Unicode Agenda for Bangla

Bidyut Baran Chaudhuri

Society for Natural Language Technology Research



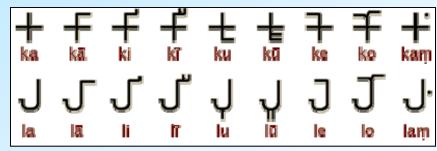
Indian Statistical Institute, Kolkata, India

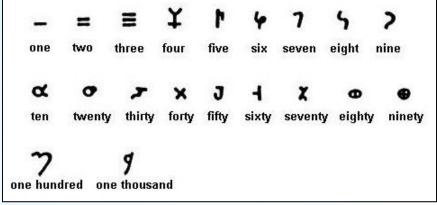
Indian Script and Bangla

- 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 ·:		ī	L u	L ū
Do	$\bigcap_{e} \bigcap_{e} \bigcap_{e$		o -m		
++ ka	} kha	√ ga	Шдһа		
d	Ф čha	[& C)	- jha	ña	
ţa	tha	- ḍa	⊸⊝ dha	T ņa	
$\left\langle \left\langle \right\rangle \right\rangle$	tha	- da	dha		
pa	© pha	♦ □	bha	X_{ma}	
$\bigcup_{y_a} \downarrow$	\ \ \ \ \ \ \ \ ra	$\bigcup_{la} \bigwedge$		→ o	
sa, ša	ل _{ة,a}	∧ ∕ \ sa	J. J.	ha	





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.
- Wutila 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 18th century modern Bangla script was ready. There were 34 consonants and 10 yowels.

Bangla Script Evolution

Kutila Script	1000- 1100 AD	1200 AD	1300 AD	1400 AD	1500 AD	1600 AD	1700 AD	Modern Bangla
म म	भ्र	भ	21	ञ	अ	丒	97	ज
भुजा	初	आ	श्रा	मा	आ	গ্রা	आ	আ
******* *******	°°	烮	色「	क्रा	छ	म	र्घ	支
₹	গ্	ৰ্ষ্	500	É	\$			竕
उउं	ъ	द	હ	ड	5	3	5	উ
উ য	\$	5		W-	v.		5	3
₹/	'থ্য	4/	₹/	₹	₹/		₽¥	श्रम
74	Ч	2	a	গ্ৰ	2	2	47	9
Ø	A)	4	% ?	V	(A)		S)	لأي
उ	ও 7	B	3	G	3	3	3	3
જે જ	জ	र्भ		(3)	3)	3	अ	3
6 6	Ø	क	B	क	ক	2.4	ক	क
Ω	য	251	SY	ख	7/	25	য	প্রা
N	٠	9	Z	হা	21	54	স	24
w	74	द्य	या	य	य	찍	घ	व
ग	5	ξ	દ	ξ	3		હ	ম্ভ
ਚ	4	ਚ	8	ਬ	ਚ	ব	4	Ъ
る	2 5	₹	Ф	Î	٤	₫.	죷.	苋
E.	天	<i>3</i> 5	37	ड	ङ	'S	₹.	5
T			又	क	ম		₹	ঝ
\$			73		P	racong tata	433	শ্ৰেচ

Contd...

Kutila Script	1100	1200 AD	1300 AD	1400 AD	1500 AD	1600 AD	1700 AD	Modern Bangla
S	2	3	প	છ	ε	Ü	ठ	B
40	0	2		0	8	ð	Ь	ठ
3	3	ड	3	3	3	उ	3	3
હ	ತ	ठ	8	2	હ	ઢ	2	5
(TL	m	M	CH	M	M	P	র	4
1	ス	3	Э	3	3	3	उ	
8	ય	થ	21	8	21	,52	25	N
22	ઢ	ą	ड	य	7,	5	T ₁	দ
ध	σ	9	Я	a	4	ध	ধ	ક
4	ন	+	न	न	न	F	न	न
Ч	য	य	ঘ	य	य	य	ध	भ
20	20	₹.	ध	হ্য	ひ	Σ̈́	ফ	ফ
4	4	a	8	a	a	4	4	a
ě	×	J	28	3	20	5	उ	Œ
25	¥	য	भ	স	म	अ	प्र	ম
IJ	ય	घ	য	য	य	य	य	A
T	ন	7	υγ	ব	ㅋ	ব	ব	র
w	8	ल	P	न	ल	ল	त	ल
4	4	ā	B	đ	ठ	đ	a	ব
28	94	3	ध	57	13	M	×	364
B	M	Я	9	ਬ	a	ख	A	78
ਖ਼	স	স	A	स	IJ	अ	म	34
di	ス	ह	a.	इ	70	25	*	2
ğ	₹5	各	Ť	मु	Ŧı	NA	325	蒸

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.
- Widyasagar 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.

