

# Unicode Agenda for Bangla

L2/09-294

**Bidyut Baran Chaudhuri**

Society for Natural Language Technology Research

&

Indian Statistical Institute,

Kolkata, India

# Indian Script and Bangla

- ❁ Most Indian Scripts are derived from Ancient Brahmi script.
- ❁ They are alpha-syllabary/abiguda class of scripts.
- ❁ Indian writing system started to evolve 3000 years ago.
- ❁ Perhaps inspired by Ancient Aramic, but have exceptional originality of Indian philologists.
- ❁ Alphabet matrix is arranged according to manner of articulation like unvoiced (unaspirated, aspirated), voiced (unaspirated, aspirated) versus place of articulation in mouth like velar, post-alveolar, alveolar, dental and bilabial.

# Brahmi Alpha Numerals

a	ā	i	ī	u	ū
e		ai	o	-ṃ	
ka	kha	ga	gha		
ca	cha	ja	jha	ña	
ṭa	ṭha	ḍa	ḍha	ṇa	
ta	tha	da	dha	na	
pa	pha	ba	bha	ma	
ya	ra	la	ḷa	va	
sa, ṣa	ṣa	śa	sa	ha	

ka	kā	ki	kī	ku	kū	ke	kē	kaṃ
la	lā	li	lī	lu	lū	le	lē	laṃ

one	two	three	four	five	six	seven	eight	nine
ten	twenty	thirty	forty	fifty	sixty	seventy	eighty	ninety
one hundred	one thousand							

# From Brahmi to Bangla

- ❁ Full-blown Brahmi script was active during the days of Christ, but its initial form started earlier.
- ❁ It branched into north and south Indian groups.
- ❁ By 800 AD a north variety named Kutila script evolved through Kushana-Gupta group of scripts.
- ❁ Kutila means complicated (the upper-caste people did not like the lower-caste people to learn writing and reading).
- ❁ By 1000 AD proto-Bangla script evolved.
- ❁ Proto modern Bangla script evolved by 1500 AD.
- ❁ By 18<sup>th</sup> century modern Bangla script was ready. There were 34 consonants and 10 vowels.





# Stabilization of Bangla Script

- ❁ Printing in Bangla started in late eighteenth century (Halhed, 1778).
- ❁ Full stop and double full stop were only punctuation marks noted in initial script.
- ❁ Other punctuation marks were borrowed from English.
- ❁ Vidyasagar introduced three more characters in mid nineteenth century by placing dot below three existing characters.
- ❁ Some characters like li and double-li became obsolete.
- ❁ This stabilized script system remained in use for 150 years.





# Further Modification of Bangla Script

- ❁ After 1900 AD Spelling correction and script correction debates gained momentum.
- ❁ Several correction suggestions were accepted through the initiative of Kolkata University.
- ❁ New Decimal monetary system, weighing standards etc were introduced around 1960s.
- ❁ Some of the older signs and symbols disappeared.
- ❁ Simplification in Representation of conjunct characters are being proposed since twenty years. There is still debate on which should be simplified.

# Development of Bangla ISCII and Unicode

- ❁ ISCII for Indian Languages were developed in 1980's through the initiatives of Dept. of Information Technology, Govt. of India.
- ❁ Bangla script too got an ISCII version.
- ❁ There has always been some problems in using Bangla ISCII for preparing electronic texts.
- ❁ The Bangla UNICODE code points appear to be based mainly on Bangla ISCII.
- ❁ So, it has problems too, though some of them are already solved.

# Unicode 5.1 for Bangla

	098	099	09A	09B	09C	09D	09E	09F
0	৐	ঐ	ঔ	ৱ	ী		ঋ	ৠ
1	ঁ		ড		ঢ		ণ	ত
2	ং		ঢ	ল	ণ		ত্ৰ	ত্ৰ
3	ং	ও	ং		ং		ং	ত্ৰ
4		ঙ	ত		ং			ত্ৰ
5	অ	ক	খ					খ
6	আ	খ	দ	শ			০	৩
7	ই	গ	ঞ	ষ	ৌ	ৌ	ূ	ূ
8	ঈ	ঘ	ন	স	ৈ		ৃ	ৃ
9	ঊ	ঙ		হ			ৄ	ৄ
A	ঊ	চ	প				৅	৅
B	ঋ	ছ	ফ		ৌ		৆	
C	৏	জ	ব	়	ৌ	ড	ড	
D		ঝ	ভ	হ	়	ঢ	ঢ	
E		ঞ	ম	া	ং		ঢ	
F	এ	ট	য	ি		য়	৏	

