Language Specific Peculiarities Document for

Bengali as Spoken in India

1. Special handling of dialects

The dialects or regional variants of Bengali can be categorized according to their geographical location within West Bengal, India. Three dialect regions have been identified[1]:

Region	Districts in West Bengal
Radha	South (Purulia, Bankura, Midnapur, Hooghly, Howrah, Kolkata, 24-Parganas North, 24-Parganas South)
Varendra	North Central (Malda, Murshidabad, Birbhum, Nadia, Burdwan)
Kamrupa	North Bengal (Darjeeling, Jalpaiguri, Coochbehar, North Dinajpur, South Dinajpur)

Standard Colloquial Bengali is the common standard dialect which is understood by most speakers. This is the national standard and was based on the dialect spoken around Kolkata[2].

2. Deviation from native-speaker principle

No special deviation – only native speakers of Bengali, born in India will be collected in this project. That is no Bangladeshi speakers will be represented.

3. Special handling of spelling

There will be no particular special handling of spelling in this collection. Business names and loan words from other languages will be spelled in Bengali script and standardized wherever possible.
The *SaralBangalaAvidhan* (Dictionary of the Bengali Language) will be used as a reference[3].

4. Description of character set used for orthographic transcription

The Bengali script will be used for the orthographic transcription of Bengali.

The Unicode range for Bengali is U+0980-U+09FF. Presentation forms of Bengali glyphs depend on the display font used, however, this does not affect the underlying Unicode. The Lohit Bengali font will be used. This Unicode-based font correctly renders all Bengali characters and can be downloaded from the following website: https://fedorahosted.org/lohit/.

[1]Bhattacharya, T. (2001). Bangla. In J Garry & C. Rubino (Eds.), *Facts about the world's languages: An encyclopedia of the world's major languages, past and present* (pp. 65-71). New York: New England Publishing Associates.
[2]Bhattacharja, S. (2007).*Word formation in Bengali: A whole word morphological description and its theoretical implications.* Munich: Lincom Europa
[3]Mitra, S. C. (2005). *SaralBangalaAvidhan (Dictionary of the Bengali Language)*. New Bengal Press Pty Ltd.

Combined presentation glyphs are represented through the use of the *hoshonto* character (U+09CD), which suppresses the inherent vowel and dictates that the characters should render together.

In some cases, the characters should render separately, including the *hoshonto* symbol, but this is generally a function of the font used to view the text. In some cases where separate rendering must be forced, such as for morphological boundaries or loan words, zero-width characters (U+200c and U+200d) may be used.

5. Description of Romanization scheme

The following is Appen Butler Hill's Romanization scheme which is fully reversible. Appen Butler Hill's Romanization schemes are being used for all Indian languages for the Indus project. These schemes are designed to be as similar in form as possible, but cannot be identical due to the different writing systems and spelling conventions in each language. Some of the symbols may seem arbitrarily assigned; however they have been chosen deliberately to fit as much as possible with other Indian languages. Transcription work is done by Bengali speakers working with the Bengali script and no Romanization; the Romanization scheme is primarily used as a reference for those unfamiliar with the Bengali script, or to find and remove duplicated symbols such as *hoshonto* (0x9cd) or *chondrobindu* (0x981), which may not be visible when viewing the Bengali script alone in a text editor.

5.1 Bengali Romanization Scheme

UNICODE	BENGALI	ROMAN	DESCRIPTION
0x981	ঁ	M	BENGALI SIGN CANDRABINDU (*chondrobindu)*
0x982	ং	W	BENGALI SIGN ANUSVARA (*onushshor)*
0x983	ঃ	9	BENGALI SIGN VISARGA (*bishorga)*
0x985	অ	a	BENGALI LETTER A
0x986	আ	A	BENGALI LETTER AA
0x987	ই	I	BENGALI LETTER I
0x988	ঈ	i	BENGALI LETTER II
0x989	উ	U	BENGALI LETTER U
0x98a	ঊ	u	BENGALI LETTER UU
0x98b	ঋ	r[	BENGALI LETTER VOCALIC R.
0x98f	এ	e	BENGALI LETTER E
0x990	ঐ	e3	BENGALI LETTER AI
0x993	ও	o	BENGALI LETTER O
0x994	ঔ	o3	BENGALI LETTER AU