The Alphabet Currently Used for Bangla

অআইঈউউঋএঐও ঔ

ক খ গ ঘ ঙ চ ছ জ ঝ ঞ ট ঠ ড ঢ ণ ত থ দ ধ ন প ফ ব ভ ম য র ল শ ষ স হ ড ঢ য ০ ০ ০ ০ ০ ০

সহড় ঢ়য় ৎং ঃ ° ককুত ভেকেককাত ক্ৰাক্ষা

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

- Significant Sig
- 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

0980 Bengali 09FF

	098	099	09A	09B	09C	09D	09E	09F
0		ઈ	ঠ	র	ী		¥	ৰ
		0990	0940	09 B0	0900		09€0	09F0
1	ै		ড ®A1		ुरु		ક્રુ ₀≆ા	ক ₹
2	ং		ঢ	ল	ূ		್ಲ	1
3	osaz ः	ও	₉₉₄₂	0982	0902 ∨ 0903		્રક્ક	%F2 Ъ
	0983	0993	09A3		0903		09E3	09F3
4		③	9		<i></i> √v ₂ 4)
5	ত	₹	থ					2
	0985	0995	09A5	mm			mm	09F5
6	<u>অ</u>	খ	ا	*T			0986	୬
7	গ্র	গ	ধ	ষ	ে	ी	>	I
	0987	0997	09A7	0987	0907	0907	09E7	09F7
8	紊	য	ন	স	्र		২	И
	0988	0998	9948	0968	0908 33333333		09E8	09F8
9	妈 §	ક્રુ		2			9	O 98F9
Α	₩ ₩	<u>Б</u>	প				8	•
^	098A	099A	09AA				09EA	09FA
В	₩	জ	ফ		ো		œ	
	0988	0998	09AB	annn	09CB	anna	09EB	
С	જ	জ	ব	়	ৌ	ড়	৬	
D		_{জ্ঞ}	^{09,60}	**************************************	्	<u>ن</u> 8000	9	
		099D	09AD	09BD	09CD	0900	09ED	<i>MIMI</i>
Е		æ	ম	া	٩		৮	
	HHHH	099E	09AE	098E	09CE	anna a	09EE	
F	এ	ট	য	ি		য়	৯	
	098F	099F	09AF	098F	HHHH	09DF	09EF	HHHHH

Unicode 5.2 for Bangla

Bengali 09B 09C 09D 09E 09F 099 09A 3 ৰ র ড 09F1 U ક 3 ত 09F4 থ 2 5 আ 0 2 গ ধ ষ য স И ৩ હ হ 0 09F9 ঊ 8 DOFA 0 ফ ব ড় ভ 2 D OSED 13 ो Е য য় ৯

0980

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 क
) as well as the Om-kar character
 .
- Solution Unnecessary existence of a code point for right side of ou-kar of.
- ⊗ No code point exists for Urdha-comma ক'রে .
- ⊕ Existence of many code points for old and obsolete symbols
 ▶ ३ ♥ ♀♀♀ in the main code table.
- Unreasonable proposal of introducing extra code for transparent and non-transparent form of vowel modifiers 18.
- Code points for various signs need discussion.

Our Proposals

- 1. Introduce a code point for ፮ in the table, after ₹ ie, at 09BA and for ॐ at 09D0.
- 2. Introduce a new code point for Ja-fala (7) say after (70) 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.

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

- 4. Release the obsolete character code points by placing them in private use area.
- 5. Stop using the code point for of unless there is other pressing reasons. It may create confusion for O-kar ((c))
- 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.

- 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 V BENGALI CURRENCY NUMERATOR SIGN FOR TWO ANNAS
 - not in current usage
- 09F6 J 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 N 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 S BENGALI GANDA MARK
 - not in current usage

Any

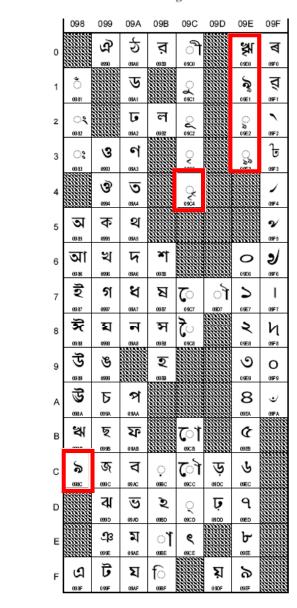
Comment?

Suggestion?

Question?