# Unicode 5.2 for Bangla

0980

Bengali

09FF

	098	099	09A	09B	09C	09D	09E	09F
0	ঐ	ঔ	ঋ	ঌ	঍	ঞ	ঠ	ড
1	ঐ	ঔ	ঋ	ঌ	঍	ঞ	ঠ	ড
2	ঐ	ঔ	ঋ	ঌ	঍	ঞ	ঠ	ড
3	ঐ	ঔ	ঋ	ঌ	঍	ঞ	ঠ	ড
4	ঐ	ঔ	ঋ	ঌ	঍	ঞ	ঠ	ড
5	অ	ক	খ					গ
6	আ	খ	দ	শ			০	৩
7	ই	গ	ধ	ষ	়ে	ৌ	১	।
8	ঐ	ঘ	ন	স	়ৈ		২	৮
9	উ	ঙ		হ			৩	০
A	ঊ	চ	প				৪	৷
B	ঋ	ছ	ফ		ৌ		৫	৶
C	ঌ	জ	ব	়	ৌ	ড়	৬	
D		ঝ	ভ	হ	়	ঢ	৭	
E		ঞ	ম	়া	ৎ		৮	
F	এ	ট	ষ	ি		য়	৯	

# Problems Remaining

- ❁ Rendition of Hasanta and two types of conjunct r + ja is clumsy with ZWJ and ZWNJ code points.
- ❁ No code point exists for (Khiya or Jukta-kha ক্ষ) as well as the Om-kar character ॐ .
- ❁ Unnecessary existence of a code point for right side of ou-kar া̣ .
- ❁ No code point exists for Urdha-comma ক'রে .
- ❁ Existence of many code points for old and obsolete symbols ৳ ৲ ৱ ৱ̣ ৱ̤ ৱ̥ ৱ̦ in the main code table.
- ❁ Unreasonable proposal of introducing extra code for transparent and non-transparent form of vowel modifiers গু ৐ .
- ❁ Code points for various signs need discussion.

# Our Proposals

1. Introduce a code point for **ক্ষ** in the table, after **হ** ie, at 09BA and for **ঐ** at 09D0.
2. Introduce a new code point for Ja-fala (**৳**) say after (**ঐ**) i.e. at 09C9 and use this to express all kinds of Ja-fala. The existing role of hasant and ZWNJ will continue. E.g.

অ + ্য + া --> অ্যা  
ত + ্য + া --> ত্যা  
র + ্ + য --> র্য  
র + ্য --> র্য  
দ + ্ + ZWNJ + গ --> দ্গ  
ন + ্ + ত + ্ + র + ্য --> ত্র্য

There will be no need for ZWJ code point in this scheme.

Contd..

3. There is no need to distinguish **ঐ** by using ZWJ.
4. Release the obsolete character code points by placing them in private use area.
5. Stop using the code point for **ঐ** unless there is other pressing reasons. It may create confusion for O-kar (**ঐ**)
6. (a) Should we use any of the existing code points for representing the upper comma which has different connotation in Bangla? We are in favor of a distinct code point.  
(b) Should we use the Devanagari code point of full-stop sign (danda) to represent Bangla full-stop also? Our suggestion is to have distinct code point for Bangla full-stops.  
(c) For representing signs for acronym, foot, inch, degree etc. for Bangla, the Unicode manual should have specific suggestions that are easily available in net.

Contd..

7. In the description of code points in Unicode manual there are several inadequacy which should be modified as follows:

09F4  BENGALI CURRENCY ~~NUMERATOR~~ SIGN FOR ONE ANNA

- not in current usage

09F5  BENGALI CURRENCY ~~NUMERATOR~~ SIGN FOR TWO ANNAS

- not in current usage

09F6  BENGALI CURRENCY ~~NUMERATOR~~ SIGN FOR THREE ANNAS


- not in current usage

09F7  BENGALI CURRENCY ~~NUMERATOR~~ SIGN FOR FOUR ANNAS

- not in current usage (A code point is needed for eight annas also)

09F8  BENGALI CURRENCY ~~NUMERATOR~~ SIGN FOR TWELVE ANNAS

- not in current usage

09F9  BENGALI CURRENCY ~~DENOMINATOR SIXTEEN~~ END MARKER  
AFTER ANNAS

- not in current usage

09FB  BENGALI GANDA MARK

- not in current usage



Any

Comment ?

Suggestion ?

Question ?

