the first word is obscure, উ্চ
চ্ছ, উচ্ছৃঙ্খল, উচ্ছেদ	স, উচ্ছুজখল, উ
জ্ঝ=জ+ঝ। যেমন-কুজ্ঝটিকা	সব=জঝ। যেমন-কুধটিকা
জ্ঞ =জ+ঞ। (যেমন- উচ্চারণ ‘গ্‌গ্যঁ’-এর মত)। যেমন	জঞ= জ+এ। যেমন-উচ্চারণ ‘গৃগ্য’-এ
ঞ্চ(ঞ্চ)= ঞ+চ। যেমন-অঞ্চল, সঞ্চয়, পঞ্চম	্চ (ঋ্চ)= এ+চ। যেমন-অঞ্চল, সঞ্চয়
ঞ্ছ=ঞ+ছ। যেমন-বাঞ্ছিত, বাঞ্ছনীয়, বাঞ্ছা	ই-এই+ছ। যেমন-বাঞ্ছিত, বাঞ্ছনীয়, বাং
ঞ্জ= ঞ+জ। যেমন-গঞ্জ, রঞ্জন,কুঞ্জ	জ=এ+জ। যেমন-গঞ্জ, রঞ্জন,কুজ
ঞ্ঝ=ঞ+ঝ। যেমন- ঝঞ্ঝা, ঝঞ্ঝাট।	=এ+ঝ। যেমন-ঝা,ঝাট।

Table 10: Conjuncts’ output by Bango OCR apps

Some points to be noted:

- Letter and conjuncts deleted -‘ঙ’ ‘ঘ’ ‘চ’, ‘ছ’ ‘ৃ’ ‘দ’ ‘ঞ’ ‘জ্ঝ’ ‘ঞ্চ’ ‘ঞ্ছ’ ‘ঞ্জ’ ‘ঞ্ঝ’

- letter and conjuncts converted- ‘ঙ’ to ‘গ’; ‘ঙ’ to ‘ঙ’; ‘জ্ঞ’ to ‘সব’; ‘ঞ্জ’ to ‘জঞ’; ‘গ্গ্যঁ’ to ‘গ্গ্য’; ‘ঞ্চ’ to ‘ঃচ’; ‘ঞ্ছ’ to ‘ই’; ‘ঞ্জ’ to ‘জ’; ‘ঞ্ঝ’ tonothiong.

6.1. 8. i2 online OCR:

We scanned the following texts like ‘BIDDROHI’ poem by Kazi Nazrul Islam through i2c online. As an editable text, we did not get the original text in Figures 35 & 36, respectively. Some words are converted into different worlds, and some words cannot be recognized by the software. We can observe the results in the following table:

Original Word/Text	Editable word/Text
হিমাদ্রীর	হিমান্ (rest of the part is obscure as we can see in figure 35)
বল বীর, বল মহাবিশ্বের মহাকাশ ফাড়ি’ চন্দ্র সুর্য গ্রহ তাঁরা ছাড়ি ভূলোক দ্যুলোক গোলক ভেদিয়া খোদার আসন আরশ ছেদিয়া উঠিয়াছি চির-বিস্ময় আমি বিশ্ব-বিধাত্রির! মম ললাটে রুদ্র ভগবান জ্বলে রাজ-রাজটীকা দীপ্ত জয়শ্রীর! বল বীর- আমি উন্নত মম শীর!	৪ বব রর 1
দুর্ব্বার	দুর্বার

আমি অনিয়ম উচ্ছশৃঙ্খল	– (no words)
no words	বছর রঙ
আমি ঝঞ্ঝা আমি ঘূর্ণী	হী হুর হী ইইউ
আমি হাম্বীর, আমি ছায়ানট, আমি হিন্দোল আমি চল-চঞ্চল ঠুমকি-ছমকি' পথে যেতে যেতে চকিতে চমকি' ফিং দিয়া দিই তিন দোল! আমি চপলা-চপল হিন্দোল!	no words

Table 11: 'বিদ্রোহী' poem output by i2OCR

- The software cannot recognize 'দ্রীর' of 'হিমাদ্রীর'. It shows 'হিমান্' instead of 'হিমাদ্রীর'.
- The software totally failed to recognize-

'বল বীর,বল মহাবিশ্বের মহাকাশ ফাড়ি'

চন্দ্র সুর্য গ্রহ তাঁরা ছাড়ি

ভূলোক দ্যুলোক গোলক ভেদিয়া

খোদার আসন আরশ ছেদিয়া

উঠিয়াছি চির-বিস্ময় আমি বিশ্ব- বিধাত্রির!

মম ললাটে রুদ্র ভগবান জ্বলে রাজ-রাজটীকা দীপ্ত জয়শ্রীর!

বল বীর-

আমি উন্নত মম শীর!'

It shows '৪ববরর1' instead of lines mentioned above

- It can not extract ‘দুর্ব্বার’ , but ‘দুর্বার’
- No words were shown for ‘আমি অনিয়ম উচ্ছশৃঙ্খল’, but‘বছর রঙ’ was shown as output for nothing.
- ‘হী হুর হী ইইউ’ is output of ‘আমি ঝঞ্ঝা আমি ঘূর্ণী’
- The following part of the poem is not recognized by the software-

‘আমি হাম্বীর, আমি ছায়ানট, আমি হিন্দোল

আমি চল-চঞ্চল ঠুমকি-ছমকি’

পথে যেতে যেতে চকিতে চমকি’

ফিং দিয়া দিই তিন দোল!

আমি চপলা-চপল হিন্দোল!’

6.1.9. Online Speech to text converter:

We see some anomalies in an online speech to text software

- We see the output ‘বাস’of ‘বাস’, and ‘বাঁশ’. Here the sound of ‘শ’ is not processed.
- ‘নিঃ’ is deleted from ‘নিঃক্ষত্রিয়’ in output
- ‘ড়গ’ converted to ‘রক’ for ‘খড়গ.’
- ‘একাত্তর’ converted to ‘71.’
- ‘লোকটি বিষ খেল’converted to ‘লক t20 খেলা’
- ‘ড়’ and ‘ঢ়’ sounds converted to ‘র’.

6.1.10. Google Translate:

There are many anomalies of google translate. Some of them are-We see the output ‘বাস’of ‘বাঁশ’

- ‘ড়’ and ‘ঢ়’ sounds converted to ‘র.’
- No output of ‘একাত্তর’
- ‘ড়গ’ converted to ‘রক’ for ‘খড়গ.’

- ‘লোকটি বিষ খেল’ converted to ‘লোকটি ব্রিজ খেলা’

Interestingly in google translate, ‘বল বীর বল উন্নত মম শীর’ translated into ‘Ball hero ball advanced mum shir, and the noticeable thing is that 71 (একাত্তর) translated into ‘61’.

6.2. Technical Challenges of Developing Bangla Language Tools:

Several technical tools are not available for developing Bangla Languages-

- Development of Bangla font technology
- Text processing tools
- Text Annotation tools and techniques
- Data archiving of texts
- Text Normalization techniques
- Text-to Speech/Speech-To-text
- Machine Translation tools
- Speech Recognition tools
- Sentiment Analysis and Recognition
- Hate Speech analysis and Recognition
- Verbal Aggression recognition

6.2.1. Development of Bengali Font Technology:

The elaborate script like Chinese, Japanese, Arabic, Tamil, and many more developed user-friendly and compatible fonts in their languages because of the availability of font technology. On the contrary, there is no available font technology, for example, Though some experts figure out that Bangla is not linear like English, so the problem arises when conjuncts come like ‘স্ক্রু’ then it is challenging to develop typeface. The mapping of Bangla, and English Alphabets have been shown below-

There are several mapping lines for designing English graphemes like top line, headline, midline, baseline, and drop line. The following figure shows the lines-

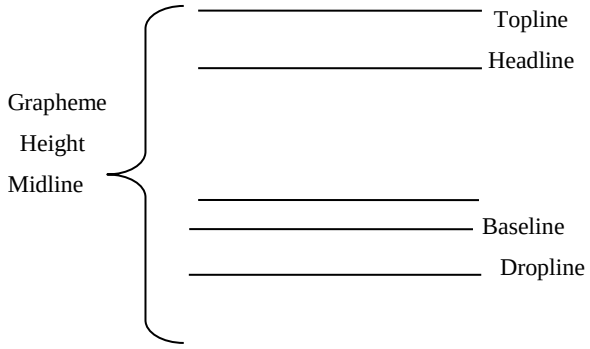

Figure 37: Basic Grapheme Height of letters

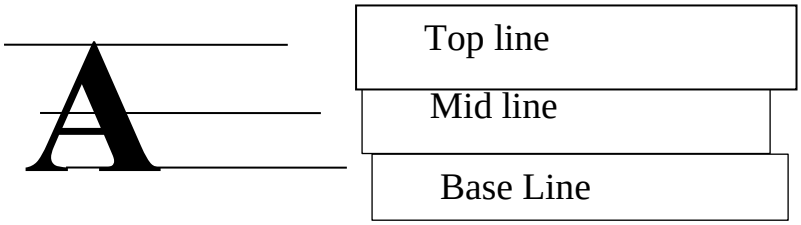

Figure 38: Mapping line for designing English grapheme

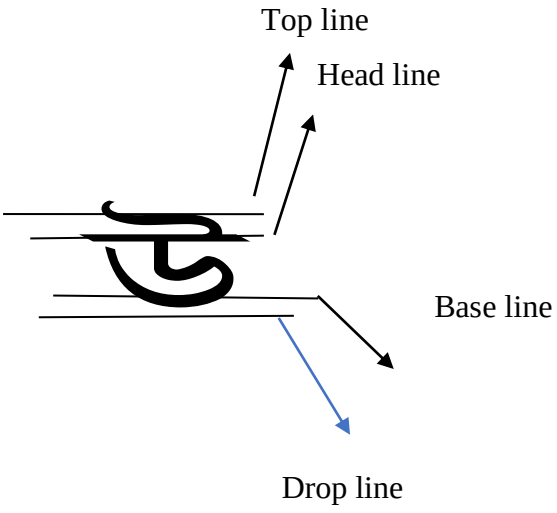

Figure 39: The picture of Mapping for 'উ'

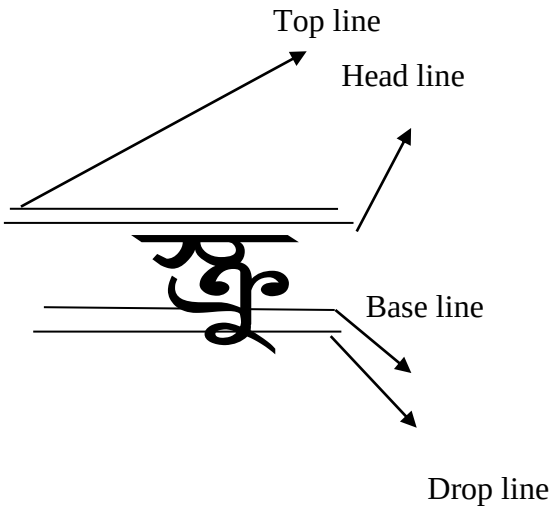

Figure 40: The picture of mapping for Bangla conjunct 'স্ক্র'.

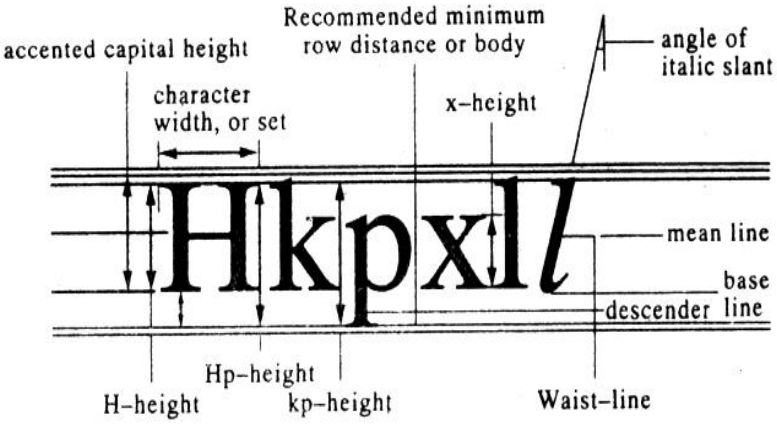

Figure 1(a). Parts and dimensions of typographic letter-forms generally.
Source: Hugh Williamson, *Methods of Book Design*, Yale U.P. 1983.

Figure41:Parts , and dimensions of English letters

In Bangla Language many graphemes go to the bottom position even after the for example,‘ক্ক, ন্মু ক্ষু, স্তetc as we see in figure 40. Again, we see that the bend form goes over the upper line, for example, ‘ি, ী, ৈ, ৌ, ঐ, ঔ, ই, ঈ, উ, ঊ’ etc. as we see in figure 39. There are ten modified vowel forms, and six modified consonant forms, with 300 possible conjuncts with a different form. Compared to Bangla, English is a linear script as we see in figure 38, and41. Another problem is that when we write in Bangla, the font size should be 14 points, in English, 12 points if we put together the languages. Moreover, 16 points are standard for mentioning vowel and consonant modifiers.

6.2.2. Lack of Standard Recorded Speech:

We could not develop well-representative BLP tools like Speech to Text Converter, Speech Recognition, Sentiment Analysis, Text to Speech Converter, and machine translation because of standard recorded speech. There are many reasons behind this-lack of high-quality recording tools, noise-free or soundproof room, high-quality voice or loud, and clear voice, using high-quality software for noise cancellation, extract exact tone or sentiment of voices. Some researcher-developed their imbalanced dataset.

6.2.3. Lack of Correct Speech Annotation:

As there isan Individual dataset of speech available, so the high-quality voice is rare. Not only voice but also experts’ bodies in the annotation process is also rare. Sometimes, memory capacity does not support the dataset. There will be no search option randomly within the dataset because of the lack of a large dataset. For h, handling real-time processing, we need accurate digital signal processing (DSP). A digital grammar with acoustic analysis and morphological analysis of speech is very significant for supporting application level. The technique or model like rule-based or selection based should be applied for extracting naturalness. The trained model even can represent the naturalness of a large volume of speech data.