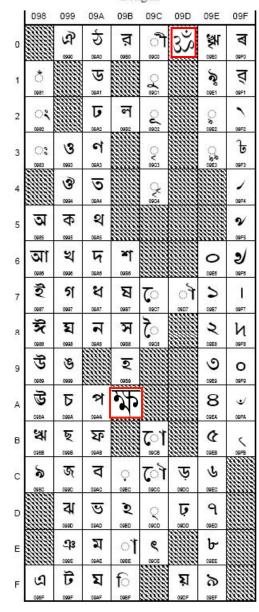
Thank You

	098	099	09A	09B	09C	09D	09E	09F
0		₽,	ঠ	র	ी		ৠ	ৰ
		0990	09A0	09B0	0900	<i>GHHHH</i>	09E0	09F0
1	ঁ		৮		ू		مج	ৱ
	0981	SHIPPINE STATES	09A1	amma	0901	HHHHH.	09E1	09F1
2	ং ®82		5	ខ	ূ •••		္စ ဖ≘း	09F2
3	ः	ઉ	ণ		ې ووون		្ណ	৳
	0983	0993	09A3		0903	minimize	09E3	09F3
4		3	9		े ्र			√ 09F4
	annanan.	944	9877		eriene.		***************************************	94.4
5	<u>অ</u> ®85	ক	ଥ					% 09∓5
	0000	9999	9949		*******	******		0.0
6	আ	খ	দ	শ ***			0986	2
	0986	0996	09A6	0916	100000		09E6	09F6
7	র	গ	ধ	ষ	<u>ে</u>	्री	>	09F7
	0987	0997	09A7	0987	0907	0907	09E7	09F7
8	₹	হ	ন	স	ी		২	И
	0988	0998	09A8	0958	0908	111111111	09E8	09F8
9	ق و	હ	0943	ર			9	O 09F9
А	ঊ	Б	প				8	•
	098A	099A	09AA	ALLEH ELLEN	HHH	HHHH	09EA	09FA
В	₩	₩ ₩	₹ 09A8		ুো		€	
				********		********		Activities to
С	৯	জ	ব	়	ৌ	ড়	હ	
	0.600	099C	09AC	09BC	09CC	09DC	09EC	<i>1000000</i>
D		ঝ	<u>9</u>	Q	্	**************************************	٩ •••	
	THE STATE OF					THE PERSON		Treasure .
E		æ æ	ম ®	ୀ	€		৮	
	*******		774		dillini.	******		de la ferra de
F	এ	ট	য	ি		য়	৯	
	09 8F	099F	09AF	098F	HHHH	090F	09EF	шини



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

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



	098	099	09A	09B	09C	09D	09E	09F
0		જ	ঠ	র	ী		첾	ৰ
		0990	09A0	0980	0900		09E0	09F0
1	ð		ড	0080	ૢૢૺ૾ૢ		مج	ৱ
	0981	HILLER	09A1	dilli.	0901	anna	09E1	09F1
2	ং		ঢ	ল	ূ		្វ	O9F2
	0982	1411141	09A2	1711117	0902	dillin.	09E2	09F2
3	ಂ	ও	ণ		>		ૢ	િ
	0983	0993	09A3	9111111	09C3		09E3	09F3
4		3	<u>5</u>		Ş			1
	dilli	0994	0984	Hillis	177777	44444	44444	09F4
5	অ	ক	থ					2
	0985	0995	09A5	1111111	SHILL S	anna	HHH	09F5
6	আ	খ	দ	*1			0	d
	0985	0995	09A6	0985	HHB		09E5	09F6
7	ই	গ	ধ	ষ	ে	ी	۵	1
	0987	0997	09A7	0987	0907	0907	09E7	09F7
8	ऋ	য	ন	স	ৈ	0907	২	И
	0988	0998	09AB	0988	0908	ann	09E8	09FB
9	₹	છ	CGAS	2	5		<u>ඉ</u>	O 09F0
	UNCH	UPPE	THILL.	0000	THE STATE	HHH	VACA	User w
Α	ভ	5	প				8	•
	098A	099A	CGAA	Hillie	*****	HHHH)	ODEA	OGFA
В	₩	ছ জে	S 8480		ুো		€	Ç
				- min				MILL
С	৯	জ	ব	Ç	ুী	ড়	હ	
	20000	699C	OSAC	0990	9900	0900	09EC	44444
D		ঝ	<u>ভ</u>	Q	্	<u>ب</u>	9	
- 8	41111	ė marianininininininininininininininininini	15000	77-77887	272	111111	1800	diete
Ε		্ৰাঃ	ম	্ৰ	٩		Ъ	
	411141	099E	OWAE	DOBE	cities.	141114	OVEE	411111
F	এ	ট	য	ি		য়	৯	
	098F	099F	09AF	098F	(HIII)	OSCF	WEF	111111