UNICODE	BENGALI	ROMAN	DESCRIPTION
0x995	ক	k	BENGALI LETTER KA
0x996	খ	K	BENGALI LETTER KHA
0x997	গ	g	BENGALI LETTER GA
0x998	ঘ	G	BENGALI LETTER GHA
0x999	ঙ	N	BENGALI LETTER NGA
0x99a	চ	c	BENGALI LETTER CA
0x99b	ছ	C	BENGALI LETTER CHA
0x99c	জ	j	BENGALI LETTER JA
0x99d	ঝ	Z	BENGALI LETTER JHA
0x99e	ঞ	J	BENGALI LETTER NYA
0x99f	ট	t`	BENGALI LETTER TTA
0x9a0	ঠ	T`	BENGALI LETTER TTHA
0x9a1	ড	d`	BENGALI LETTER DDA
0x9a2	ঢ	D`	BENGALI LETTER DDHA
0x9a3	ণ	n`	BENGALI LETTER NNA
0x9a4	ত	t	BENGALI LETTER TA
0x9a5	থ	T	BENGALI LETTER THA
0x9a6	দ	d	BENGALI LETTER DA
0x9a7	ধ	D	BENGALI LETTER DHA
0x9a8	ন	n	BENGALI LETTER NA
0x9aa	প	p	BENGALI LETTER PA
0x9ab	ফ	P	BENGALI LETTER PHA
0x9ac	ব	b	BENGALI LETTER BA
0x9ad	ভ	B	BENGALI LETTER BHA

UNICODE	BENGALI	ROMAN	DESCRIPTION
0x9ae	ম	m	BENGALI LETTER MA
0x9af	য	Y	BENGALI LETTER YA
0x9b0	র	r	BENGALI LETTER RA
0x9b2	ল	l	BENGALI LETTER LA
0x9b6	শ	S	BENGALI LETTER SHA
0x9b7	ষ	s`	BENGALI LETTER SSA
0x9b8	স	s	BENGALI LETTER SA
0x9b9	হ	h	BENGALI LETTER HA
0x9be	া	A2	BENGALI VOWEL SIGN AA
0x9bf	ি	I2	BENGALI VOWEL SIGN I
0x9c0	ী	i2	BENGALI VOWEL SIGN II
0x9c1	ু	U2	BENGALI VOWEL SIGN U
0x9c2	ূ	u2	BENGALI VOWEL SIGN UU
0x9c3	ৃ	r2	BENGALI VOWEL SIGN VOCALIC R.
0x9c7	ে	e2	BENGALI VOWEL SIGN E
0x9c8	ৈ	e4	BENGALI VOWEL SIGN AI
0x9cb	ো	o2	BENGALI VOWEL SIGN O
0x9cc	ৌ	o4	BENGALI VOWEL SIGN AU
0x9cd	্	+	BENGALI SIGN VIRAMA (*hoshonto)*
0x9ce	ৎ	t2	BENGALI LETTER KHANDA TA
0x9dc	ড়	R	BENGALI LETTER RRA
0x9dd	ঢ়	r`	BENGALI LETTER RHA
0x9df	য়	y	BENGALI LETTER YYA
0x9e0	ৠ	R[	BENGALI LETTER VOCALIC RR. Rare character.

6. Description of method for word detection

Word boundaries in the orthography are determined by localization of white spaces (blank, tab, etc.).

In terms of word boundary issues, words in Bengali (such as compound words and stems with grammatical endings) often combine together to form one word. In such cases, words will be spelled without white spaces and will use the traditional spelling alterations associated with this phenomenon. Note, however, that in some cases those words will appear next to each other with white space (this carries a different meaning).

Spelling of words as either a compound (without white spaces) or as separate words is checked and standardized throughout the transcription project by identifying and reviewing word which have been spelled both together and apart. Occurrences of words both with and without white spaces typically carry different meanings.

Hyphens are used in compounds to join components together when any of the components do not carry meanings on their own.

7. Table containing all phonemes in the stipulated notation

The phonemic transcription of the words in this database uses X-SAMPA symbols, which can be found at http://www.phon.ucl.ac.uk/home/sampa/x-sampa.htm. The total number of phonemes is 54. There are 33 consonants, 2 semi-vowels, 10 vowels (8 monophthongs and 2 diphthongs) and 9 nasal vowels (7 monophthongs and 2 diphthongs). 6 of these are foreign phones (/v, z, Z, w, j, @/) which are not part of the native Bengali sound system but are commonly heard in English words.

