LANGUAGE IN INDIA

Strength for Today and Bright Hope for Tomorrow

Volume 8 : 5 May 2008

ISSN 1930-2940

A Proposal for Standardization of English to Bangla Transliteration and Bangla Editor

Joy Mustafi, M.C.A. and B. B. Chaudhuri, Ph.D.

A Proposal for Standardization of English to Bangla Transliteration and Bangla Universal Editor

Joy Mustafi, M.C.A. & B. B. Chaudhuri, Ph.D.

1. Introduction

Indian language technology is being more and more a challenging field in linguistics and computer science. Bangla (also written as Bengali) is one of the most popular languages worldwide [Chinese Mandarin 13.69%, Spanish 5.05%, English 4.84%, Hindi 2.82%, Portuguese 2.77%, Bengali 2.68%, Russian 2.27%, Japanese 1.99%, German 1.49%, Chinese Wu 1.21%]. Bangla is a member of the New Indo-Aryan language family, and is spoken by a vast population within the Indian subcontinent and abroad. Bangla provides a lot of scope for research on computational aspects.

Efficient processors for Bangla, which exhaustively deal with all the general and particular phenomena in the language, are yet to be developed. Needless to say transliteration system is one of them. To represent letters or words in the corresponding characters of another alphabet is called *transliteration*.

English to Bangla transliteration has no standard till now. Some early systems like Lekho [1], Pata [2], Bangla Pad [3] are not very user-friendly having complex rules for character mapping. Some keyboard layouts like Ekushey [4], Avro [5], and Bijoy [6] have Unicode [7] or ASCII [8] or ISCII [9] mappings, which are again very hard to use, particularly when these systems deal with compound clusters of consonant characters. The main problem is that there is no particular rule for English to Bangla transliteration.

The system described here proposes a standard, a definite rule, and application program for writing, editing, storing, reusing and viewing Bangla text in a digital media. The English text can be stored as simple plain text file in any platform and may be used for other research activities like machine translation, information retrieval, spell checker, optical character recognition, speech technology and other Bangla language technologies [10].

This system is designed for English to Bangla character conversion and representation of Bangla in Unicode. The mapping of characters follows the morphological structure and the spelling rule of Bangla. Though some systems were developed earlier for phonological representation, but, for visual editor or storage of Bangla corpus, the spelling is the more important than the pronunciation.

A universal editor for Bangla is also proposed here which follows the standard and represents Bangla in Unicode [11] with suitable Bangla open type font. The text in English script is used for the input, which can be browsed from any location, and the

Bangla Universal Editor converts the text into Bangla and displays it in the specified window.

The Bangla Unicode output can be used for the development of Bangla software like operating system, compiler, word-processor, dictionary, web-page [12] and other software. It is useful in writing emails, messages, blogs in Bangla. The standards, methodologies and applications are described here.

1.1 Objective

The main objective of the work is to introduce a standard for English to Bangla transliteration system. Advanced research on Bangla language technology [13][15][16][17] by us is already established. Bangla corpus is used for developing many language technology systems.

Some early research on Bangla also proposed some transliteration rules or character mapping [14]. As there is no standard for Bangla transliteration, Bangla corpus cannot be stored in a specific format. As a result, the researchers get different representation of Bangla text from different sources. If one can write Bangla text in English script, and can store data in plain text, it may not require any other specific software for Bangla. A simple ASCII editor (like Notepad, gedit, nedit) will work. It will become platform independent also.

To view and edit the English script written to represent Bangla, a Universal editor is introduced here, which can be used for correction or modification of the Bangla text written in English script. However, in this editor, one can see the Bangla text in English and as well as in Bangla font simultaneously in separate frames of the same application program.