6.3. Linguistic Challenges of Developing Bangla Language Processing Tools:

There are several linguistic issues that are not available for developing Bangla Languages-

- Standards, and well-representative speech, and text corpus
- Digital Lexical databases
- Frequency-based Lexical Databases
- Concordance List of most frequently used words
- Parsed Syntactic Databases
- Digital dictionaries and thesauri
- Digital Bengali grammar

6.3.1. Standard and well-representative Speech and Text corpus:

There is no standard corpus in Bangla with a superior quality of the speech recordings. This will need to be developed to aid in spelling, grammar checking, and speech reconstruction. Enlarging the corpus increases the probability that acceptable units will be found in the database (e.g., the 16-h corpus ATRECCS provided for the Blizzard Challenge 2006 (Bennett, 2006). There is some example, of corpus development which is collected personally. In 2017, the government took the initiative to develop andstandard written and speech corpus. A scholar, Professor NiladriShekhor Das at ISIT, designed a plan for developing the Bangla Corpus development of Bangladesh from India. However, it is regrettable that the proposed plan still needs clearance. Not only speech, and written corpus there are no available corpora like Parallel text corpus, Multimodal Text, and Speech Corpus, Annotated Text, and Speech Corpus, Processed Text, and Speech Corpus.“Till now, there exists no standard synthetic speech evaluation technique. This might be achieved by moving away from forced-choice tests using abstract emotion words towards tests measuring the perceived naturalness of an utterance given an emotion defining context (Haque, 2013).”

6.3.1.1. Nasalization and Nasal Sounds:

There are many features while we speak. Linguistically, we call them suprasegmental features, such as stress, intonation, pitch, tone, amplitude, length, etc. These features should be included in the speech synthesis model or technique. Besides this, nasalization is a significant feature of Bangla vowel sounds- 'অঁ, আঁ, ইঁ, উঁ এঁ অ্যাঁ, ওঁ'. By conversion, nasal sounds are consonants- 'ঙ, ম, ন'. There are no well-developed features for annotating the nasalized vowels of the Bangla language.

6.3.1.2. POS should be defined according to Context:

Speech is a continuous process. It changes every second, so the role of POS may vary according to users. It should be defined accordingly. For example, we have a different meaning and POS role of 'মাথা' in Bangla-

'আমার মাথা ব্যথা করছে' (Means- 'an organ of upper-part of body, POS-Noun)

'তিনি আমদের বিভাগের মাথা' (Means- 'Head of the department', POS-Noun)

'তার প্রোগ্রামিং এ মাথা ভালনা' (Means- 'Intelligence' POS-Noun)

'এ কাজে মাথা খাটাও' (Means- 'Intelligence' POS-Verb)

So, a large number of defined lexicons according to context is a must for correct speech recognition

6.3.1.3. Flexibility of Sentence Structure and Dependency Parsing:

Another Linguistic challenge is the flexibility of Bangla Sentence structure. For example, the sentence 'আমি তোমাকে ভালোবাসি' can be written in the following ways-

- 'তোমাকে আমি ভালবাসি'
- 'ভালোবাসি তোমাকে আমি'

- ‘ভালোবাসি আমি তোমাকে’
- ‘তোমাকে ভালোবাসি আমি’
- ‘আমি ভালোবাসি তোমাকে’

but in English language we cannot say like following:

*I you love; *You I love; *Love I you; * Love you I etc. So, we need to analyze sentences and then coding. One of the big challenges is dependency parsing in Bangla where all component’s relation with verb of a sentence are to analyze.

6.4. Spell Checker:

Researchers are struggling to develop a Bangla Spell-checker in terms of homonyms and Conjuncts of Bangla Words. No notable works are found in
the researchfields.

According to Naima Islam Nodi et al., 2020-

The problem that is faced related to the spell checker work are:

- Phonetic similarity of Bangla Characters
- The difference between grapheme representation and phonetic utterances
- Bangla is a language of complex orthographic rules; as results, there exists a large gap between the spelling and pronunciation of a word
- A large number of words in Bangla is originated from Sanskrit, an ancestral predecessor of Bangla. However, these words are either modified in terms of pronunciation or spelling and pronunciation. Thus, there exists a gap between spelling and pronunciation requiring complex orthographic rules.

After Naima Islam Nodi et al., 2020, Some work limitations of Bangla Spellchecker are shown in the following table-

Title	Authors	Work Done	Limitations
A Comprehensive Bangla Spelling Checker	(Naushad UzZaman & Mumit Khan, 2006)	Solves the misspelled word with a maximum edit distance of two, using phonetic	Unable to solve Nth character error, space-related errors, homonym errors, conjuncts with

		encoding.	unusual punctuation, and different pronunciation in a different context
Reversed word dictionary and phonetically similar word grouping based spellchecker to Bangla text	(Bidyut Baran Chaudhuri, 2001)	A modified dictionary has been used for phonetically similar words. A reversed dictionary is used for unmatched words, and by comparing these two dictionaries, the odd area and error positionis identified, and thus the correct result is being provided.	There is no discussion of conjugate words known as“juktakkors” in Bangla.
Coding System for Bangla Spell Checker	(Md Tamjidul Hoque & Md Kaykobad, 2002)	SQL queries to detect frequent errors, devise coding, and prioritize minor mistakes, improve the suggestion list and its presentation, code for detecting minor or delta errors, and break-up words.	The word having similar sounding-based suggestions will come later.
A Generic Spell	(ABA Abdullah &	Add-in approached	The used recursive

Checker Engine for South Asian Languages	Ashfaq Rahman, 2003)	has been used to detect highly misspelled words and provides a suggestion for them.	simulation algorithm has lots of bugs and fails to offer correct recommendations.
A Bangla Phonetic Encoding for Better Spelling Suggestions	(Naushad Uz Zaman & Mumit Khan, 2004)	Phonetically error correction and provided Unicode for Bangla letters are phonetically the same as spell.	Does not offer Unicode for all the vowels. Some of the English letters more different letters in Bangla that may cause an ambiguous situation.
A double Metaphone encoding for Bangla and its application in spelling checker	(Naushad UzZaman & Mumit Khan, 2005)	Solves error with edit distance one using double Metaphone encoding.	Remains some ambiguity in the case of multiple pronunciations of the same letter.
Clustering-based Bangla Spell Checker	(Prianka M, andal& BM Mainul Hossain, 2017)	Reduces time and space complexity by Clustering the dictionary. The clustered dictionary is constructed on Structural Similarity and phonetic similarity of words.	Makes no solve space-related errors and homonym errors.
A Spell Checker and Corrector for	(LA Grobbelaar & JDM Kinyua,	A step has been taken to make a	Wordwise correction is being provided by

the Native South African Language South Sotho	2009)	multithreaded spell checker searched in the dictionary to give the correct result.	separating the prefect Translation of more than one phrase simultaneously is difficult in this approach. Excluding "Full stop," other punctuations have not been considered.
A Non-Learning Approach to Spelling Correction in Web Queries	(Jason Soo, 2013)	Adverse environment spelling correction algorithm, Segments has been used, a hybrid approach that uses N-gram, and a series of substring generation rules, does not require training data. Also, language and domain-independent.	Algorithm complexity is higher. Web crawling can provide misspelled words sometimes. In this case,the result in accuracy decreases.
Spell Checker	(V.V Bhaire, A. A. Jadhav, & P. A. Pashte, 2015)	A spell has been checked using edit distance and comparing it with the corrected dictionary. Digital tree data structured has been used to store a dynamic set	It does not solve homonym errors, name (noun) errors, unable to solve multiple errors in a word. It's not a language-independent solution.

		or associative array of English words. An effort has been made to omit space related errors.	
Automated word prediction in Bangla language using Stochastic language models	(Md. Masudul Haque, Md. Tarek Habib, & Md. Mokhles, 2015)	N-gram language model, back off, and deleted interpolation techniques have been used to predict Bangla words in a sentence.	TheSolution provides less accuracy for a large data set.

Table 12: Limitations of some Bangla Spell Checker

6.5. Anomalies in popular keyboard including Bijoy Bayanno 2020 and Avro Keyboard 5.5.0:

Some notable problems occur in Bijoy 2020 software like-

1) If we type unconsciously 'ছ্,' and then delete 'ু'. Afterward add 'ু' then the modified vowel goes little far from the letter. To avoid this problem, we had to delete whole 'ছু'' and retype the 'ছু'
2) There are problems in ligature forms
3) Another problem is in bold typeface. Sometimes the conjunct looks obscure.
4) If the windows version changes, then the fonts break down
5) Bijoy Bayanno does not support google meet
6) There is no inbuilt spell or grammar checker

Some significant problems occur in Avro 5.5.0 like:

1) If two-letter inserted side-by-side, then it automatically makes conjuncts, which is a big problem
2) It is not fully phonetically based but seems like phonetic
3) Sometimes more stroke need to type a word, which is time-consuming

4) The google meet platform ‘Hasanta’ cannot be written under a single word; it automatically constructs conjuncts with the following word.
5) It does not support Abode photoshop, Elastrator, Premium, etc.

6.6. Research Observations:

While conducting this study, we observed the following things-

6.6.1. Lack of Corpus:

First of all, we have no benchmark dataset or a representative corpus of the Bangla language. Most of the available canon are imbalanced.

6.6.2. Lack of Collaboration:

Collaboration among linguists and technologists is very rare. Most of the tools are developed by a technologist who has little knowledge of the language. So, we should focus on collaboration.

6.6.3. Lack of Information:

There is no comprehensive information regarding Bangla language processing. While searching for information, get some scattered data. We have no dataset on researchers, technical experts, linguists, developers, and software companies in this field. That is why our students are not much interested in Bangla Language Processing

6.6.4. Lack of users of BLP tools:

Most of the students, teachers, and technologists do not use Bangla Language processing tools. They even do not know much about Bangla language processing. We will not get actual feedback unless the devices reach the users.

6.6.5. Lack of Experts:

The scarcity of experts in technology and linguistics is another major problem in Bangla Language Processing. Any linguist does not validate some students both in public and private institution conduct project on BLP. But the data they used or simply is imbalanced. We do not have prominent linguists after Dr. Muhammad Shahidullah, Dr. Mofazzal Haider, Professor Kazi Din-Muhammad, Professor Rafiqul

Islam, Professor Monsur Musa, Professor Abul kalam Monjur Morshed, Professor Daniul Huq.

6.6.6. Lack of Researcher:

Researchers are not available in this field up to dem, and of this field. In 1992, the Department of linguistics started, but the research on Bangla language Processing seems rare. A few of the students conducted M. A. or M.Phil. Research on BLP. As far we know, not a single Ph.D. research is completed on BLP in this department. , but six students are conducting Ph.D. in computational linguistics in SUST.

6.6.7. Lack of Interest:

Lack of interest is another barrier to step forward in the BLP sector. Some researchers think that our paper will not publish in renowned journals because of this type of field. So, they changed their lot later. The researcher findsit difficult to conduct the study as lack of dataset.

6.6.8. Lack of Availability of BLP tools:

We see many papers available in ResearchGate or academia.edu, but the tools are not reachable. When we searched, a few software can be download like ‘OCR,’ ‘Keyboard layout,’ ‘Fonts.’ Only a few tools like an online spellchecker, OCR, Font converter are available. Sometimes, the software does not work, hang the device like a mobile phone or desktop/laptop. All are not open-source software.

6.6.9. Lack of Analyzation of Anomalies:

Most of the BLP tools are not up-to-date as there is no analysis of bugs. As we earlier mentioned that the number of BLP tool Users is deficient, including linguists. That is why the proper limitation of the specific tool is hidden.

6.6.10. Lack of Compatibility Issue:

Many tools have a problem of incompatibility. For example, document typed in Bijoy 2007, 10 0r 16 does not fully support the next version. Fonts are broken by converting

Bijoy to Unicode and vice-versa. If one has no Bijoy interface, s/he cannot read the file. ASCII does not support IPA transcription software and google meet. By conversion, Avro does not support Adobe photoshop, Elastrator, Premium, etc.

6.6.11. Lack of Cooperation:

While conducting this research, we knocked several language experts and technologists who have not responded. In ResearchGate, we request almost 150 papers for study, but around seven requests are accepted. Here, we see that both the technologist and linguist treat themselves like I know better.

6.6.12. Lack of Funding:

We need funds to work better in this sector. For the first time, we see an immense amount of around 160 Crore Tk to complete a project on BLP in Bangladesh Computer Council by the government. University and private sector can invest in this sector to move our country digitally advanced.

6.6.13. Lack of Digital Grammar:

So far, we have no complete Bangla Grammar. We need a Bangla grammar where all the functions of the words are required to be discussed. For example- there are many words same spelling but have a different role in sentences. Lemmatization is very important to tokenize the root word so, we need a digital lexicon dictionary, including functions of lemma.

CHAPTER 7
PROPOSITIONS: MEASURES TO BE TAKEN TO OVERCOME THE CHALLENGES OF BANGLA LANGUAGE PROCESSING

7.1. Language Planning and Policy:

We have already celebrated 49 years of Independence Day. However, till now, language planning and policy are not settled. It is a matter of sorrow that our freedom fighter sacrificed their life in 1971, motivated by language martyrs. Still, the policymakers are only discussing this issue, but the action is absent. We will not be able to overcome the challenges in Bangla Language Processing until we form the highest-body monitoring of the Bangla language and giving us direction.

7.2. National Committee:

A national committee should be present for 2 or 3 years and play a vital role regarding the Bangla Language Processing. The members of the committee that can be conformed are as following:

1) Policy Makers
2) Technology Experts & Programmers
3) Linguists (Bangla)& Language Experts (Both from English & Sanskrit)

7.3. Bangla Corpus:

In the 21st century, there is no Bangla Corpus. ManyBangla language Processing field researchers claimed that the software tools are not up to users' demand because of a well-representative corpus. Linguists should focus on this area right away.

7.4. Research Collaboration between Experts:

So far, full-fledged compatible Bangla Language Processing tools are not available. One of the main reasons is the lack of collaboration between experts, including linguists, technology experts.

7.5. Collaboration among Government, Universities, and Software companies:

There is very little collaboration in our country. Recently, APURBA Technologies signed an MoU with two universities, including Daffodil International University (DIU), North South University (NSU). In DIU, they fund three teachers and some students. In the future, they will include more students for funding, according to the CEO and founder of the companies. Collaboration among government, linguistic experts, and IT experts are seen in Bangladesh Computer Council from 2016. No other partnership is mentionable. Maximum Bangla Language Processing tools have been developed personally.

7.6. Funding Opportunities:

The funding opportunity is a significant problem or barrier for developing user-friendly Bangla Language Processing software. Recently, the Bangladesh computer council is developing 16 components of Bangla Language Processing software funded (Approximately 260 crore Tk) by the government.Ministry of Information and Technology give yearly fund based in CSE department of some public universities. Private and public universities offer internal funds to research, but the amount is not always sufficient to produce a better tool. So, funds should be increased in this sector.