**Thank You**

0980

Bengali

09FF

	098	099	09A	09B	09C	09D	09E	09F
0	ঐ	ঔ	ঊ	ঋ	ঌ		঍	঎
1	ঁ		ড		ঢ		ণ	ত
2	ং		ঢ	ল	ণ		ত	র্
3	ঃ	ও	ঐ		ঊ		ঋ	ঌ
4		ঔ	ঊ		ঋ			র্
5	অ	ক	খ					গ
6	আ	খ	দ	শ			০	ঞ
7	ই	গ	ঘ	ষ	়ে	়ী	ূ	ৃ
8	ঈ	ম	ন	স	়ৈ		ৄ	৅
9	ঊ	ঙ		হ			৆	ে
A	ঋ	চ	প				ৈ	৉
B	ঌ	ছ	ফ		়ৌ		৊	
C	঍	জ	ব	়	়ৌ	ড	ঢ	
D		ঝ	ভ	হ	়	ঢ	ণ	
E		ঞ	ম	া	ং		ত	
F	এ	ট	য	ি		য়	ৌ	

BACK

	098	099	09A	09B	09C	09D	09E	09F
0	ঐ	ঔ	ঋ	ঌ	঍	঎	এ	ঐ
1	ঐ	ঔ	ঋ	ঌ	঍	঎	এ	ঐ
2	ঐ	ঔ	ঋ	ঌ	঍	঎	এ	ঐ
3	ঐ	ঔ	ঋ	ঌ	঍	঎	এ	ঐ
4	ঐ	ঔ	ঋ	ঌ	঍	঎	এ	ঐ
5	অ	ক	খ					ং
6	আ	খ	দ	শ			০	ঐ
7	ই	গ	ঘ	ষ	়ে	়ো	১	।
8	ঈ	ম	ন	স	়ৈ		২	৮
9	ঊ	ঙ		হ			৩	০
A	ঊ	চ	প				৪	৬
B	ঋ	ছ	ফ		়ৌ		৫	
C	ঌ	জ	ব	়	়ৌ	ড	৬	
D		ঝ	ভ	হ	়	ঢ	৭	
E		ঞ	য়	়া	ং		৮	
F	এ	ট	য	়ি		য়	৯	

BACK

In Devanagari, khiya is formed by combining two characters. In Bangla also, the current practice is to form it as follows:

ক+্+য → ক্ষ

However, in Bangla it is considered as single character and in Bangla dictionary it is ranked in between **ক** and **খ** . So, there should be a separate code point for it.

[BACK](#)

	098	099	09A	09B	09C	09D	09E	09F
0	ঐ	ঔ	ঋ	ঌ	঍	এ	উ	ঊ
1	ঋ	ঌ	঍	এ	উ	ঊ	ঋ	ঌ
2	঍	এ	উ	ঊ	ঋ	ঌ	঍	এ
3	এ	উ	ঊ	ঋ	ঌ	঍	এ	উ
4	উ	ঊ	ঋ	ঌ	঍	এ	উ	ঊ
5	ঊ	ঋ	ঌ	঍	এ	উ	ঊ	ঋ
6	ঋ	ঌ	঍	এ	উ	ঊ	ঋ	ঌ
7	এ	উ	ঊ	ঋ	ঌ	঍	এ	উ
8	এ	উ	ঊ	ঋ	ঌ	঍	এ	উ
9	এ	উ	ঊ	ঋ	ঌ	঍	এ	উ
A	এ	উ	ঊ	ঋ	ঌ	঍	এ	উ
B	এ	উ	ঊ	ঋ	ঌ	঍	এ	উ
C	এ	উ	ঊ	ঋ	ঌ	঍	এ	উ
D	এ	উ	ঊ	ঋ	ঌ	঍	এ	উ
E	এ	উ	ঊ	ঋ	ঌ	঍	এ	উ
F	এ	উ	ঊ	ঋ	ঌ	঍	এ	উ

BACK

	098	099	09A	09B	09C	09D	09E	09F
0		ঐ	ঔ	ঝ	ঞ		ঠ	ড
1	ঠ		ড		ঢ		ণ	ত
2	ঢ		ঢ	ণ	ত		থ	দ
3	ণ	ও	ঐ		ঐ		ঐ	ঐ
4		ঐ	ঐ		ঐ			ঐ
5	অ	ক	খ					গ
6	আ	খ	দ	শ			০	ঐ
7	ই	গ	ঘ	ষ	়ে	ৌ	ূ	।
8	ঐ	ঘ	ন	স	়ৈ		ূ	ূ
9	উ	ঐ		হ	়		ৃ	ৃ
A	উ	চ	প				ৄ	ৄ
B	ঋ	ৄ	ফ		ৌ		৅	৅
C	ৄ	জ	ব	়	ৌ	ড়	৆	
D		ঝ	ভ	হ	়	ঢ	ে	
E		ঞ	ম	়া	ৎ		ৈ	
F	এ	ট	ষ	়ি		য়	৉	

BACK