1.2 Justification of the Proposal

The proposal for the standard is necessary as there is no standard available for Bangla transliteration. Some important points discussed here are:

i. Plain Text Storage Mode for English (platform, encoding, font independent)
ii. One-to-One Character Mapping (using lower [a~z] and upper [A~Z] cases (26 X 2) 52 characters of English → 50 Basic characters of Bangla (Vowels, Consonants, Special Characters); 1 ‘hasanwa’ [Q]; and 1 unused [L])
iii. Phonetic Character Chart (English characters are chosen very close to the phonetics of Bangla characters, but the word construction rule obeys the spelling of correct Bangla words. In most cases the phonetic character chart is maintained, but there are a few exceptions)
iv. Simple Representation of Character Clusters (easy to parse the input text).
v. Morphological Word Construction Rule (using spelling not pronunciation, to overcome ambiguities)

English Script (Existing Methods)	English Script (Proposed Standard)	Bangla Script
kh d'h	K D	খ ঢ
H/m	hm	হ্ম
ka'n/d'a	kAnda	কান্ড

Table 1. Comparison Chart of Existing Methods and Proposed Standard

Some existing transliteration rules have many-to-one character mapping, which makes the converter complex when the system compiles the input text in English script. Also some existing rules include other special symbols like single quote (’), dot (.) or slash (/) along with [a~z] and [A~Z]. It is very difficult and sometimes impossible to parse such multiple characters, which are also mixed with special symbols. It does not give guarantee of the uniqueness of the character representation.

2. English to Bangla Transliteration Standard

The proposed standard describes the character mapping, representation of different Bangla characters (individuals and clusters), construction of Bangla words, and some rules to handle all possible complex salutations of writing in Bangla.

2.1 Character Mapping

a অ	A আ	i ই	I ঈ	u উ	U ঊ	q ঋ
e এ	E ঐ	o ও	O ঔ			
k ক	K খ	g গ	G ঘ	f ঙ		
c চ	C ছ	j জ	J ঝ	F ঞ		
t ট	T ঠ	d ড	D ঢ	N ণ		
w ত	W থ	x দ	X ধ	n ন		
p প	P ফ	b ব	B ভ	m ম		
Y য	y য়	r র	l ল			
S শ	R ষ	s স	h হ			
v ড়	V ঢ়		Z ৎ			
M ং	H ঃ	z ঁ	Q ্			

Table 2. Proposed Standard for English to Bangla Transliteration

The main idea of the proposal is one-to-one character representation. Thus it is very easy to remember and to compile the input text. Though the system is designed for Bangla text, but the English characters are so chosen for mapping close to the phonetics of Bangla characters.

English q, Q, f, F, w, W, x, X, R, v, V, z, Z to corresponding Bangla mappings are not close to phonetics. The characters w, W, x and X are mapped to Bangla as WX-Notation used in earlier transliteration system for Hindi [18]. Still these are easy to memorize as exceptions. These characters are not related to Bangla phonetics also, rather absent in pronunciation. It can be also noticed that the shape of ‘Q’ is very similar to a ‘round’ and attached ‘hasanwa’.

2.2 Morphological Word Construction Rules

The word construction rule is designed so that the correct Bangla spelling is always there. Though the character chart is designed close to phonetics, but for word construction the morphological approach is adopted to solve the ambiguity problem. Thus Bangla corpus or huge lexicon database can be built with the sweetness of the traditional Bangla spelling. The rules are as follows:

i. Vowels

Vowels can be represented independently. One English character is sufficient to write a vowel in Bangla.

Examples: a, A, i, I, u, U, q, e, E, o, O

অ, আ, ই, ঈ, উ, ঊ, ঋ, এ, ঐ, ও, ঔ

ii. Consonant

Consonant can not appear without a vowel, so at least one vowel is concatenated (like: 'k+a') to the corresponding consonant.

Examples: ka, kA, ki, kI, ku, kU, kq, ke, kE, ko, kO

ক, কা, কি, কী, কু, কূ, কৃ, কে, কৈ, কো, কৌ

By default 'a' is attached with the corresponding consonant, if there is no specific vowel.

iii. Consonant Cluster (Compound Character)

When consonants are added to other consonants to form a cluster, they are simply concatenated one after another according to the order of the compound character.

Examples: kka, kta, kwa, kYa, kra, kla, kRa, ksa, kza (with consonant 'ka')

ক্ক, ক্ট, ক্ত, ক্য, ক্র, ক্ল, ক্ষ, ক্স, ক্ঁ

Following these rules complete words can be easily constructed. No matter whether there are difficult glyphs in Bangla script, one can represent them in the English characters using the standard.