7.7. Courses should be an offer like Bangla Language Processing in Universities:

Recently Shahjalal University of Science and Technology (SUST) started computational linguistic courses. By conversion, the linguistics department at the University of Dhaka taught some studies related to this field. They offered courses like Phonetics & Phonology, Morphology, Syntax, Semantics, Pragmatics, Sociolinguistics, Cognitive linguistics, Corpus linguistics, etc. Without these two institutions, similar courses are not available anywhere in Bangladesh. Some introductory courses should be offered at the undergrad level in all universities, including public and private. By providing these types of studies, we can motivate students to work dedicatedly in this field. In Bangladesh, students choose subjects

according to the demand of the job sector. That is why it is high time we made some job sectors in BLP.

7.8. Open Source platform like Linux should be used to Support BLP tools:

Maximum BLP tools support only the Windows platform. We should develop tools that can helpwith Linux, Android,and IOS. We should keep in mind that open-source platforms are becoming popular, so we need to build more BLP tools for this platform.

7.9. Models need to be Justified and Well-trained:

There are many models for computing Bangla Language. Every model has its benefits and limitations. A perfect model needs to figure out for natural processing language succesfully and get the output according to the expectation. For example, a deep neural network is more popular than a stattistic model because of its benefits. But also we should develop more features to train well the model for Bangla Language.

7.10. Recommendations for ICT Ministry:

1. ICT ministry should form a department like Bangla Language Processing.
2. They should offer to fund research on BLP.
3. Form a monitoring cell for BLP tools. This committee makes sure of collaboration between linguists and technologists while developing BLP tools.
4. They can offer Research Awards to encourage the researchers.
5. They can form a central website or achieve where all information regarding Bangla language processing will be available.

7.11. Recommendations for Linguist:

1. The linguist should make a representative Bangla grammar
2. They should know what programmer dem, and for Bangla language processing
3. POS tagging correctly

4. Find out the lemma correctly
5. There is no digital Bangla grammar in this digital era so that they can focus on this area.
6. They should make a representative corpus, including speech and written corpus of Bangla Language.

7.12. Recommendations for Technologist:

1) The technologist should have good knowledge of the Bangla Language.
2) They should collaborate with linguists while developing BLP tools
3) They should know the updated methodology, model, or technology for Bangla Language Processing
4) They should ensure high-quality devices while working in this sector

7.13. Recommendations for Researcher:

1) The researcher should find out the research gap in the field of Bangla language processing
2) They should publish quality research more
3) Collaboration is a must for researching this sector
4) Encourage some prominent researcher in this sector
5) They should dedicate themselves to this field
6) Share their investigations, among others

7.14. Recommendation for the Developer:

1) Developer should develop Bangla language processing tools according to the demand.
2) They should apply the most updated technology
3) They can offer training or intern on BLP under their software companies
4) Recruit some linguist as a consultant
5) To conduct survey

7.15. Recommendation for Users:

1) User should use Bangla language Processing Software frequently
2) They can use mobile apps and try to find out bugs (a problem that needs to be fixed)
3) They can express their opinions via social media or direct to the developer or researcher as their feedback is essential.

7.16. Recommendation for Bangla Conjuncts and Ligature redesigning:

1) Conjuncts and Ligature are the select properties of the Bangla language, but sometimes it is very tough to set a model for those unique form like- ‘ঞ্চ, ঞ্ছ, ক্ষ্ণ শ্ন ষ্ণ হ্ণ দ্ভ দ্ভ ক্ত ষ্ম ড়্ ঢ়্ ঢ্ভ ঙ্ক্ষ হ্ন’ etc., and ligatures like- ড়্ ঢ়্ ঢ্

We need to redesign the shape of conjuncts and ligatures

2) There should be an option in every Bangla Keyboard Interface to write conjuncts in simple form as well. For example-‘শ্ণ’ instead of ‘শ্ন.’

We should develop font technology to overcome the opacity of conjuncts and ligatures.

CHAPTER 8

CONCLUSION

This paper presented Bangla Language Processing (BLP) in Bangladesh regarding trends challenges and recommendations for overcoming the existing obstacles. Bangla Language Processing in Bangladesh still is a prominent place to explore. We discussed available BLP tools and tools under development as well. We tested some software from online and Android platforms to find out anomalies or limitations. Our selected software is Optical Character Recognition (OCR), Bangla Spell Checker, Bangla Keyboard Interface, Fonts, Machine Translation (MT), Text to Speech Converter (T2S), Speech to Text Converter (S2T), Bangla Search Engine, Bangla Corpus Development, IPA Transcription, Sentiment Analysis, and Bangla POS tagging. We collected data from four interviewee groups, including developers, researchers, technical experts, language experts, and users. Despite our best efforts, we could not reach the policymaker.

If we look at Bangla Language processing tools like Bangla font, we see many anomalies. Especially in many conjuncts mentioned in the data presentation chapter, we cannot display them electronically transparently. Likewise, SutonnyMJ the conjuncts- ণ্ড, ঙ্ক, ণ্ড, ত্ন, ত্ম, দ্রু, ন্ত, ন্মু, ব্ল, ভ্রু, ষ্ম, স্থ, ড়ু, ঢ়ু are not clear in all presented fonts, and their sizes are different in 14 points size considered standard for Bangla Type. Command to type Bold or regular does not work correctly for some Bangla fonts. Another essential thing is the aesthetic view of fonts. Some fonts are too small in standard 14 points size, and some are comparatively larger. There is debate on Bangla Keyboard Layout like which is convenient- Bijoy Layout or Phonetic Layout. We cannot write Bangla fully phonetically with the help of the present English Keyboard. Ligature and conjuncts are broken while convert from ASCII to Unicode and vice-versa. Speech to Text, along with google translate software, does not recognize the pronunciation of ড়, ঢ়, ঞ্জ, ঞ্চ, etc. and interestingly, the pronunciation of ‘একাত্তর’ annotates as ‘61’. Bangla OCR does not recognize many words, conjuncts, and different formats like pdf as well. Spell checker like Avro suggests wrong spelling and also more than one or two false suggestions. IPA transcription software does not support ASCII or android along with silent letters in some cases. The

challenges are limitation of technology, lack of experts, lack of datasets, a lack of linguistic knowledge, lack of collaboration, lack of funding, and many more. We see that we have no benchmark or standard corpus, including speech and written form. We need permission to use datasets from private sources.

However, we have not applied font technology for the Bangla language well; by conversion, there are so many complex scripts like Chinese, Japanese, Tamil, Arabic languages that developed their font technology and can display their written form. Even the writing system is different from some languages like Arabic (Right to Left). Nowadays, the Avro keyboard gets more popular than Bijoy because the new generation compares English letters to Bangla letters and finds phonetic layout like Provat, Ridmik, Google-Board, Ekushey, etc., convenient. Unpublished research showed that 90% of young (approximately 5,000 students were participants) are comfortable with the phonetic layout. As ASCII format does not support maximum field, most Unicode fonts are now available, but the major problem is the compatibility issue. But another thing is that Unicode does not support photoshop, illustrator, premium, etc. Typing is always with typo error, so in English and other languages spell checker is available; conversely, we did not get any Bangla spell checker and grammar checker. If we want to develop a user-friendly speech-to-text and vice-versa software, we need to correct annotated addresses from the audio file of standard pronunciation. Separately or personally, some works have been done in this field. Still, those words and the articulation of Bangla words in google translate are not validated or authenticated by any linguist.

Moreover, there is a debate, whether the book represents the standard pronunciation of Bangla words of Dhaka City or Bangladesh, on the book named 'উচ্চারণের অভিধান' published by Bangla Academy. The application of such rules in terms of pronunciation should be marked by digitalization and preservation of Standard Pronunciation of Bangla words frequently used in all sectors. Some accents data have not been inserted into the software. There is no digital audio record of Bangla words. Bangla OCR, Speech to Text, and vice-versa software are not user friendly. Sentiment analysis software is at armature level. POS tagging software, stemmer, lemmatization, word categorization are not developed up to the user dem, and only because of not publishing a digital dictionary. We got IPA transcription software for the first time in Bangla, but it does not support ASCII format and the android version. Some words

cannot be transcribed correctly. Regarding the corpus, there is no complete representative Bangla Language Corpus. Some developers made some canons only based on the specific field; for example, a developer developed a tool like speech to text. He annotated datasets from different users of different contexts. But the major problem is whether the corpus is validated or reviewed by a linguist or language experts. Without the current project running in Bangladesh Computer Council. There is no complete collaboration in this field. Many organisations and institutions developed their own datasets, but those are not open for publicly accessible for example, google developed a datasets, but it is not open or even validated by linguists or experts. Prof. Muhammad Zafar Ikbal took an excellent initiative personally to apply Computational Linguistics in the academic area in SUST. CRBLP of Brac University started to develop some BLP tools funded by a Canadian organization, but unfortunately, that stopped. So, funding and collaboration seem rare. Students of the computer science or linguistic department are not interested in this field as there is no complete Bangla Language corpus. Especially a public institution might play an important role in this sector. We requested almost 150 academic papers for getting full text on BLP through Researchgate, but around 5-6 researchers send the articles. So, the lack of sharing research is another big problem in our country. We could not reach many researchers in this field after knocking via mail or phone-contact, even after confirmation of the appointment. A project offered by the government (2016-2021) continuing in Bangladesh Computer council makes us think positively regarding Bangla Language processing. For the first time, a big budget is being spent, including collaboration. Tender is also another problem because of the big-budget. It took a long time to start the project. We can fix most issues like bugs and anomalies by making a well representative standard Bangla corpus and more funding and collaboration. According to the standard Bangla Corpus, we should reprogram our BLP tools, and tools should be open source, including active online and offline. The researcher should work actively in this field. We did not mention exactly how many software companies are working in this sector; their purposes are not understanding. Paid and open-source software should be tested by linguists. The researcher should analyze the anomalies of BLP tools up to the demand of the developer. A complete Bangla Corpus should be developed, and find out the more appropriate model for Bangla Language processing. Digital Bangla Grammar is a must for understanding

Bangla Language quickly. Language should have got priority as our martyr sacrificed their lives to protect the right of mother tongue. Still, our government and experts in this field are silent for many years after independence. That is why we cannot move forward. Almost all the works are done based on written corpus; there are a few correct annotated datasets on speech. We only studied the open-source tools but paid-version Bijoy Bayanno 2020. Paid tools should be analyzed. The dataset for the devices has not been discussed fully as those are not available. We did not make a comparison for all BLP tools without a few like fonts, spellchecker. We need some detailed linguistic challenges and solutions to overcome limitations. Researcher in this field will get an overall idea regarding Bangla language Processing in Bangladesh and motivate to explore many areas of this subject. It is high time we worked together to update or fix bugs up to the users' expectations.

References

ইসলাম, রফিকুল (১৯৯২)। *ভাষাতত্ত্ব।* ঢাকা: বুক ভিউ

চ্যার্টাজী, এস. কে. (১৯১২)। *বাংলা ভাষার উদ্ভব ও বিকাশ,* নিউ দিল্লী: রুপা এন্ড কো.

পবিত্র সরকার (১৯৮৬), *'বাংলা বানান সংস্কার ও সম্ভাবনা ও সীমাবদ্ধতা',* (পশ্চিম বাংলা আকাদেমি (সম্পাদিত), প্রসঙ্গ বাংলা ভাষা), কলিকাতা: তথ্য ও সংস্কৃতি মন্ত্রনালয় পশ্চিম বঙ্গ, পৃষ্ঠা-৮৯

জব্বার, মোস্তফা, (২০০৫), ডিজিটাল বাংলা। ঢাকা: আগামী প্রকাশনী, পৃষ্ঠা-১-১৭

মোরশেদ, আবুল কালাম মনজুর (২০০২)। *আধুনিক ভাষাতত্ত্ব*। ঢাকা: মাওলা ব্রাদার্স

সরকার, পবিত্র এবং হক, মোহাম্মদ দানীউল, *'বাংলা ভাষা'* সিরাজুল ইসলাম (প্রধান সম্পাদক), বাংলাপিডিয়া। খণ্ড ৬ (ঢাকা: বাংলাদেশ এশিয়াটিক সোসাইটি, ২০০৩), পৃ-৩৪১

সাখাওয়াত আনসারী (২০০১), *বাংলা লেখ্য রুপের সংস্কার: প্রাসঙ্গিক বিবেচনা,* প্রীতি কুমার মিত্র (সম্পাদিত), আইবিএস জার্নাল, ৮ম সংখ্যা, রাজশাহী বিশ্ববিদ্যালয়, পৃষ্ঠা-৬৮-৬৯

সিকদার, সৌরভ (২০১৪)। *বাংলা ভাষা ও বাংলাদেশের ভাষা*। ঢাকা: প্রতীক প্রকাশনা সংস্থা, পৃষ্ঠা-৬৪

হক, মোহাম্মদ দানীউল (২০০৩)। ভাষাবিজ্ঞানের কথা। ঢাকা: মাওলা ব্রাদার্স

রামেশ্বর শ' (২০০৪), সাধারণ ভাষাবিজ্ঞান ও বাংলা ভাষা, কলকাতা: পুস্তক বিপণি, পৃষ্ঠা-৪৭৯-৪৮৩

Abdullah, M., & Mumin, A. (2019). *u sin g F a c to re d T ra n sla tio n M o d e l. April*.

Abujar, S., S., M., Rahman, A., & Sattar, A. (2017). State of the Art Research for Bangla Text to Speech on Android Platform. *International Journal of Computer Applications*, *170*(1), 19–23. https://doi.org/10.5120/ijca2017914650

Ahmed, B. S., Sakib, A. K. M. N., Mahmud, I., & Rahman, S. (2012). *The Anatomy of Bangla OCR System for Printed Texts using Back Propagation Neural Network. 12*(6).

Ahmed, S., Sadeq, N., Shubha, S. S., Nahidul Islam, M., Adnan, M. A., & Islam, M.

Z. (2020). Preparation of bangla speech corpus from publicly available audio & text. *LREC 2020 - 12th International Conference on Language Resources and Evaluation, Conference Proceedings*, *May*, 6586–6592.

Aktar, Rumana (2013)**.** *Bangla Language Processing: A Linguistic Analysis, Overview and use of Bangla glyphs and Typology in Computer Programmes* [Unpublished master's thesis]. University of Dhaka.

Al Mahmud, N., & Amjad Munni, S. (2020). Qualitative Analysis of PLP in LSTM for Bangla Speech Recognition. *The International Journal of Multimedia & Its Applications*, *12*(5), 1–8. https://doi.org/10.5121/ijma.2020.12501

Al Mamun, Md Abdullah and Ahmed, Mohammed Iftekhar and Bhuian, Mohammed Alauddin and Selim, Mohammed Riaz and Iqbal, Z. (2001). An Implementation of Machine Translation between Bangla and English. *Iccit*, *April*.