BENGALI PHONE CHART

TYPICAL BENGALI CORRESPONDENCE	UNICODE	ROMAN	IPA	SAMPA	COMMENTS
CONSONANTS					
□	0x9aa	p	p	p	
□	0x9ab	P	f	f	Has /p_h/ as an allophone. Mainly occurs in English words e.g. "**ph**one, **f**inal".
			pʰ	p_h	allophone of /f/, mainly as naive pronunciation in English words. May be standard pronunciation in a few native Bengali words.
□	0x9ac	b	b	b	
□	0x9ad	B	bʰ	b_h	Has /v/ as an allophone
			v	v	allophone of /b_h/, mainly occurs in English words e.g. "**v**ideo, **v**ery".
□	0x9a4	t	t̪	t	
□	0x9ce	t2			occurs word-finally
□	0x9a5	T	t̪ʰ	t_h	
□	0x9a6	d	d̪	d	
□	0x9a7	D	d̪ʰ	d_h	
□	0x99f	t`	ʈ	t`	
□	0x9a0	T`	ʈʰ	t`_h	
□	0x9a1	d`	ɖ	d`	
□	0x9a2	D`	ɖʰ	d`_h	
□	0x995	k	k	k	
□	0x996	K	kʰ	k_h	
□	0x997	g	g	g	

TYPICAL BENGALI CORRESPONDENCE	UNICODE	ROMAN	IPA	SAMPA	COMMENTS
☐	0x998	G	gʰ	g_h	
☐	0x99a	c	ʧ	tS	
☐	0x99b	C	ʧʰ	tS_h	
☐	0x99c	j	ʤ	dZ	has allophones /z/ and /Z/
☐	0x9af	Y			
☐	0x99c	j	z	z	allophone of /dZ/ in English words, such as "**z**oo"
			ʒ	Z	allophone of /dZ/ in English words, such as "mea**s**ure, vi**si**on"
☐	0x99d	Z	ʤʰ	dZ_h	
☐	0x9ae	m	m	m	
☐	0x9a8	n	ɳ	n	
☐	0x9a3	n`			
☐	0x99e	J			
☐	0x999	N	ŋ	N	
☐	0x982	W			*onushshor*, indicates a velar nasal consonant.
☐	0x9b6	S	ʃ	S	has allophone /s/
☐	0x9b7	s`			
☐	0x9b8	s	s	s	allophone of /S/
☐	0x9b9	h	h	h	
☐	0x9b0	r	ɾ, r	r	
☐	0x9dc	R	ɽ	r`	
☐	0x9dd	r`			
☐	0x9b2	l	ɭ	l	
ENGLISH SEMI VOWELS					

TYPICAL BENGALI CORRESPONDENCE	UNICODE	ROMAN	IPA	SAMPA	COMMENTS
◻	0x9df	y	j	j	occurs in English words, such as "**u**nite, **y**es". May be substituted with native vowel /i/ or /e/
◻	0x987	I			
◻◻	0x993 0x9df	oy	w	w	occurs in English words, such as "**w**alk, **w**eb". May be substituted with native vowel /u/ or /o/.
◻◻	0x9cb 0x9df	o2y			
ORAL VOWELS					
◻	0x985	a	ɔ	O	
			ə	@	occurs in English words, such as "**a**bout, s**u**ppose". May be substituted with native vowel /O/.
◻	0x986	A	a	a	
◻	0x9be	A2			
◻	0x987	I	i	i	
◻	0x9bf	I2			
◻	0x988	i			
◻	0x9c0	i2			
◻	0x989	U	u	u	
◻	0x9c1	U2			
◻	0x98a	u			
◻	0x9c2	u2			
◻	0x98f	e	e	e	
◻	0x9c7	e2			
◻	0x98f	e	æ	{	
◻	0x9c7	e2			

TYPICAL BENGALI CORRESPONDENCE	UNICODE	ROMAN	IPA	SAMPA	COMMENTS
□□	0x9df 0x9be	yA2			
□	0x993	o	o	o	
□	0x9cb	o2			
□	0x990	e3	oi	oi	
□	0x9c8	e4			
□	0x994	o3	ou	ou	
□	0x9cc	o4			
NASAL VOWELS					
□□	0x985 0x981	aM	ɔ̃	O~	
□□	0x986 0x981	AM	ã	a~	
□□	0x987 0x981	IM	ĩ	i~	
□□	0x988 0x981	iM			
□□	0x989 0x981	UM	ũ	u~	
□□	0x98a 0x981	uM			
□□	0x98f 0x981	eM	ẽ	e~	
□□	0x98f 0x981	eM	æ̃	{~	may be rarer than other nasal vowels
□□	0x993 0x981	oM	õ	o~	
□□	0x990 0x981	e3M	oĩ	oi~	may be rarer than other nasal vowels
□□	0x994 0x981	o3M	oũ	ou~	may be rarer than other nasal vowels