Examples: afka, Sulka, arka, bAkYa, naksA, lakRa, lakRmI

অঙ্ক, শুল্ক, অর্ক, বাক্য, নক্সা, লক্ষ, লক্ষ্মী

An illustrated example (difficult Bangla words having lots of compound characters) of transliteration is shown as bellow. This is a poem by Nobel Laureate Rabindranath Tagore, which contains a large number of consonant compounds:

namo Yanwra, namo Yanwra,	নমো যন্ত্র, নমো যন্ত্র, নমো যন্ত্র, নমো যন্ত্র।
namo Yanwra, namo Yanwra।	
wumi cakramuKaramanxriwa,	তুমি চক্রমুখরমন্দ্রিত,
wumi bajrabahnibanxiwa,	তুমি বজ্রবহ্নিবন্দিত,
waba baswubiSbabakRoxaMSa	তব বস্তুবিশ্ববক্ষোদংশ
XbaMsa-bikata xanwa।	ধ্বংস-বিকট দন্ত।
waba xIpwa agni Sawa SawaGnI	তব দীপ্ত অগ্নি শত শতঘ্নী
biGnabijaya panWa।	বিঘ্নবিজয় পন্থ।
waba lOhagalana SElaxalana	তব লৌহগলন শৈলদলন
acala-calana manwra।	অচল-চলন মন্ত্র।
kaBu kARTaloRtraiRtakaxqVa	কভু কাষ্ঠলোষ্ট্রইষ্টকদৃঢ়
GanapinaxXa kAyA,	ঘনপিনদ্ধ কায়া,
kaBu BUwala-jala-anwarIkRa-	কভু ভূতল-জল-অন্তরীক্ষ-
lafGana laGumAyA,	লঙ্ঘন লঘুমায়া,
waba Kani-Kaniwra-naKa-bixIrNa	তব খনি-খনিত্র-নখ-বিদীর্ণ
kRiwi bikIrNa-anwra,	ক্ষিতি বিকীর্ণ-অন্ত্র,
waba paFcaBUwa-banXanakara	তব পঞ্চভূত-বন্ধনকর
inxrajAla wanwra।	ইন্দ্রজাল তন্ত্র।
rabInxranAWa TAkura	রবীন্দ্রনাথ ঠাকুর

Table 3. English to Bangla Transliteration Illustration

Thus the stored text contains correct Bangla spelling always. This is irrespective of regional or local impact on colloquial speech, which leads to ambiguity. Most of the characters are mapped phonetically. When same spelling has different pronunciation, the system emphasizes on the spelling. For an example, the words ‘kamala’ has two different meanings (‘Decreased’ and ‘Lotus’) and pronunciations (‘komlo’ and ‘kamol’ - not following the standard, as written here phonetically):

কমল - কোমলো

কমল - কমোল

Here it is written as ‘kamala’ only to keep the correct spelling. Again ‘komala’ (‘Soft’) has different spelling as well as different pronunciation. That’s why the morphological word construction rule is followed. This method disambiguates the problem.

3. Bangla Universal Editor

The Bangla Universal Editor is designed for viewing, editing, and storing of Bangla text. It can store Bangla corpus in English script irrespective of the Unicode or system

dependencies. The system is developed in object-orientated programming language Java in UNIX operating system. The desktop application program is platform independent and can be run in any other popular operating system (like Windows) which supports Unicode. The program requires at least one Unicode Bangla font (Open type) already installed in the operating system to display the Bangla scripts in the specified window. The Algorithm is simple but very unique in nature. Some important methods along with the basic algorithm are described here.

3.1 Bangla Unicode Standard

Bangla Unicode Standard is available in the range of **U0980 ~ U09FF**. The main categories are consonants, vowels and special symbols.

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
U+098x		ঁ	ং	ঃ		অ	আ	ই	ঈ	উ	ঊ	ঋ	ঌ			এ
U+099x	ঐ			ও	ঔ	ক	খ	গ	ঘ	ঙ	চ	ছ	জ	ঝ	ঞ	ট
U+09Ax	ঠ	ড	ঢ	ণ	ত	থ	দ	ধ	ন		প	ফ	ব	ভ	ম	য
U+09Bx	র		ল				শ	ষ	স	হ			়	ঽ	া	ি
U+09Cx	ী	ু	ূ	ৃ	ৄ			ে	ৈ			ো	ৌ	্	ৎ	
U+09Dx								ৗ					ড়	ঢ়		য়
U+09Ex	ৠ	ৡ	ৢ	ৣ			০	১	২	৩	৪	৫	৬	৭	৮	৯
U+09Fx	ৰ	ৱ	৲	৳	৴	৵	৶	৷	৸	৹	৺					