Al Mumin, M. A., Seddiqui, M. H., Iqbal, M. Z., & Islam, M. J. (2019). Neural Machine Translation for Low-resource English-Bangla. *Journal of Computer Science*, *15*(11), 1627–1637. https://doi.org/10.3844/jcssp.2019.1627.1637

Alam, F., Nath, P., & Khan, D. (2007). *Text to speech for Bangla language using festival*. *May 2014*.

Alom, M. Z., Sidike, P., Hasan, M., Taha, T. M., & Asari, V. K. (2018). Handwritten Bangla Character Recognition Using the State-of-the-Art Deep Convolutional Neural Networks. *Computational Intelligence and Neuroscience*, *2018*. https://doi.org/10.1155/2018/6747098

Asher, R.E. Simpson, JMY (edited). 1994. *The Encyclopedia of Language and Linguistics,* voll 1-10. Pergman press, England, p-4805.

BanglaSpellChecker_AStateoftheArt_BAUSTJ.pdf. (n.d.).

Basu, S., Kundu, M., & Nasipuri, M. (2012). *Development of OCR Techniques for Handwritten Bangla Text: OCR Techniques for Bangla Text*. *January*.

Biswas, M., & Hoque, M. M. (2019). Development of a Bangla Sense Annotated Corpus for Word Sense Disambiguation. *2019 International Conference on Bangla Speech and Language Processing, ICBSLP 2019*, *July 2020*. https://doi.org/10.1109/ICBSLP47725.2019.201516

Chowdhury, A. A., Ahmed, E., Ahmed, S., & Hossain, S. (2002). *Optical Character Recognition of Bangla Characters using neural network : A better approach*. *January 2002*.

Computer, I., Acoutic, D. I., Tts, C., Synthesis, S., & Tts, T. (2016). *Text to speech synthesis 20*. *January*.

Dash, N. S., & Arulmozi, S. (2018). History, features, and typology of language corpora. In *History, Features, and Typology of Language Corpora*. https://doi.org/10.1007/978-981-10-7458-5

Dash, N. S., & Ramamoorthy, L. (2018). Utility and application of language corpora. In *Utility and Application of Language Corpora*. https://doi.org/10.1007/978-981-13-1801-6

Debnath, D. (n.d.). *Bangla Character Recognition using Artificial Neural Network Step : Feature Selection Submitted by*.

Doermann, D., & Tombre, K. (2014). Handbook of Document Image Processing and Recognition. In *Handbook of Document Image Processing and Recognition* (Issue May). https://doi.org/10.1007/978-0-85729-859-1

Habib, S. M. M., & Khan, M. (n.d.). A High Performance Domain Specific. *Novel Algorithms and Techniques In Telecommunications, Automation and Industrial Electronics., 2008*, 174–178.

Hasan, K. M. A., Hozaifa, M., Dutta, S., & Rabbi, R. Z. (2014). A framework for Bangla text to speech synthesis. *16th Int'l Conf. Computer and Information Technology, ICCIT 2013, December*, 60–64. https://doi.org/10.1109/ICCITechn.2014.6997307

Hasan, M. A., Alam, F., Chowdhury, S. A., & Khan, N. (2019). Neural machine translation for the bangla-english language pair. *2019 22nd International Conference on Computer and Information Technology, ICCIT 2019, December*. https://doi.org/10.1109/ICCIT48885.2019.9038381

Hasnat, A., & Khan, M. (2009). *Elimination of Splitting Errors in Printed Bangla Scripts*.

Hosain Sumit, S., Al Muntasir, T., Arefin Zaman, M. M., Nath Nandi, R., & Sourov, T. (2018). Noise Robust End-to-End Speech Recognition for Bangla Language. *2018 International Conference on Bangla Speech and Language Processing, ICBSLP 2018, October*. https://doi.org/10.1109/ICBSLP.2018.8554871

Huque, S., Habib, A., & Babul, M. (2016). Analysis of a Small Vocabulary Bangla Speech Database for Recognition. *International Journal of Computer Applications*, *133*(6), 22–28. https://doi.org/10.5120/ijca2016907827

Islam, M. I. K., Habib, M. T., Rahman, M. S., Rahman, M. R., & Ahmed, F. (2018). A context-sensitive approach to find optimum language model for automatic Bangla spelling correction. *International Journal of Advanced Computer Science and Applications*, *9*(11), 184–191. https://doi.org/10.14569/ijacsa.2018.091126

Ismail, S., & Rahman, M. S. (2014). Bangla word clustering based on N-gram language model. *1st International Conference on Electrical Engineering and Information and Communication Technology, ICEEICT 2014, April*. https://doi.org/10.1109/ICEEICT.2014.6919083

Isthiaq, A., & Saif, N. A. (2020). OCR for printed bangla characters using neural network. *International Journal of Modern Education and Computer Science*, *12*(2), 19–29. https://doi.org/10.5815/ijmecs.2020.02.03

Karim, M. A., Kaykobad, M., & Murshed, M. (2013). Technical challenges and design issues in Bangla language processing. In *Technical Challenges and Design Issues in Bangla Language Processing*, p-xiv, 4, 6, 9, 11, https://doi.org/10.4018/978-1-4666-3970-6

Khairul Islam, M. I., Meem, R. I., Abul Kasem, F. Bin, Rakshit, A., & Habib, M. T. (2019). Bangla Spell Checking and Correction Using Edit Distance. *1st*

International Conference on Advances in Science, Engineering and Robotics Technology 2019, ICASERT 2019, 2019(Icasert), 1–4. https://doi.org/10.1109/ICASERT.2019.8934536

Khairullah, M., & Ratul, M. A. S. (2017). Steganography in Bengali Unicode text. *SUST Journal of Science and Technology*, *27*(1), 71–80.

Khan, D. M. F. (2017). Creation and Analysis of a New Bangla Text Corpus BDNC01. *International Journal for Research in Applied Science and Engineering Technology*, *V*(XI), 260–266. https://doi.org/10.22214/ijraset.2017.11038

Khatun, A., Rahman, A., Chowdhury, H. A., Islam, M. S., & Tasnim, A. (2020). *A Subword Level Language Model for Bangla Language*. 385–396. https://doi.org/10.1007/978-981-15-3607-6_31

Mahmud, N. Al. (2020). *Performance Analysis of Different Acoustic Features based on LSTM for Bangla Speech Recognition. 12*(1), 17–25.

Mittra, T., Nowrin, S., Islam, L., & Roy, D. C. (2019). A Bangla Spell Checking Technique to Facilitate Error Correction in Text Entry Environment. *1st International Conference on Advances in Science, Engineering and Robotics Technology 2019, ICASERT 2019*. https://doi.org/10.1109/ICASERT.2019.8934461

Mohaimen, A., & Chakraborty, D. (2017). *Bangla OCR Post Processing - Word Based Longest Common Subsequence Shahjalal University of Science and Technology Department of Computer Science and Engineering*. *February 2018*. https://doi.org/10.13140/RG.2.2.26377.75360

Mridha, M. F., Rana, M. M., Hamid, M. A., Khan, M. E. A., Ahmed, M. M., & Sultan, M. T. (2019). An Approach for Detection and Correction of Missing Word in Bengali Sentence. *2nd International Conference on Electrical, Computer and Communication Engineering, ECCE 2019*, *February*, 1–5. https://doi.org/10.1109/ECACE.2019.8679416

Mridha, M. F., Saha, A. K., Adnan, A., Hussein, M. R., & Das, J. K. (2015). Design and implementation of an efficient enconverter for bangla language. *ARPN Journal of Engineering and Applied Sciences*, *10*(15), 6543–6548.

Nasir, M. K. (2013). Hand Written Bangla Numerals Recognition for Automated Postal System. *IOSR Journal of Computer Engineering*, *8*(6), 43–48. https://doi.org/10.9790/0661-0864348

Nayeem, A., Rahaman, R., Das, A., Nayen, M. Z., & Rahman, S. (2011). *A Novel Technique of Segmentation and Features Extraction of Bangla Speech for Speech to Text Conversion*. 46–47.

Rahman, M. M., Roy Dipta, D., & Hasan, M. M. (2018). Dynamic Time Warping Assisted SVM Classifier for Bangla Speech Recognition. *International Conference on Computer, Communication, Chemical, Material and Electronic Engineering, IC4ME2 2018*. https://doi.org/10.1109/IC4ME2.2018.8465640

Raju, R. S., Bhattacharjee, P., Ahmad, A., & Rahman, M. S. (2019). A Bangla Text-to-Speech System using Deep Neural Networks. *2019 International Conference*

on Bangla Speech and Language Processing, ICBSLP 2019. https://doi.org/10.1109/ICBSLP47725.2019.202055

Rashid, M. M., Hussain, M. A., & Rahman, M. S. (2009). Diphone preparation for Bangla text to speech synthesis. *ICCIT 2009 - Proceedings of 2009 12th International Conference on Computer and Information Technology, August 2018*, 226–230. https://doi.org/10.1109/ICCIT.2009.5407135

Rashid, M. O. (2020). *Bangla Language Pre-processing Phases : Noise Removal and Normalization Bangla Language Pre-processing Phases : Noise Removal and Normalization*. *December*. https://doi.org/10.13140/RG.2.2.14919.11686

Rownak, A. F., Rabby, M. F., Ismail, S., & Islam, M. S. (2016). An efficient way for segmentation of Bangla characters in printed document using curved scanning. *2016 5th International Conference on Informatics, Electronics and Vision, ICIEV 2016*, 938–943. https://doi.org/10.1109/ICIEV.2016.7760138

Roy, S. (2019). Improved Bangla Language Modeling with Convolution. *1st International Conference on Advances in Science, Engineering and Robotics Technology 2019, ICASERT 2019, January*. https://doi.org/10.1109/ICASERT.2019.8934469

Sarkar, A. I., Shahriar, D., Pavel, H., & Khan, M. (n.d.). *Automatic Bangla Corpus Creation*.

SAYEM, A. (2014). Speech Analysis for Alphabets in Bangla Language: Automatic Speech Recognition. *International Journal of Engineering Research*, *3*(2), 88–93. https://doi.org/10.17950/ijer/v3s2/211

Selim, M. R., & Iqbal, M. Z. (2014). SUMono : A Representative Modern Bengali Corpus. *SUST Journal of Science and Technology*, *21*(1), 78–86.

Selim, M. R., & Ismail, S. (2014). *Mapping Bangla Unicode Text to Keyboard Layout Specific Keystrokes Mapping Bangla Unicode Text to Keyboard Layout Specific Keystrokes*. *January 2009*.

Shamim Ahmed, M. A. K. (2013). Enhancing the Character Segmentation Accuracy of Bangla OCR using BPNN. *International Journal of Science and Research (IJSR)*, *2*(12), 157–161. https://www.ijsr.net/archive/v2i12/MDIwMTM1NTk=.pdf

Sharif, O., & Hoque, M. M. (2020). Automatic Detection of Suspicious Bangla Text Using Logistic Regression. In *Advances in Intelligent Systems and Computing* (Vol. 1072, Issue September). Springer International Publishing. https://doi.org/10.1007/978-3-030-33585-4_57

Siddique, S., Ahmed, T., Rifayet Azam Talukder, M., & Mohsin Uddin, M. (2020). English to Bangla Machine Translation Using Recurrent Neural Network. *International Journal of Future Computer and Communication*, *9*(2), 46–51. https://doi.org/10.18178/ijfcc.2020.9.2.564

Sultana, S., Akhand, M. A. H., Das, P. K., & Hafizur Rahman, M. M. (2012). Bangla Speech-to-Text conversion using SAPI. *2012 International Conference on Computer and Communication Engineering, ICCCE 2012, July*, 385–390. https://doi.org/10.1109/ICCCE.2012.6271216

Tausif, M. T., Chowdhury, S., Hawlader, M. S., Hasanuzzaman, M., & Heickal, H. (2018). Deep Learning Based Bangla Speech-to-Text Conversion. *Proceedings - 5th International Conference on Computational Science/Intelligence and Applied Informatics, CSII 2018*, *7*(3), 49–54. https://doi.org/10.1109/CSII.2018.00016

Uzzaman, N., & Khan, M. (2006). A comprehensive bangla spelling checker. *BRAC University*.

Yeasmin Omee, F., Shabbir Himel, S., & Naser Bikas, A. (2011). A Complete Workflow for Development of Bangla OCR. *International Journal of Computer Applications*, *21*(9), 1–6. https://doi.org/10.5120/2543-3483

Websites

(PDF) A Complete Workflow for Development of Bangla OCR. (n.d.). Retrieved February 15, 2021, from https://www.researchgate.net/publication/222109100_A_Complete_Workflow_for_Development_of_Bangla_OCR

maxresdefault.jpg (1280×720). (n.d.). Retrieved February 15, 2021, from https://i.ytimg.com/vi/k1squGjesUs/maxresdefault.jpg

500px-KB-Bengali-Shahidlipi.svg.png (500×167). (n.d.). Retrieved February 15, 2021, from https://upload.wikimedia.org/wikipedia/commons/thumb/e/e7/KB-Bengali-Shahidlipi.svg/500px-KB-Bengali-Shahidlipi.svg.png

article_0101_2_845.jpg (845×475). (n.d.). Retrieved February 15, 2021, from https://www.wipo.int/export/sites/www/ipadvantage/images/article_0101_2_845.jpg

UniBijoy-Layout.jpg (1600×608). (n.d.). Retrieved February 15, 2021, from https://priozone.com/wp-content/uploads/2019/12/UniBijoy-Layout.jpg

Ridmik Keyboard for Android - APK Download. (n.d.). Retrieved February 15, 2021, from https://apkpure.com/ridmik-keyboard/ridmik.keyboard

screen-4.jpg (480×800). (n.d.). Retrieved February 15, 2021, from https://image.winudf.com/v2/image1/Y29tLm1heWFiaXNvZnQuaW5wdXRtZXRob2QubGF0aW5fc2NyZWVuXzRfMTU2Njk5NzA5Nl8wNzE/screen-4.jpg?fakeurl=1&type=.jpg

www.fokia.tk : Avro Keyboard 5.1.0.0: Free Download. (n.d.). Retrieved February 15, 2021, from https://sifatblog.blogspot.com/2012/10/a-full-unicode-supported-bangla-typing.html

চিত্র:Avro Phonetic Keyboard Layout.png - উইকিপিডিয়া. (n.d.). Retrieved February 15, 2021, from https://bn.wikipedia.org/wiki/চিত্র:Avro_Phonetic_Keyboard_Layout.png