OTHER SYMBOLS	
#	word boundary
.	syllable break

Notes:

- Note that the consonants /t, d, t_h, d_h, n, l/ are dental, as indicated by the IPA symbols used above. We have opted to use a simplified SAMPA representation /t, d, t_h, d_h, n, l/ for these phonemes since the difference between alveolar and dental is not contrastive in this language.
- The nasalized vowels are quite rare, but do occur in certain words and have phonemic status in Bengali. This phone set allows for a nasalized counterpart for all of the native vowels.
- The inherent vowel may be pronounced as /O/ or /o/. Care is taken in the creation of the lexicon to ensure that the appropriate pronunciation is given.
- The phoneme represented by SAMPA /O/ may also be pronounced as /Q/ (IPA [ɒ]) by some speakers. It is also the sound used in English words such as "hot".

8. List of rare phonemes

8.1. List of rare phones

The following phonemes are rare:

d`_h
ou
a~
i~
u~
e~
{~
o~
oi~
ou~
O~

8.2. List of foreign phones

The following phones are foreign (English):

@
j
w
v
Z
z

9. Other language specific items

9.1. Table of Digits

Digit	Digit Bengali	Bengali	Romanization
0	০	শূন্য	Su2n+Y

1	১	এক□	ek
2	২	দুই□	dU2I
3	৩	তিন□	tI2n
4	৪	চার□	cA2r
5	৫	পাঁচ□	pA2Mc
6	৬	ছয়□	Cy
7	৭	সাত□	sA2t
8	৮	আট□	At`
9	৯	নয়□	ny

9.2. Other Numbers

Number	Number Bengali	Bengali	Romanization
10	১০	□□	dS
100	১০০	শতক একশ	Stk ekS
10,000	১০,০০০	দশহাজার	dShA2jA2r
100,000	১,০০,০০০□	একলক্ষ□ □□□	eklk+s` lA2K
10 million	১,০০,০০,০০০□	এককোটি□ ক্রো□	ekko2t`I2 k+ro2R

9.3. Presentation of digits

Script Number, PIN Number and Isolated Digits (C1, C4, I1 and I2) digit strings have been written out in Bengali words.

Telephone numbers, credit card number, money amount and natural number (C2, C3, C5, M1 and N1) digits strings are written in Arabic numerals followed by the Bengali digit in parentheses.

Respondents were instructed to read out the telephone numbers, credit card numbers, money amounts and natural numbers in a natural manner. This may elicit either the native language or English (respondents were instructed not to code-switch).

9.4. Dates

Bengali has a traditional solar calendar which is 593 years behind the Gregorian calendar. Although the Bengali calendar is still used for public holidays, it has been almost entirely replaced by the Gregorian calendar in everyday use[4].

[4]Wikipedia, Bengali Calendar. http://en.wikipedia.org/wiki/Bengali_calendar

Prompted date items (D1) have been presented using the Gregorian calendar only. For spontaneous date items (D1), native speakers will be instructed to answer naturally and both Bengali and Gregorian calendar responses were accepted.

10. References

Cardona, George and Jain, Dhanesh. 2003. The Indo-Aryan Languages. Routledge.

Bhattacharya, Tanmoy. "Bangla" chapter in Garry, Jane and Rubino, Carl (eds). 2001. "Facts About the World's Languages". New English Publishing Associates, U.S.A.

Bhattacharja, S. 2007. Word formation in Bengali: A whole word morphological description and its theoretical implications. Munich: Lincom Europa.

Immihelp.com, Telephone numbering scheme in India www.immihelp.com/nri/phone-number-scheme-india.html.

Klaiman, M. H. "Bengali" chapter 4 in Comrie, Bernard (ed). 1990. The Major Languages of South Asia, The Middle East and Africa. Routledge.

Mitra, S. C. 2005. SaralBangalaAvidhan (Dictionary of the Bengali Language). New Bengal Press Pty Ltd.

Rajapurohit, B. B. 1986. Acoustic Studies in Indian Languages. Central Institute of Indian Languages, Mysore.

Seely, Clinton. 2002. Intermediate Bangla. LINCOM EUROPA.

Wikipedia, Bengali Calendar http://en.wikipedia.org/wiki/Bengali_calendar.