Table 4. Bangla Unicode 5.1 Standard

Bangla characters are very complex in glyph view, the vowel can appear independently and with consonant, and the glyphs are different. In second case it is called the ‘mAwrA’. However the consonant must have a vowel associated with it. It is attached with the specific vowel to form a different glyph. Usually ‘mAwrA’ is added in the top, bottom, left and right (or both side). By-default ‘a’ is attached, for which there is no change in the glyph of the consonant. Consonants can also form clusters. Some of the clusters are simple in view (vertical concatenation of the corresponding consonants), and some have totally different look, which has no connection with the original glyphs. Special symbols like ‘canxra-binxu’, ‘anusbara’, ‘bisarga’, ‘hasanwa’, and numerals have their own Unicode and corresponding glyph.

3.2 Algorithm for Unicode Conversion

```
1.      Initialize CONSONANT-FLAG=false
2.      Repeat Steps [3-4] till the entire Input String is processed.
```

```
3.      Extract a Character from Input String.
4.      If the Character is in the range [65-91] or [97-122] or [48-57] (ASCII)
        a) Check the Character-Type
                Case 'NUMERAL":
                        Write the corresponding Character.
                Case 'CONSONANT':
                        i) Check CONSONANT-FLAG
                                Case 'true':
                                        Write 'hasanwa' and then the Character.
                                Case 'false':
                                        Write the Character and set the CONSONANT-FLAG=true
                        ii) If VOWEL_MODIFIER-FLAG=true then VOWEL_MODIFIER-FLAG=false
                Case 'VOWEL':
                        i) check CONSONANT-FLAG
                                Case 'true':
                                        i) Check the Input Character
                                        Case 'a':
                                                If VOWEL_MODIFIER-FLAG=true then
                                                        VOWEL_MODIFIER-FLAG=false
                                        Case 'A/i/I/u/U/q/e/E/o/O':
                                                Write the Modifier corresponding to Vowel.
                                                Set VOWEL_MODIFIER-FLAG=true
                                        ii) Set CONSONANT-FLAG=false
                                Case 'false':
                                        Write the Vowel corresponding to Input Character.
                Case 'SPECIAL': (For canxra-binxu/anusbara/bisarga)
                        i) Check CONSONANT-FLAG
                                Case 'true':
                                        Print Error Message 'canxra-binxu/anusbara/bisarga'
                                        cannot be preceeded by Consonant+hasanwa combination
                                Case 'false':
                                        Write the corresponding Character.
                        ii) If VOWEL_MODIFIER-FLAG=true then VOWEL_MODIFIER-FLAG=false
                Case 'Default':
                                Case '@':
                                        i) Extract characters till the next white space or
                                        new line and write them as it is
                                        ii) Set VOWEL_MODIFIER-FLAG=false
        b) Write the Character (English).
5.      Exit
```

3.3 Design and Implementation

This Bangla Universal Editor is at present a desktop application. The GUI design is very user-friendly. It has two text areas, the first one (left-side) is for English input text. This text appears after clicking the ‘Browse’ button, and browsing the input text file from the storage device. By clicking the ‘Convert’ button, the Bangla output text appears in the second text area (right-side). The ‘Reset’ button clears both text areas. If there is some spelling error in typing, one can edit and save the input file in any text editor (like Notepad), browse it and convert again to get the best result. The ‘Help’ button shows the Transliteration Standard. ‘About Us’ is as usual the developers’ details.

The web-based application for the similar system is being developed and will be available shortly. Using this tool, one can convert the Bangla text in ASCII English script to Unicode Bangla script online. It will also work to publish web pages, write e-mails, and create blogs in Bangla. However, the current version of this desktop application can accomplish all these by simply copying the content of the right-side text area and pasting it into the target application.