চিত্র:KB-Bengali-Probhat.svg - উইকিপিডিয়া. (n.d.). Retrieved February 15, 2021, from https://bn.m.wikipedia.org/wiki/চিত্র:KB-Bengali-Probhat.svg

Friederici, A. D. (2011). *Anatomical, and cytoarchitectonic details of the left*

hemisphere. Retrieved February 15, 2021, from https://doi.org/https://doi.org/10.1152/physrev.00006.2011

Brodmann areas: Anatomy and functions | Kenhub. (n.d.). Retrieved February 15, 2021, from https://www.kenhub.com/en/library/anatomy/brodmann-areas

Natural Language Processing steps | Download Scientific Diagram. (n.d.). Retrieved February 15, 2021, from https://www.researchgate.net/figure/Natural-Language-Processing-steps_fig1_311705165

Bijoy - Unicode Converter | বিজয় - ইউনিকোড কনভার্টার. (n.d.). Retrieved February 15, 2021, from https://bsbk.portal.gov.bd/apps/bangla-converter/index.html

google translate - Google Search. (n.d.). Retrieved February 15, 2021, from https://www.google.com/search?q=google+translate&oq=google+&aqs=chrome.0.69i59j69i57j69i59j0i271l2j69i60l3.4070j0j15&sourceid=chrome&ie=UTF-8

i2OCR - Free Online Gujarati OCR. (n.d.). Retrieved February 15, 2021, from https://www.i2ocr.com/free-online-gujarati-ocr

IPA. (n.d.). Retrieved February 15, 2021, from https://nlp.egeneration.co/

Speech to Text Bengali | Type by Speak in Bengali | Bengali Speech Typing. (n.d.). Retrieved February 15, 2021, from https://speechtyping.com/voice-typing/speech-to-text-bengali

বাংলা ভাষা প্রকল্প. (n.d.). Retrieved February 21, 2021, from https://bcc.gov.bd/site/page/74683337-931a-4344-ab00-34cf2527acc6/বাংলা-ভাষা-প্রকল্প

APPENDIX I

Interview and deliberation/colloquium

Daniul Huq, Professor (Retd.) of Bangla Department of Jahangirnagar University:

Bangla Language Processing in Bangladesh seems plodding progress in Bangladesh compared to the other languages like Arabic, Chinese, Japanese, etc. We have no compatible keyboard layout and fonts. Other tools development is at in amateur level. In a recent workshop, we found around 300 conjuncts of the Bangla language, which are not added to the software. Avro is a syllabic keyboard, so we cannot write many letter combinations. As it has spelling suggestions, the user can get confused whether the correct one. It does not support Adobe Photoshop, Illustrator, etc.

Muhammad Zafar Ikbal, Professor of CSE, Shahjalal University of Science and Technology:

It is my great pleasure to let you know that six students are researching Bangla Language Processing at SUST. We are to take assist from a linguist in every step. We have almost all kinds of technology available for the computing of the Bangla Language. As the development of the corpus is time-consuming, we tell the researcher to make their datasets. For example, -a respondent utters ten sentences in ten different tones to represent naturalness for Speech to Text converter. They make sure the high-quality voice.

On the other hand, Bangladesh Computer Council is working very well on this platform. Government fund Tk 160cr for the project. Half of the budget is for making a standard corpus of the Bangla language. Though sometimes, they change their mindset to go abroad for a better opportunity. Afterward, they did not come back to Bangladesh.

Abdullah-Al Mumin, Professor of CSE, Shahjalal University of Science and Technology:

Lack of corpus is one of the causes that we are not getting convenient tools. We developed a text corpus, Sumono, the largest of its kind at SUST, consisting of 27 million words. We had developed another English-Bangla parallel corpus, SUPara, at SUST that consists of 2,000000 words either in languages. We are collaborating with

some of the Professors of Bangla at SUST. We need more collaboration. Government or Institutions should increase the funding opportunities. However, we fund some Ph.D. students through the ICT ministry. Now, I am working on collecting audio from the Sylhet dialect. I plan to develop a corpus on 'Rabindranath Tagore' like a corpus developed on 'Shakespeare,' and write a book on Bangla Language Processing. Professor Niladri Shekhor Dash of ISI at India did excellent work in this field. If I get a chance to collaborate with linguists, we can develop some necessary applications on BLP. We will publish our datasets and tools on websites as open-source resources.

Shakhawat Ansari, Professor of linguistics, University of Dhaka:

It is a matter of sorrow that we do not give much importance to Bangla Language in Bangladesh. Still, we do not have any language planning and policy. There is no incorporation between spelling and articulation of the Bangla Language. If incorporation is not possible, then there will be more option for a single word like 'মরীচিকা'. The government can make a highest-body who will work for incorporating Bangla language. Experts from Bangla, English, and Sanskrit will be the members of the committee. The name of the highest authority can be 'VASHA COMMISSION' or a relevant term. Bangla Academy is working on Bangla Language Development, but there are no linguists of Bangla Language. We did not get a good number of linguists of Bangla Language. We are working on Indigenous languages, but the mainstream's language problems should be solved first.

Niladri Shekhar Dash, Linguistic Research Unit, Indian Statistical Institute:

Corpus is one of the significant areas for BPL. We do not have a complete and representative Bangla Corpus in Bangladesh and West-Bengal as well. It is very time-consuming to develop a representative corpus. If we want to build a speech corpus of the Bangla Language, we must consider things like speech annotation correctly. We need high-quality tools for recording high-quality voices. There are 16 different components for a speech like intonation, tone, pitch, length, etc. We need to analyze those components for sentiment analysis. In Bangladesh, I have been invited to collaborate with a BLP project. Where one chapter of my book recommended corpus development guidelines. Now, Bangladesh Computer Council is working on a project. They are trying to develop many tools—some works. There are many challenges in

developing BLP tools. For example-Technical (not available for the Bengali Language)

Development of Bengali font technology, Text processing tools, Text Annotation tools and techniques, Development of Bengali font technology, Data archiving of texts, Text Normalization techniques, Text-to-Speech/Speech-To-text, Machine Translation tools, Speech Recognition tools, Sentiment Analysis and Recognition, Hate Speech analysis and Recognition, Verbal Aggression recognition, and many more. On the other hand, Linguistic (Not available for the Bengali Language)- Standard and well-representative text corpus, Standard and well-representative Speech corpus, Parallel text corpus, Multimodal Text, and Speech Corpus, Annotated Text and Speech, Corpus Processed Text and Speech Corpus, Digital Lexical databases, Frequency-based Lexical Databases, Concordance List of most frequently used words, Parsed Syntactic Databases, Digital dictionaries and thesauri Digital Bengali grammar and many more. We can develop the tools above since complex scripts like Chinese, Japanese, Arabic (Writing systems different from Bangla and English) have already created them.

Nabeel Mohammad, Associate Professor of CSE, North South University:

While we go to work on Bangla Language, the problem we face is the lack of datasets. Suppose we want to make a corpus where 100 linguists will work and 2-3 years need to finish the canon. But we cannot move forward because of a good number of linguists. We can develop tools for the Bangla language but, we need validation from linguists. We need sufficient funds for creating tools. To represent the naturalness of devices, we need access to a different context. But some places are restricted. For example- we cannot use the language of our Parliament easily. Diversity and volume of datasets are other areas to explore. We managed data in one-dimension to train the system, but linguists work data in many dimensions. Lack of collaborations seems another problem in developing BLP tools. Dependency parsing is one of the challenging tasks for linguists to analyze.

Mamun Or Rashid, Associate Professor of Bangla Department at Jahangirnagar University and Language Consultant at Bangladesh Computer Council:

I am working on a project at Bangladesh Computer Council as a Language Consultant. Hence, we are trying to develop some BLP tools. We already developed

spell and grammar checkers, OCR, IPA transcription software, Sentiment analysis, Machine Translation based on ten languages. We are collaborating with the technologist. Nearly ten teachers of the linguistics department are working on this BLP project funded by the government. We are trying to create a convenient Bangla Keyboard layout. It is impressive that the software company named 'Vendor' found nearly eight lakh unique tokens of the Bangla language. There are some problems in negotiation with Unicode Consortium; for example- they directed us to write 'য় ড় ঢ়' as first we have to insert 'য ড ঢ', then we have to put '.' (Nukta). But, we proposed to Unicode Consortium to include 'য় ড় ঢ়' as a single/atomic character. We should work hard for dependency parsing as it may solve many problems on BLP.

Mashrur Imtiaz, Assistant Professor of Linguistics Department, University of Dhaka:

Bangla Language Processing still needs to be focused remarkably. Once upon a time, CRBLP started working with the primary hub of Bangla Language, but it stopped because of funding. There is a debate on keyboard layout designing, but I think it depends on the skill; for example- I am habituated with typing Avro keyboard, so now I did not find any problems typing Bangla Language. I can also type with Bijoy. Sometimes, we compare the Bangla Writing system with typing, but we should keep in mind that typing is a different platform, so the wring style may vary. Ridmik or Avro got famous on the web for writing Bangla. Now professionally, the UniBijoy layout is being used as well. So, we can make incorporate of keyboard layout. We could not develop user-friendly software because of the corpus. Bangladesh Computer Council is trying to establish a canon, but it is in the initial phase. We can start corpus with standard language, so we need to create the base from written documents. The commonly used fonts are SutonnyMJ, Kalpurush, Shiyam Rupali, Nikosh, etc. Some Newspaper uses their fonts like Prothom Alo uses the 'Prothoma' font. All the newspapers are using Unicode-based font. Overall, we need collaboration to overcome these problems. Still, somehow our programmer thinks language is a simple entity, but language itself is a complex entity, so we need to analyze the languages for computation.

Sajjadul Islam, Professor of Bangla Department, Jahangirnagar University:

We do not see software like Bangla Pronunciation. It is essential to have this kind of tool for our students and users of the Bangla Language. In Bangla Academy, they published many dictionaries on Bangla Language. But, as per my knowledge, there are no linguists in their committee. We should include a journalist in such a committee and know that newspaper represents the everyday language. We have no original Bangla Grammar yet. We could not solve a problem like which one we write 'রেস্টুরেন্ট' or 'রেস্তোরাঁ'. 'রেস্টুরেন্ট' is the English word 'রেস্তোরাঁ' is a French word. So, we need the incorporation of using words. We could not make a digital dictionary. But it is possible to write a digital one.

Tarik Manzoor, Professor of Bangla Department, University of Dhaka:

We see a few numbers of BLP tools available in Bangladesh. I do not have much experience using those available tools. We have no spell checker in Bangla. It's is a necessary tool for use. Bijoy or Avro can develop this as an inbuilt tool. There are many anomalies in font display. I am using Bijoy Bayanno 2020, where some anomalies like If we type unconsciously **'চু,'** and then delete **'ু'**. Afterward**,** add **'ু'**, then the modified vowel goes a little far from the letter. To avoid this problem, we had to delete whole 'চু,' and retype the 'চু'. It is time-consuming. A developer also develops a system where automatically space comes after punctuation marks.

Bangla Language Processing cannot move forward because of a lack of incorporation in spelling, pronunciation, etc. There might be 2 or 3 options for word spelling. The same thing can best apply to accents.

Mohammad Azam, Professor of Bangla Department, University of Dhaka:

Many researchers are conducting BLP. But we could not solve the problem whether we use Bangla or foreign words in our language. So, our Bangla grammar is not complete yet. We use Bangla words but add Sanskrit suffix. In this case, how we differentiate words. I think it's better to treat terms used in Bangla as Bangla words. Our country has got 3 or 4 renowned linguists. But we need more. The linguistic department should come forward to research in this field. A committee should be formed under Bangla Academy who will work for the development of the Bangla Language. Some letters will not use shortly like- 'ৎ, ণ, ষ' etc.

Rifat Hassan Jihad, Mechanical Engineering, KHUET:

The main challenge I faced while developing a Bengali layout is to map the মহাপ্রাণ (Aspirated) & অল্পপ্রাণ (Unaspirated) phoneme according to keystroke frequency. On most of the existing designs, these types of characters are mapped under one key. Again, all vowel letters have to be mapped under the Alt-Gr state to keep all keys in a uniform style. It is logical, but it is not possible to maintain the keystroke frequency efficiency in this style. All vowels can be generated with the Hasanta key, but the Dead Key feature is not available in Linux-based OSs. Mapping Bengali ligatures in a layout is problematic. As the current input method system (XKB) in Linux-based OSs does not support this kind of key-mapping.

Arif Ahmed, Information science, Leading University, Sylhet:

The 1st challenge is a good dataset/corpus. Since most of the researches is now being conducted in deep learning methods, a suitable dataset is essential. We didn't have any, so we had to prepare our own, which took a lot of time—2nd challenge: computing resources. Again, deep learning algorithms need to be run on powerful machines. Luckily, our department provided us a GPU-based server for our experiments. Lack of previous researches: There are not many pieces of research done in Bangla with state-of-the-art techniques. So, we couldn't take much from the literature review.
Expert linguists: Although we didn't need in-depth linguistic knowledge for our tasks, I would appreciate some help from linguists at the earlier stage of my studies.

A. S. M. Shakil Haider, Ph.D. candidate at Texas Tech University, USA:

Natural Language Processing is an exciting field. But Bangla Language Processing is in the primary stage. I use some tools to write Bangla on the web. Usually, I use ‘SARCHAKKRA’ for writing Bangla. There is no spelling correction option in that interface. It would be better if the developer adds this facility. Pronunciation of words is essential so that it might be added as well. Many conjuncts are not clear while writing in Bangla. Professionally 90% of people use ASCII format, but it does not support many platforms like google meet, Turnitin, etc. I searched tools like speech to text and vice versa in the play store, but most of them are worthless. The researcher should come forward in this field.

Sabbir Ahmed, Final Year Student, CSE of North South University:

Dhaka University Institutional Repository

As a student of CSE, we are to complete many projects. Some students take projects on Bangla Language Processing; for example- I am now developing a system where voice can be verified. The main challenge is to find out the perfect model to apply. We can create many tools on BL by some available datasets. But we should identify first what model goes with what datasets. Many people think that Bangla coding is significant for computing Bangla Language, but it is not valid. The model applicable is necessary to know first. Coding in Bangla is possible, but the international standard should maintain in this case. We do not get good research because some students do the project to get a good grade. The quality of their task is deficient. So, we do not have much knowledge of linguistics, so we need collaboration to produce a validated tool.

Sanjid Alom, Arabic department, University of Rajshahi:

I have some experience using BL (Bangla Language) apps. Especially, speech to text software. The problem is that sometimes, it does not work correctly. It got hang or stop. Bangla fonts are not marginable with English font. More apps on Bangla grammar should be developed. There is an alignment problem while Bangla and English are written together in the same line. Sometimes, words are typed in Bold format. It cannot be changed in a regular form easily. Fonts are broken while changing the MS word

versions. Sometimes, the command does not work like if I want to bold some Bangla Typeface, then it remains normal and vice-versa. Some Bangla ligatures and conjuncts are not clear in different fonts. There are problems in the aesthetic view of some fonts.