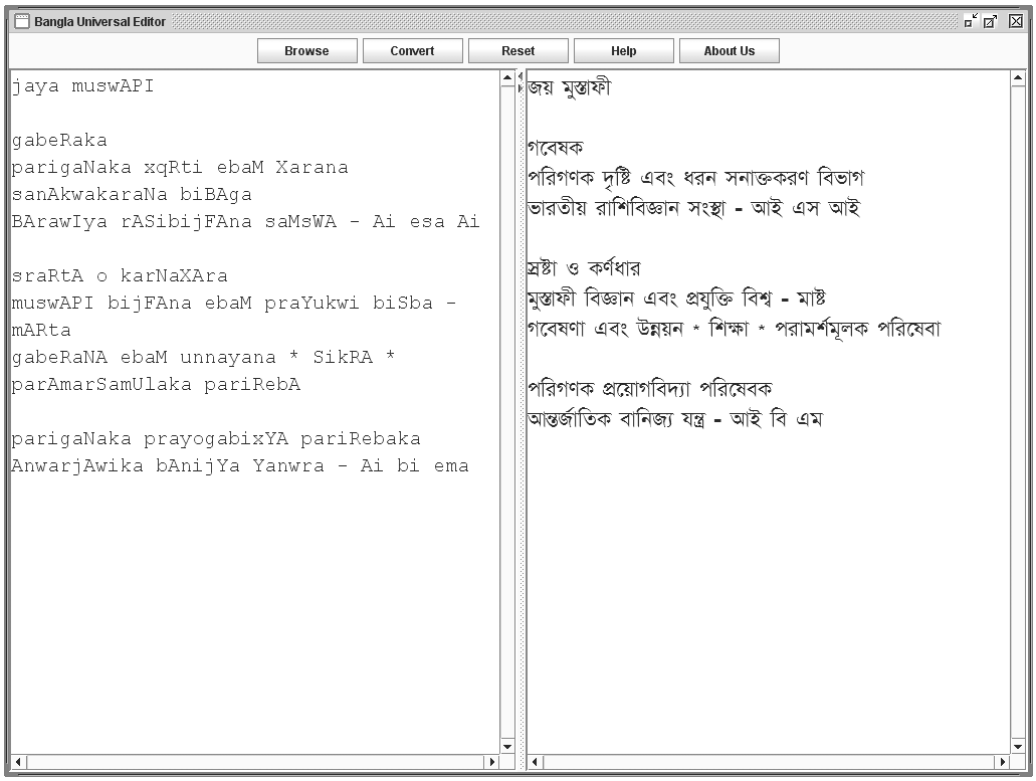

Figure1. Screenshot of the Desktop Application Program

4. Scope and Limitations

This system has a potential to be a part of Bangla language technology. Future work is open for all kind of edit options, spell check, and format support of other rich texts. Multi-lingual editor can be developed (like Bangla and Hindi) which can represent and store a document having more than one language or script in the same piece of text. This is helpful for other research activities like machine translation, information retrieval, spell checker, optical character recognition, speech technology etc.

The Bangla Universal Editor has been tested with sufficient quantity of text. The system is almost free from technical errors. Some morphological limitations are there for Unicode version 5.1. Though, they can not occur if correct Bangla spelling rule is followed.

1. aYA → this vowel is unusual in Bangla but mostly used in English as ‘A’ in the words like ‘Apple’, ‘America’ etc. But this vowel is used in Bangla for pronunciation only not for the spelling [19]. It is usually written ‘A’ or ‘e’ instead of ‘a+Y+A’. However consonants with ‘+Y+A’ is acceptable like ‘bYA’ in ‘bYAkaraNa’. And a+Y+A is used in other way like ‘aYAciwa’, and thus disambiguated.

2. rYa → this cluster leads to ambiguity, whether it is ‘refa’ over the ‘Ya’ or ‘Ya-PalA’ with ‘ra’. English words have this pronunciation in ‘Rat’, ‘Wrapper’ etc. However

the first case is considered as second case is absent in pure conventional Bangla spelling [20]. For the second case ‘rA’ is used, ‘Y’ is removed from the cluster.