Sayada Jahan, MEd (continuing) in Education, IER, University of Dhaka:

Many Bangla Language Processing tools are now available in the play store. All are not supported on all platforms. Some apps need the updated android version. I downloaded a software text to speech, but it did not work correctly. Even typed Bangla in windows 19 does not support well in the earlier version. When I transfer Microsoft doc from 19 to 2007, then all the Bangla Words glue to each other, for example- 'আমিতোমাকেভালোবাসি.' Sometimes, Bangla scripts break down into different versions of MS word.

Dhaka University Institutional Repository

On the other hand, in the excel sheet same font in the same size seems different in some columns. We should immediately develop tools like Bangla spell and grammar checker. We are going through a pandemic situation where students participate in online classes, but there are no Bangla compatible tools like Zoom, google meet, or Duo. Some devices are paid versions like spell checker of CRBLP, Bijoy of Ananda Computers. These tools should be publicly accessible for further research also.

Dhaka University Institutional Repository

Appendix II

Some BLP Tools

Unicode Supported Fonts:

1. Amar-Desh	32. NikoshLight-Ban
2. Adorsholipi	33. NikoshBAN
3. Aponalothi	34. NikoshGrameen
4. Akashnormal	35. Nikosh Light
5. Amar-Bangla	36. Nikosh
6. Azad	37. Punatbhaba
7. Amar-Bangla-Bold	38. Rupali
8. Bangla-Kolom-Bold	39. Sornaly
9. Bangla	40. Sulekha Bangla Font
10. Bensen Handwriting	41. Sumit Bangla Font
11. Buriganga	42. Sumon Unicode Madhob Bangla Font
12. Bangla-Kolom	
13. Bensen	43. Sumon Unicode Rajani Bangla Font
14. 16-December	
15. Ekushey-kolom	44. Sumon Unicode Slipi Bangla Font
16. Ekushey-sumon	
17. Ekushey-Kolom.Bold	45. Sutom Bangla Font
18. Ekushey-Bangla	46. Solaimani Lipi
19. Godhuli	47. SuttonNy Bangla Font
20. Ghorautro	48. SolaimanLipi_Bold
21. Kalpurush	49. SolaimanLipi_20-04-07
22. Likhon Normal	50. SolaimanLipi_22-02-2012
23. Lal-Salu	51. SolaimanLipi_29-05-06
24. Lal-Sabuj	52. Sagar-Normal
25. Lal-sabuj-Normal	53. Sornali
26. Mitra	54. Sharifa
27. Muktinarrow	55. Sumit
28. Mohua	56. Siyamrupali
29. Mukto	57. Shimanto
30. Mukti	58. SutonnyMJ
31. Mukti_1.99_PR	59. Vrinda

Table 1: Unicode supported fonts

Dhaka University Institutional Repository

Some Latest fonts developed by 'LIPIGHAR'

মায়াবী RRR

রিফাত আর রহমান

আমার সোনার বাংলা

ইউনিকোড এবং ANSI | 2 স্টাইল

নিলাদ্রী হ্যালহেড

নিলাদ্রী শেখর বালা

আমার সোনার বাংলা

তোহা বর্ণ

তৌফিকুর রহমান

আমার সোনার বাংলা

ইউনিকোড এবং ANSI | 4 স্টাইল

খালিদ মেঠো পথ

এইচ. এম. খালিদ

খালিদ মেঠোপথ

ইউনিকোড এবং ANSI | 2 স্টাইল

আজাদ ফেনী

আবুল কালাম আজাদ

আজাদ ফেনী

ইউনিকোড এবং ANSI | 2 স্টাইল

খালিদ কালকিনি

Dhaka University Institutional Repository

এইচ. এম. খালিদ

ইউনিকোড এবং ANSI | 2 **স্টাইল**

খালিদ মিয়ারহাট

এইচ. এম. খালিদ

ইউনিকোড এবং ANSI | 2 **স্টাইল**

বালু দা

Baloo DaCreated by: Noopur Datye and Ek TypeUnicode3 styles

নোটো সান্‌স বেঙ্গলি

Noto Sans BengaliUnicode13 styles

সিয়াম রুপালি

Siyam RupaliCreated by: Md. Tanbin Islam SiyamUnicode1 style

সাগর

SagarUnicode1 style

নিকশ

NikoshUnicode1 style

কালপুরুষ

KalpurushCreated by: Md. Tanbin Islam Siyam Unicode1 style

হিন্দ শিলিগুড়ি

Hind SiliguriUnicode5 styles

গালাদা

GaladaCreated by: Latin by Pablo Impallari, Bengali by Jeremie Hornus, Yoann Minet, and Juan Bruce Unicode1 style

একুশে গোধূলি

Ekushey GodhuliUnicode1 style

বেনসেন

BenSenCreated by: Subrata SenUnicode1 style

বাংলা

BanglaUnicode1 style

আত্মা

অপনালোহিত

Dhaka University Institutional Repository

AponaLohitUnicode1 style

আকাশ

AkaashUnicode1 style

আদর্শলিপি

একুশে সুমিত

Ekushey SumitUnicode1 style

একুশে শরিফা

Ekushey SharifaUnicode1 style

একুশে সরস্বতী

একুশে লোহিত

Ekushey LohitUnicode1 style

একুশে দুর্গা

Ekushey DurgaUnicode1 style

একুশে আজাদ

Ekushey AzadUnicode1 style

সোলাইমানলিপি

SolaimanLipiCreated by: Solaiman KarimUnicode1 style

নোটো সেরিফ্ বেঙ্গলি

Noto Serif BengaliCreated by: Indian Type FoundryUnicode2 styles

নোটো সান্স বেঙ্গলি ইউআই

Noto Sans Bengali UIUnicode9 styles

মুক্তি

MuktiCreated by: Dr Anirban MitraUnicode1 style

মিনা

MinaUnicode2 styles

চারুকলা ইউনিকোড

Charukola UnicodeCreated by: Chandan AcharjaUnicode6 styles

Dhaka University Institutional Repository

চারুকলা রাউন্ডেড হেড ইউনিকোড

Charukola Round Head UnicodeCreated by: Chandan AcharjaUnicode

চারু চন্দন হার্ডস্ট্রোক ইউনিকোড

CharuChandan HardStroke UnicodeCreated by: Chandan AcharjaUnicode4 styles

চারু চন্দন ইউনিকোড

Charu Chandan UnicodeCreated by: Chandan AcharjaUnicode4 styles

চারু চন্দন ব্লাড ড্রিপ ইউনিকোড

Charu Chandan BloodDrip UnicodeCreated by: Chandan AcharjaUnicode4 styles

চারু চন্দন থ্রিডি ইউনিকোড

Dhaka University Institutional Repository

Some Bangla apps of Android phone:

There are many Bangla apps are available in android or smart phone. Some app's name has been given below-

1. Bangla to English Dictionary
2. English to Bangla Dictionary
3. Bangla Speech to Text
4. Bangla Text to Speech
5. Banga OCR
6. Puthi OCR
7. Subachan speech to text
8. Bangla Dictionary offline
9. Bangla Panini keyboard
10. Bangle SMS android app
11. Bangla calendar android app
12. Bangla currency converter android app
13. Bangla vai android app
14. Bangla Rashifol android app
15. Bangle calculator
16. Recipe (Bangla)
17. Bangladeshi National Anthem
18. Maths for kid in Bengali
19. English to Bengali flashcards vocabulary app
20. Bijoy keyboard app
21. Avro keyboard app
22. Bangla-TV android app
23. Mayabi Keyboard
24. Ridmik Keyboard

Bengali Spelling Checker:

1. Nikosh Bangla
2. Ankur
3. Mozilla add-on
4. Avro
5. Shuddhoshabdo
6. Srishty

Bangla Keyboard Layout:

Dhaka University Institutional Repository

2. Shahid Lipi 2. Inscript Keyboard 3. Jatiyo keyboard 4. Probhat 5. Ekushey keyboard 6. Munier keyboard 7. Bijoy 8. Uni Gitanjali 9. Rupali keyboard 10. UniBijoy keyboard 11. Ridmik keyboard 12. Mayabi keyboard 13. Bangla word keyboard 14. Baishakhi keyboard 15. Avro keyboard

Dhaka University Institutional Repository

<table>
<tr><th>ক্রমিক নং</th><th>বিষয়বস্তু</th><th>বিবরণ</th></tr>
<tr><td>১</td><td>প্রকল্পের নাম</td><td>গবেষণা ও উন্নয়নের মাধ্যমে তথ্য প্রযুক্তিতে বাংলা ভাষা সমৃদ্ধকরণ প্রকল্প</td></tr>
<tr><td>২</td><td>মেয়াদ</td><td>জুন ২০১৬ - জুন ২০২১</td></tr>
<tr><td>৩</td><td>প্রাক্কলিত ব্যয়</td><td><table><tr><th>অর্থের উৎস</th><th>পরিমান</th></tr><tr><td>জিওবি</td><td>১৫৮৯৬.৬৯ (লক্ষ টাকায়)</td></tr><tr><td>মোট</td><td>১৫৮৯৬.৬৯ (লক্ষ টাকায়)</td></tr></table></td></tr>
<tr><td>৪</td><td>জনবল</td><td>কমকর্তা: ১১ জন
কর্মচারী: ৭ জন</td></tr>
<tr><td>৫</td><td>পটভূমি</td><td>বাংলা ব্যবহারের দিক থেকে পৃথিবীতে প্রভাবশালী ভাষাগুলোর একটি। বাংলা ভাষাভাষীর রয়েছে রক্তস্নাত ভাষা-আন্দোলনের ইতিহাস। দেশ ও ভাষার মর্যাদা রক্ষায় এই জাতির রয়েছে গৌরবময় ঐতিহ্য, রয়েছে ভাষার প্রতি দরদ, ভাষাকে সমুন্নত রাখার চেতনা। কিন্তু দুঃখজনক হলেও সত্য, বাংলা ভাষাকে প্রযুক্তি বান্ধব করার ক্ষেত্রে প্রয়োজনীয় ভিত্তি তৈরি হয়নি, বিশেষ করে কম্পিউটিংয়ে বাংলা ভাষাকে অভিযোজিত করার ক্ষেত্র-- খুব বেশি অগ্রসর হয়নি । আনন্দের বিষয় যে, গণপ্রজাতন্ত্রী বাংলাদেশ সরকার আন্তর্জাতিক পর্যায়ে বাংলা ভাষাকে তুলে ধরার জন্য ‘গবেষণা ও উন্নয়নের মাধ্যমে তথ্য প্রযুক্তিতে বাংলা ভাষা সমৃদ্ধকরণ’ প্রকল্প অনুমোদন করায় বর্তমানে একটি কর্মযজ্ঞ শুরু হয়েছে।
কর্মযজ্ঞ সুসম্পন্ন হলে আশা করা যায়, মুখে উচ্চারিত বাংলা ভাষা স্বয়ংক্রিয়ভাবে কম্পোজ হয়ে যাবে, লিখিত টেক্সট কম্পিউটার পড়ে শোনাবে, মুদ্রিত বই-দলিল দ্রুত সফটকপিতে রূপান্তরিত হবে, বাংলা ভাষা সঠিক যান্ত্রিক অনুবাদ পাওয়া যাবে, বাংলা ভাষার বিশাল মৌখিক ও লিখিত নমুনা (করপাস) গড়ে উঠবে। এমন ১৬ টি সুবিধা ১৬ টি উপাংশের মাধ্যমে সম্পন্ন করা হবে এই প্রকল্পে।
দেশকে ‘ডিজিটাল বাংলাদেশ’ হিসেবে গড়ে তোলার একটি প্রধান শর্ত বাংলা ভাষাকে প্রযুক্তিবান্ধব করা। তথ্য ও যোগাযোগ প্রযুক্তিতে বাংলা ভাষার ব্যবহার ও প্রয়োগ হলে দেশের প্রশাসনিক ব্যবস্থা, শিক্ষা ব্যবস্থা ও যোগাযোগ কাঠামোতে নতুন পরিবর্তন সূচিত হবে। আন্তর্জাতিক পর্যায়ে বাংলা ভাষাকে যথাযথ মর্যাদাদান ও উৎকর্ষে পৌঁছানো সম্ভব হবে।</td></tr>
<tr><td>৬</td><td>লক্ষ্য ও উদ্দেশ্য</td><td>প্রকল্পের লক্ষ্য ও উদ্দেশ্য হচ্ছে আন্তর্জাতিক পরিসরে (Global Platform-এ) নেতৃস্থানীয় ভাষা হিসেবে বাংলাকে প্রতিষ্ঠা করা। বিশেষ করে, কম্পিউটিং ও আইসিটিতে বাংলা ভাষাকে অভিযোজিত করা বা খাপ খাইয়ে নেয়া- এ প্রকল্পের প্রধান উদ্দেশ্য। গবেষণা ও উন্নয়নের মাধ্যমে তথ্য প্রযুক্তিতে বাংলা ভাষা সমৃদ্ধকরণের লক্ষ্যে প্রকল্পটির উদ্দেশ্য বাংলা ভাষার জন্য বিভিন্ন প্রযুক্তিমাধ্যমে (ওয়েব, মোবাইল, কম্পিউটার) ব্যবহারযোগ্য বিভিন্ন সফটয়্যার/টুলস/রিসোর্স উন্নয়ন করা, যাতে বাংলা ভাষা কম্পিউটারে ব্যবহার করতে কোনো প্রতিবন্ধকতা না থাকে।</td></tr>
<tr><td>৭</td><td>প্রকল্পের উল্লেখ্যযোগ্য কম্পোনেন্ট</td><td>বাংলা ভাষার জন্য ১৬টি সফটয়্যার/টুলস/রিসোর্স উন্নয়ন করা হবে। এরফলে আন্তর্জাতিক পর্যায়ে বাংলা ব্যবহারের সুযোগ তৈরি হবে। সম্পূর্ণ বাংলা করপাস এবং বাংলা স্টাইল-শিট সম্পন্ন হলে বিশ্বমানের বাংলা কম্পিউটিং-এর ভিত্তি তৈরি করা যাবে। ১৬টি উপাংশের সংক্ষিপ্ত পরিচয় বাংলা ও ইংরেজি শিরোনামসহ (ইংরেজি মূল, বাংলা সহজে অনুধাবনের জন্য ঈষৎ পরিমার্জিত) নিম্নে উল্লেখ করা হলো :