Some correct Bangla spelling for the above examples:

অযাচিত, আমেরিকা, খেলা, ব্যাকরণ, আচার্য

However, the above two problems can be solved by introducing a new Bangla Unicode for ‘Y-modifier’ [Ya-PalA]. As ‘Y-modifier’ in Bangla is used as a semi-vowel modifier (Just like how the ‘y’ and ‘w’ play the role in English). This semi-vowel in Bangla is usually represented as in first case discussed here. Introduction of this new Unicode also solves the second case by concatenating ‘r’ and ‘Y-modifier’. The ‘Y-modifier’, a +‘Y-modifier’ + A, and r + ‘Y-modifier’ + A are shown as follows:

◌্য অ্যা র‍্যা

In the Bangla Universal Editor, along with the Bangla alphabet, numerals and some special symbols can also be represented by the English numerals 0, 1, 2, 3, 4, 5, 6, 7, 8, 9 and symbols !, “, #, $, %, &, ‘, (,), *, +, ,, -, ., /, :, ;, <, =, >, ?, |, \ like:

০ ১ ২ ৩ ৪ ৫ ৬ ৭ ৮ ৯

!	"	#	$	%
&	’	(	)	*
+	,	-	.	/
:	;	<	=	>
?				
।	॥			

Overall the standard and the system can be used in different field of Bangla language technology like natural language processing, machine translation, digital document processing, optical character recognition, electronic dictionary, spell checker, word processor, information retrieval, speech technology, font design, digital desktop publishing, web application and lots more, whenever the Bangla text in English script is required as the input or output to the corresponding system. Also it encourages research and development of similar system for other Indian languages.

References

[1] **Lekho;** http://lekho.sourceforge.net/
[2] **Pata;** http://www.naushadzaman.com/pata.html
[3] **Bengali Pad;** http://www.bengalipad.com/
[4] **Ekushey;** http://ekushey.org/?page/bangla_unicode_layout
[5] **Avro Keyboard;** http://omicronlab.com/avro-keyboard.html
[6] **Bijoy;** http://www.angelfire.com/tx/rezaul/bijoy.htm
[7] **Unicode;** http://www.unicode.org/charts/
[8] **ASCII;** http://en.wikipedia.org/wiki/ASCII
[9] **ISCII;** http://tdil.mit.gov.in/standards.htm
[10] **MUST Bangla Language Technology;** http://must.bangla.googlepages.com/
[11] **Bangla Unicode;** http://unicode.org/charts/PDF/U0980.pdf
[12] **Bangla Wikipedia;** http://bn.wikipedia.org/
[13] S. Goyal; **Example Based Parsing for Resource Deficient Languages.** *Ph.D. Dissertation,* Indian Institute of Technology - Delhi, 2007. pp. 4
[14] P. Sengupta; **On Lexical and Syntactic Processing of Bangla Language by Computer.** *Ph.D. Dissertation,* Indian Statistical Institute - Kolkata, 1993. pp. 166-168.
[15] J. Mustafi; **Ideas to Develop a Morphological Parser for Bangla.** *Proceedings of Symposium on Indian Morphology, Phonology and Language Engineering,* Indian Institute of Technology - Kharagpur, 2004. pp. 110-111.
[16] J. Mustafi, B. B. Chaudhuri; **Morphological Generation of Bangla Verbs for Machine Translation.** *Proceedings of Symposium on Indian Morphology, Phonology and Language Engineering,* Indian Institute of Technology - Kharagpur, 2005. pp. 90-93.
[17] J. Mustafi, S. Kar, M. Basu, B. B. Chaudhuri; **Categorization of Bangla Root Verbs for Morphological Sentence Generation.** *Proceedings of Workshop on Morphology,* Indian Institute of Technology - Bombay, 2005.
[18] **WX-Notaion;** http://mirrors.ibiblio.org/pub/mirrors/mozdev.org/indicime/wx_keyboard.html
[19] S. K. Chatterjee; **Bhasha Prakash Bangala Byakaran.** (Bangla) *Rupa & Co. Calcutta*. 1988.
[20] S. Biswas; **Samsad Bengali-English Dictionary.** 3rd ed. Calcutta, Sahitya Samsad, 2000. http://dsal.uchicago.edu/dictionaries/biswas-bengali/

Joy Mustafi, M.C.A.
Technology Integration and Management
IBM India Pvt. Ltd
Kolkata - 700156. India
jmustafi@in.ibm.com

B. B. Chaudhuri, Ph.D.
Computer Vision and Pattern Recognition Unit
Indian Statistical Institute
Kolkata - 700108. India
bbc@isical.ac.in