ক) আন্তর্জাতিক মান বজায় রেখে পূর্ণাঙ্গ বাংলা করপাস উন্নয়ন (Development of Complete Bangla Corpus following international Standard)
এই উপাংশের মাধ্যমে বাংলা ভাষার বিশাল সংগ্রহশালার নমুনা প্রস্তুত করা হবে। এই করপাসের মধ্যে বাংলা শব্দভাণ্ডার, বাংলা বাক্য সংগ্রহ, মধ্যযুগ ও আধুনিক যুগের সাহিত্যের পূর্ণাঙ্গ ভাষিক তথ্য, বাংলাদেশে প্রকাশিত সকল ও সাময়িকীর তথ্য, ব্রিটিশ পর্ব থেকে বর্তমান পর্যন্ত দলিলের ভাষিক তথ্য, বাংলাদেশের প্রমিত ভাষা ও বিভিন্ন অঞ্চলের মানুষের মৌখিক ভাষার তথ্য অন্তর্ভুক্ত। আরো বলা যায়, বাংলা করপাস নিশ্চিত করবে সঠিক বাংলা ভাষার শব্দের উচ্চারণ, সব উচ্চারণ বিবরণ ও ফরম্যাটিং এর নিদর্শন।
বিশাল এই ভাষিক তথ্য ভান্ডার সহজে ফিল্টার, সর্টিং ও সার্চিং করার ব্যবস্থা এবং বিভিন্ন ডিজিটাল মাধ্যমে পঠনযোগ্য অবস্থায় থাকবে।
করপাসের শব্দভাণ্ডারের POS ট্যাগিংসহ কম্পিউটিশনাল ভাষাবিজ্ঞান সমর্থিত অন্যান্য ধাপগুলো সম্পন্ন করা থাকবে। যা পরবর্তী সময় বিভিন্ন ভাষা প্রযুক্তি টুলসে ব্যবহার করা হবে।

খ) তথ্য ও যোগাযোগ প্রযুক্তি বিভাগ কর্তৃক তৈরীকৃত বাংলা OCR এর আরও উন্নতিসাধন এবং এর সাথে হাতের লেখা শনাক্তকরণ পদ্ধতি একীভূত করা (Further improvement of Bangla OCR developed by ICTD & integrating hand writing recognition system)
ওসিআর হলো অপটিক্যাল ক্যারেক্টার রিডার যা মুদ্রিত নথি বা বইয়ের হরফকে কম্পিউটারে সংশোধন ও সম্পাদনাযোগ্য অবস্থায় রূপান্তরিত করে। আইসিটি বিভাগ কর্তৃক তৈরীকৃত বাংলা OCR শুধুমাত্র টাইপ নথি চিহ্নিত করতে পারে এবং এটির ৮৭% সঠিক আউটপুট প্রদান করে। প্রস্তাবিত OCR টি হাতের লেখা ডকুমেন্ট শনাক্ত করতে সক্ষম হবে এবং বাংলা হরফ শনাক্তকরণ ক্ষমতা ৯৯% এ উন্নীত করা হবে। ফলে এই অ্যাপ্লিকেশনটির মাধ্যমে টাইপ না করে দ্রুত কম্পোজ ও ডিজিটাইড করা যাবে পুরানো ও বিরল বাংলা বই, হস্তলিখিত নথি, মধ্যযুগের পুথি, ব্রিটিশ সময়ের দলিল। এর ফলে প্রকাশনা, শিক্ষা-গবেষণা ও দাপ্তরিক কাজে মূল্যাবান সময় ও শ্রম বাঁচবে। এই উপাংশের ফলে পিডিএফ বা ছবি থেকে হরফ শনাক্ত করে সহজে ওয়ার্ড প্রসেসরে রূপান্তর করা যাবে।

গ) কথা থেকে লেখা এবং লেখা থেকে কথায় রূপান্তর সফটওয়্যার উন্নয়ন (Development of Bangla speech to text & text to speech software)</td></tr>
</table>

Dhaka University Institutional Repository

ক্রমিক নং	বিষয়বস্তু	বিবরণ
		স্পিচ টু টেক্সট (STT) সফটওয়্যার হলো উচ্চারিত কথামালাকে টেক্সটে রূপান্তর করা। এই অ্যাপ্লিকেশনটি সম্পন্ন হলে ভাষণ ও বক্তব্য দ্রুত লিখিত তথা কম্পোজ অবস্থায় পাওয়া যাবে। বিভিন্ন সাক্ষাৎকার, বিবৃতি দ্রুত যন্ত্রের মাধ্যমে অনুলিখন করা যাবে, যার ফলে অনেক অর্থ-সময় ও শ্রম বাঁচবে। পক্ষান্তরে টেক্সট টু স্পিচ (TTS) অ্যাপ্লিকেশন হলো ডিজিটাল টেক্সটকে উচ্চারিত শব্দে রূপান্তর করা। এই অ্যাপ্লিকেশন যাদের dyslexia বা পড়ার অসুবিধা, reading challenges বা দৃষ্টি-বৈকল্য আছে তাদের উপকারে আসবে । এর ফলে স্বয়ংক্রিয়ভাবে যন্ত্রের মাধ্যমে সরকারি জরুরি বিজ্ঞপ্তি, নির্দেশনা, পত্রিকার শিরোনাম/ তাজা খবর শোনা যাবে। ওয়েবসাইটে প্রকাশিত লেখা সহজে শোনা যাবে। ঘ) জাতীয় কিবোর্ড (বাংলা) এর উন্নয়ন (Improvement of the National Keyboard (Bangla) ২০০৪ সালে বাংলাদেশ কম্পিউটার কাউন্সিল কর্তৃক 'জাতীয় বাংলা কিবোর্ড' (BDS-1538:2004) তৈরি করা হয়েছে। অদ্যাবধি, এর ব্যবহার ও প্রয়োগ খুব সীমিত পর্যায়ে রয়েছে। এটি আরো কার্যকর করার জন্য, তার সীমাবদ্ধতাগুলো চিহ্নিত করা প্রয়োজন এবং উন্নতির জন্য প্রয়োজনীয় পদক্ষেপ গ্রহণ করা প্রয়োজন।পরিমার্জিত ও পরিবর্ধিত কিবোর্ডটি বিভিন্ন ডিভাইস উপযোগী হবে এবং দ্রুত সহজে নির্ভুল বাংলা কম্পোজের উপযোগী হতে হবে। ঙ) বাংলা ভাষাশৈলীর নীতি প্রমিতকরণ (স্টাইল গাইড উন্নয়ন) (Development of Bangla style guide) প্রযুক্তির সঙ্গে বাংলাভাষার সম্মিলনের প্রথম ও পূর্বশর্ত হলো ভাষার রীতি এবং ভাষা ব্যবহারের নীতি ঠিক করা। এর আওতায় রয়েছে ন্যাচারাল ল্যাংগুয়েজ প্রসেসিং-এ কার্যকর হবে এমনভাবে বাংলা বাক্য ও শব্দ বিশ্লেষণ, বানান প্রমিতকরণ, বাংলা উচ্চারণ প্রমিতকরণ, বাংলা ক্যারেকটার প্রমিতকরণ, বাংলা বিরাম চিহ্ন প্রয়োগ-রীতি প্রমিতকরণ, অন্যভাষার সঙ্গে বাংলা ভাষার ব্যবহারের রীতি (ইংরেজি, আরবি, সংস্কৃত, চাকমা প্রভৃতি) নির্ধারণ, টীকা ব্যবহারের রীতি (ফুটনোট ও এন্ডনোট), গ্রন্থপঞ্জি ও নির্ঘণ্ট লেখার নিয়ম নির্ধারণ প্রভৃতি। চ) বাংলা ফন্ট আন্তঃক্রিয়া/ রূপান্তর ইঞ্জিন) (Development of the Bangla font interoperability Engine) বিভিন্ন ডিজিটাল মাধ্যমে যেমন, কম্পিউটার অপারেটিং সিস্টেম, ওয়েব ও মোবাইল প্লাটফর্ম-এ বাংলা ফন্ট স্থানান্তরের সময় ভেঙে যায়। কখনো কখনো এক অ্যাপ্লিকেশন থেকে অন্য অ্যাপ্লিকেশনে (যেমন, ওয়ার্ড থেকে এক্সেল বা পাওয়ার পয়েন্ট)লেখা স্থানান্তরের সময় ফন্ট ভেঙে যায়। এ ফন্টভাঙা সমস্যা থেকে মুক্তি পাওয়ার জন্য কম্পিউটিং প্ল্যাটফর্মে ইন্টারাঅপারেবল করার জন্য একটি/(একাধিক) আদর্শ তথ্য এনকোডিং প্রতিষ্ঠা প্রয়োজন। এই এনকোডিং এর ওপর ভিত্তি করে কিছু ফন্ট এনকোডিং কনভার্টার তৈরি করতে হবে, যা ফন্ট ইন্টারঅপারেবল ইঞ্জিন হিসেবে কাজ করবে। অর্থাৎ বিভিন্ন অপারেটিং সিস্টেম, ডিভাইস, সফটওয়্যার, মুদ্রিত ও ডিজিটাল মাধ্যমে ফন্ট ভাঙবে না। ভাঙলে তা এই ইঞ্জিনের মাধ্যমে সঠিক অবস্থায় আনা যাবে। ছ) বাংলা CLDR উন্নয়ন এবং ইউনিকোড কনসোরটিয়মে জমা দেয়া (Development of Bangla CLDR resource and submit to Unicode) ইউনিকোড কমন লোকাল ডাটা রিপোজিটরি (Unicode Common Locale Data Repository বা CLDR) হলো বিশ্বের প্রধান ভাষাসমূহের সহায়ক সফটওয়্যার হিসেবে মূল বিল্ডিং ব্লক যোগানদাতা। এটি স্থানীয় ইউনিকোড বিষয়ে বৃহত্তম ও প্রমিত তথ্য ভান্ডার। আন্তর্জাতিক কোম্পানিসমূহ তাদের সফটওয়্যার আন্তর্জাতিকায়ন ও স্থানীয়করণে এই তথ্য ভান্ডার ব্যবহার করে থাকে এবং ডিএলডিআর প্রদত্ত মান অনুসরণ করেন । বাংলা ভাষার ক্ষেত্রেও এই উপাংশের মাধ্যমে সিএলডিআর ভান্ডার উন্নয়ন ও প্রমিতকরণ করে তা ইউনিকোড কনসোরটিয়মে জমা দিতে হবে এবং অনুমোদনের প্রয়োজনীয় পদক্ষেপ নিতে হবে। জ) বাংলা বানান ও ব্যাকরণ পরীক্ষক উন্নয়ন (Development of Bangla Spell & Grammar checker) স্বয়ংক্রিয় বানান পরীক্ষক ব্যবহার করে শব্দ সঠিকভাবে সম্পাদন করা সম্ভব। মোবাইল, কম্পিউটার, ওয়েবসহ অন্যান্য মাধ্যমে প্রমিত বানানের ভুল চিহ্নিত করবে এবং প্রযোজ্যক্ষেত্রে সঠিক বানানের পরামর্শ দেবে- এমন বানান পরীক্ষক উন্নয়ন করা হবে। বাংলা প্রায় সমোচ্চারিত, সমার্থক ও সমদর্শী (হোমোনিম, হোমোগ্রাফ, হোমোফোন প্রভৃতি) চিহ্নিত করতে পারবে- এমন বানান পরীক্ষক উন্নয়ন করা হবে। ব্যাকরণ পরীক্ষক ভুল বাংলা বাক্য জানাতে সাহায্য করবে। সরল ও জটিল বাক্যের প্রচলিত সাধারণ ভুলগুলো চিহ্নিত করে পরামর্শ/ সাজেশন দিতে সক্ষম ব্যাকরণ পরীক্ষক উন্নয়ন করতে হবে। বানান এবং ব্যাকরণ পরীক্ষক প্রুফরিডারের কাজ করবে, যা দ্রুত নির্ভুল রচনা নিশ্চিত করবে। ঝ) বাংলা যান্ত্রিক অনুবাদক উন্নয়ন (Development of the Bangla Machine Translator (MT)) যান্ত্রিক অনুবাদের মাধ্যমে দ্রুত বাংলা ভাষা বিভিন্ন ভাষায় অনুবাদ করা সম্ভব হবে। বর্তমানে প্রচলিত যন্ত্র-অনুবাদ পদ্ধতির চেয়ে অধিকতর কার্যকর ও সফল অনুবাদ পদ্ধতির উন্নয়ন ঘটানো হবে। এর ফলে তথ্যমূলক বাংলা, প্রাতিষ্ঠানিক রচনা/ডকুমেন্টস/ নথি, সংবাদ বিজ্ঞপ্তি, আবহাওয়া সংবাদ দ্রুত নির্ভুলভাবে অনুবাদ করা সম্ভব হবে। এই অনুবাদ-কৌশলের ফল প্রকল্পের অন্যান্য উপাংশেও ব্যবহৃত হবে। ঞ) স্ক্রিন রিডার সফটওয়্যার উন্নয়ন (Development of Screen Reader software) স্ক্রিন রিডার সফটয়্যার এর মাধ্যমে মনিটরের পর্দায় প্রদর্শিত হচ্ছে তথ্য (বাংলা লেখা বা চিহ্ন)তা শনাক্ত ও ব্যাখ্যা করা যায়। এই ব্যাখ্যা তারপর পুনরায় উপস্থাপিত হয় টেক্সট টু স্পিচ, শব্দ আইন বা ব্রেইল আউটপুট ডিভাইস ব্যবহারকারীর জন্য। স্ক্রিন রিডার একটি সহায়ক প্রযুক্তি যা দৃষ্টিশক্তিহীন, ক্ষীণদৃষ্টি, নিরক্ষর বা শেখার অক্ষম মানুষ ব্যবহার করতে পারবে।

Dhaka University Institutional Repository

ক্রমিক নং	বিষয়বস্তু	বিবরণ
		ট) বিশেষ চাহিদাসম্পন্ন / 'প্রতিবন্ধী' ব্যক্তির ভাষিক যোগাযোগের জন্য সফটয়্যার উন্নয়ন (Development of software for disable people) সাধারণ মানুষের কাছে সহজ কিন্তু বিশেষ ইন্দ্রিয় বা অঙ্গ ব্যবহারে অক্ষম মানুষের কাছে ভাষা ব্যবহারের পদ্ধতি দুঃসাধ্য বা অসম্ভব হতে পারে। এমন পরিস্থিতিতে বাংলা ভাষা উপযোগী সফটয়্যার বাংলা ভাষাভাষী মানুষের কল্যাণ বয়ে আনতে পারে, বিশেষ করে দৃষ্টি ও শ্রবণে স্বাভাবিকভাবে অক্ষম মানুষের জন্য। এজন্য ব্রেইল বোর্ডসহ কয়েকটি বাংলা সফটওয়্যার ও টুলস তৈরি করা হবে যেন মূক-বধির-অঙ্গ চালনে অক্ষম মানুষ ভাষা/ বাংলা ভাষা ব্যবহারের প্রতিবন্ধকতা দূর করতে হবে। বাংলা ভাষা-অনুকূল সাইন ল্যাংগুয়েজ বুঝতে সক্ষম ও প্রদর্শনে সক্ষম সফটয়্যার তৈরি করা হবে। ঠ) বাংলা অনুভূতি বিশ্লেষণের সফটয়্যার উন্নয়ন (Development of sentiment analysis software in Bangla) সেন্টিমেন্ট অ্যানালাইসিস টুলস ভাষিক তথ্য বা টেক্সট বিশ্লেষণের মাধ্যমে মানুষের অনুভূতি বিশ্লেষণ করে থাকে যা opinion mining নামেও পরিচিত। এর সাথে সম্পর্কযুক্ত হলো প্রাকৃতিক ভাষা প্রক্রিয়াকরণ, টেক্সট বিশ্লেষণ ও গণনীয় ভাষাতত্ত্বের ব্যবহার। এই সফটয়্যার তৈরি হলে কোনো বিবৃতির কন্টেন্ট বিশ্লেষণ করে তার সারমর্ম, মূলভাব বের করা যাবে। কোনো বক্তব্য নেতিবাচক, অস্তিবাচক বা নিরপেক্ষ কিনা তাও যন্ত্রের মাধ্যমে বের করা যাবে। এর মাধ্যমে দ্রুত বাজার-জরিপ, জনমত জরিপ করা, নির্বাচন উত্তর জনমত যাচাই যন্ত্রের মাধ্যমে দ্রুত করা যাবে। শিক্ষা ও গবেষণায় বিশেষ করে কোয়ানটেটিভ রিসার্চে এর প্রয়োগ করা যাবে। ড) একটি বহুভাষিক কন্টেন্ট রূপান্তর পদ্ধতি ও প্ল্যাটফর্ম উন্নয়ন করা (Developing a service platform combining language processing tools to build processing pipelines for value adding tasks in multilingual content processing) ভ্যালু অ্যাডিং টাস্ক হিসেবে একটি সেবা প্ল্যাটফর্ম নির্মাণ করা যেখানে ভাষা প্রক্রিয়াকরণ টুলস্ একত্রিত করে একটি বহুভাষিক কন্টেন্ট রূপান্তর পদ্ধতি ও প্ল্যাটফর্ম উন্নয়ন করা হবে এই উপাংশে। স্বয়ংক্রিয়ভাবে বহুভাষায় রিয়েলটাইম ট্রান্সক্রিপশন ও অনুবাদ করতে পারে এই অ্যাপ্লিকেশন ও প্লাটফরম। এর ফলে বক্তৃতা, সভার এবং টেলিফোন কথোপকথন থেকে অন্য ভাষার লেখা ও কথায় রূপান্তর করা সম্ভব হবে এবং এইসব রূপান্তরিত তথ্য জমা রাখার ব্যবস্থা থাকবে। ঢ) সবচেয়ে জনপ্রিয়/ব্যবহৃত সাইটগুলি আন্তর্জাতিক ভাষায় অনুবাদ (Translation of most popular/used sites into international language) বাংলাদেশে সাংস্কৃতিক ও অর্থনৈতিকভাবে গুরুত্বপূর্ণ বিষয়গুলো আন্তর্জাতিক পর্যায়ে ও ওয়েবে তুলে ধরার জন্য প্রয়োজন বাংলা-সংশ্লিষ্ট বিষয়গুলোকে ইংরেজিসহ পৃথিবীর অন্যান্য ভাষায় তুলে ধরা। জিআই ট্যাগিংয়ে গুরুত্বপূর্ণ বিষয়সমূহ যেমন জামদানি শাড়ি, মসলিন, ফজলি আম, নকশি কাঁথা, গাজীর পটচিত্র, জারিগান, সাহিত্য, চিত্রকলা, নৃত্য প্রভৃতি বিষয়গুলো এবং বিষয়-ধারক ওয়েবসাইটগুলো ইংরেজি, স্প্যানিশ, আরবি, ফরাসি, মান্দারিন, জাপানি, হিন্দি, উর্দু, বার্মিজ, নেপালি, জার্মান, ফারসি, পর্তুগিজ, কোরিয়, রাশিয়া প্রভৃতি ভাষায় ম্যানুয়ালি অনুবাদ করা হবে। এবং বিভিন্ন জনপ্রিয় সাইট/ক্ষেত্র অনুযায়ী আপলোড করতে হবে। এই অনুবাদের সাথে যুক্ত রয়েছে বাংলাদেশি পণ্যের রপ্তানির সম্ভাবনা। পণ্য রপ্তানির সঙ্গে বাংলা ভাষা ও সংস্কৃতিরও আন্তর্জাতিকায়ন হবে। এই উপাংশে উল্লিখিত বিষয়সংশ্লিষ্ট কন্টেন্টগুলো অনুবাদ করা হবে। প্রধান ও ব্যবহারে জনপ্রিয় গুরুত্বপূর্ণ ওয়েবসাইটগুলোকে দুইভাবে অনুবাদ করতে হবে। প্রথমত, সাইটগুলোর কন্টেন্ট অনুবাদ। দ্বিতীয়ত, ল্যাংগুয়েজ ফাইল (যেমন, bn_bd) তৈরি। জনপ্রিয় সাইটগুলোসহ বিভিন্ন ব্রাউজার, সিএমএস, ওস লোকালাইজেশন এই কার্যক্রমের অন্তর্ভুক্ত। ন) বাংলা ভিন্ন দেশের অন্য প্রচলিত ভাষা/ ক্ষুদ্র-নৃগোষ্ঠীর ভাষার জন্য প্রমিত কিবোর্ড (Standard Keyboard for Tribal Languages) বাংলা ছাড়াও বাংলাদেশে আরো অনেক ভাষা রয়েছে। এর মধ্যে কয়েকটি সচল ও শক্তিশালী, কয়েকটি বিপন্ন। এমন সব ভাষার বিশেষ করে, বাংলাদেশের ক্ষুদ্র-নৃগোষ্ঠী ভাষা নিজস্ব বর্ণমালা আছে। এই ভাষাগুলোকে প্রযুক্তি বান্ধব করা প্রয়োজন। এজন্য বিভিন্ন ব্রাউজার, ওয়ার্ড প্রসেসিং অ্যাপ্লিকেশন, ওয়েবে ব্যবহারযোগ্য কিবোর্ড লে-আউট ও কিবোর্ড সফটয়্যার উন্নয়ন করা প্রয়োজন। এর ফলে দেশের বিভিন্ন ভাষার মানুষ ফেসবুক, টুইটারসহ অনলাইন সামাজিক মাধ্যমে লিখতে পারবে। ত) বাংলা ভাষা সহায়ক IPA ফন্ট ও সফটয়্যার উন্নয়ন (Incorporating Bengali IPA fonts and software to world language linguistic List) স্বয়ংক্রিয় ট্রান্সস্ক্রিপশন পদ্ধতি তৈরি করতে হলে আইপিএ নিয়ে গবেষণা হবে মূল ভিত্তি । বাংলা ভাষার জন্য সহায়ক IPA ফন্ট ও সফটয়্যার উন্নয়ন হলে এই প্রকল্পের অন্যান্য উপাংশ তৈরি সহজ হবে। আন্তর্জাতিক ধ্বনিমূলক বর্ণমালা/ International Phonetic Alphabet (IPA) মানুষের দ্বারা উচ্চারিত প্রায় সব ধ্বনির লিখিত রূপকে প্রকাশ যা হিসাবে আন্তর্জাতিক ফোনেটিক এসোসিয়েশন দ্বারা স্বীকৃত ও নিয়ন্ত্রিত হয়। বাংলা ভাষাকে আইপিএতে প্রকাশ করা প্রয়োজন অন্যান্য টুলসকে প্রয়োজনীয় সমর্থন জোগানোর জন্য । সাধারণত IPA অভিধান রচয়িতা, বিদেশি ভাষার ছাত্র-শিক্ষক, ভাষাবিদ, স্পিচ-ল্যাংগুয়েজ প্যাথলজিস্ট, গায়ক, অনুবাদকদের দ্বারা ব্যবহৃত হয়। বাংলা ইন্টারন্যাশনাল ফোনেটিক বর্ণমালা হবে বাংলা বর্ণমালার উপর ভিত্তি করে ফোনেটিক ট্রান্সক্রিপশন সিস্টেম, যা IPA উপযোগী বাংলা স্ক্রিপ্ট। বলা যায়, এই উপাংশগুলোর কাজ সম্পন্ন হলে দেশ ও জাতি এর সুফল পাবে। প্রযুক্তিতে বাংলা ভাষা আর কোনো প্রতিবন্ধক

Dhaka University Institutional Repository

ক্রমিক নং	বিষয়বস্তু	বিবরণ
		হবে না, বরং সহায়ক হবে। আন্তর্জাতিক পর্যায়ে বাংলা ভাষার প্রয়োগ ও বিস্তৃতির বাস্তবিক ভিত্তি তৈরি হবে। যা হবে এক সত্যিকারের বিপ্লব।
৮	বাস্তবায়ন অগ্রগতি	▪ ৪টি কম্পোনেন্টের সফটওয়্যার ডেভেলপমেন্ট কার্যক্রম চলমান রয়েছে। ▪ বাকি কম্পোনেন্টসমূহের ক্রয় কার্যক্রম চলমান রয়েছে।
৯	সেমিনার/কর্মশালা/ আয়োজিত ইভেন্ট ও প্রতিযোগিতা	▪ মাননীয় প্রধানমন্ত্রী কর্তৃক ০৫ আগস্ট ২০১৮ খ্রি: এবং ০১ নভেম্বর ২০১৮ খ্রি: তারিখে ভিডিও কনফারেন্সিং এর মাধ্যমে ৪৫ জেলার ১৪০০টি ইউনিয়নের কানেক্টিভিটি উদ্বোধন করা হয়েছে। ▪ মাননীয় প্রধানমন্ত্রীর তথ্য ও যোগাযোগ প্রযুক্তি বিষয়ক মাননীয় উপদেষ্টা কর্তৃক ১১ এপ্রিল ২০১৮ খ্রি: তারিখে ০৬ জেলার ১৬০টি ইউনিয়নের কানেক্টিভিটি উদ্বোধন করা হয়েছে। ▪ মাননীয় প্রতিমন্ত্রী, তথ্য ও যোগাযোগ প্রযুক্তি বিভাগ ০৫ আগস্ট ২০১৭ খ্রি: তারিখে সিলেটে HDPE Duct স্থাপন কার্যক্রম পরিদর্শন, ০৮ ফেব্রুয়ারি ২০১৯ খ্রি: তারিখে কক্সবাজারে, ১৪ মার্চ ২০১৯ খ্রি: তারিখে রাজশাহীতে, ০৩ এপ্রিল ২০১৯ খ্রি: তারিখে খুলনায় এবং ২০ জুন ২০১৯ খ্রি: তারিখে সিলেটে অনুষ্ঠিত মতবিনিময় সভায় সভাপতিত্ব করেন।

প্রকল্পের চলমান কিছু কার্যক্রমের স্থিরচিত্র

১. গত ২০ জুন, ২০১৭ তারিখে বাংলাদেশ কমিউটার কাউন্সিলের মিলনায়তনে ‘তথ্য প্রযুক্তিতে বাংলা ভাষার ব্যবহারঃ চ্যালেঞ্জ ও করণীয়’ শিরোনামে প্রকল্পের উদ্যোগে অনুষ্ঠিত সেমিনার।

২. গত ১৭ সেপ্টেম্বর, ২০১৭ তারিখে রাজশাহী প্রকৌশল ও প্রযুক্তি বিশ্ববিদ্যালয়ে প্রকল্পের পক্ষ থেকে একটি সেমিনার অনুষ্ঠিত হয়।

Dhaka University Institutional Repository

৩. ০১ নভেম্বর, ২০১৭ তারিখে প্রকল্পের পক্ষ থেকে খুলনা বিশ্ববিদ্যালয় একটি সেমিনার অনুষ্ঠিত হয়।

৪. খুলনা প্রকৌশল ও প্রযুক্তি বিশ্ববিদ্যালয়ে ০২ নভেম্বর, ২০১৭ তারিখে প্রকল্পের পক্ষ থেকে একটি সেমিনার অনুষ্ঠিত হয়।

৫. ২০ সেপ্টেম্বর, ২০১৭ তারিখে প্রকল্পের পক্ষ থেকে অনুষ্ঠিত হয় একটি সেমিনার। যেখানে উপস্থিত ছিলেন বিশ্ববিদ্যালয়ে শিক্ষক ও শিক্ষার্থীসহ আরও অনেকে।

Dhaka University Institutional Repository

৬. বাংলা মেশিন ট্রান্সলেটর নিয়ে ডিজিটাল ওয়ার্ল্ড, ২০১৭ তে অনুষ্ঠিত হয় একটি সেমিনার, যেখানে মূল আলোচক হিসাবে উপস্থিত ছিলেন ড. নীলাদ্রী শেখর, যাদবপুর বিশ্ববিদ্যালয়, ভারত।

৭. 'প্রতিবন্ধী ব্যক্তিদের জন্য সফটওয়ারের এর রুপরেখা প্রণয়নের জন্য ০৮ নভেম্বর, ২০১৮ তারিখে জনতা টাওয়ার সফটওয়ার টেকনোলজি পার্কের সভাকক্ষে অনুষ্ঠিত হয় দিনব্যাপী কর্মশালা। যেখানে উপস্থিত ছিলেন ডাক, টেলিযোগাযোগ ও তথ্য প্রযুক্তি মন্ত্রণালয়ের মাননীয় মন্ত্রী জনাব মোস্তাফা জব্বার, তথ্য ও যোগাযো প্রযুক্তি বিভাগের সচিব জনাব জুয়েনা আজিজ।

৮. এসডি ১৯: ইন্টিগ্রেটেড সার্ভিস প্লাটফর্ম নিয়ে ১৫ নভেম্বর, ২০১৮ তারিখে জনতা টাওয়ার সফটওয়ার টেকনোলজি পার্কের সভাকক্ষে অনুষ্ঠিত হয় দিনব্যাপী কর্মশালা।