9. *Unicode Standardization**

9.1 Bangla Code Chart

	098	099	09A	09B	09C	09D	09E	09F
0		ঐ 0990	ঠ 09A0	র 09B0	ী 09C0		ৠ 09E0	ৰ 09F0
1	ঁ 0981		ড 09A1	ৱ 09B1	ু 09C1		ৡ 09E1	ৱ 09F1
2	ং 0982		ঢ 09A2	ল 09B2	ূ 09C2		ৢ 09E2	৲ 09F2
3	ঃ 0983	ও 0993	ণ 09A3		ৃ 09C3		ৣ 09E3	৳ 09F3
4		ঔ 0994	ত 09A4		ৄ 09C4		৤ 09E4	৴ 09F4
5	অ 0985	ক 0995	থ 09A5	৵ 09B5			৥ 09E5	৵ 09F5
6	আ 0986	খ 0996	দ 09A6	শ 09B6			০ 09E6	৶ 09F6
7	ই 0987	গ 0997	ধ 09A7	ষ 09B7	ে 09C7	ৗ 09D7	১ 09E7	৷ 09F7
8	ঈ 0988	ঘ 0998	ন 09A8	স 09B8	ৈ 09C8		২ 09E8	৸ 09F8
9	উ 0989	ঙ 0999		হ 09B9			৩ 09E9	৹ 09F9
A	ঊ 098A	চ 099A	প 09AA	় 09BA			৪ 09EA	৺ 09FA
B	ঋ 098B	ছ 099B	ফ 09AB		ো 09CB		৫ 09EB	৻ 09FB
C	ঌ 098C	জ 099C	ব 09AC	ঽ 09BC	ৌ 09CC	ড় 09DC	৬ 09EC	ৼ 09FC
D		ঝ 099D	ভ 09AD	া 09BD	্ 09CD	ঢ় 09DD	৭ 09ED	৽ 09FD
E		ঞ 099E	ম 09AE	ী 09BE			৮ 09EE	
F	এ 098F	ট 099F	য 09AF	ি 09BF		য় 09DF	৯ 09EF	

** In continuation from issue nos. 4 & 5 for other Indian Languages.*

9.1.1 Code Chart Details of Bangla

Code Point	Character	Description
Various signs		
0981	ঁ	BENGALI SIGN CANDRABINDU
0982	ং	BENGALI SIGN ANUSVARA
0983	ঃ	BENGALI SIGN VISARGA
Independent vowels		
0985	অ	BENGALI LETTER A
0986	আ	BENGALI LETTER AA
0987	ই	BENGALI LETTER I
0988	ঈ	BENGALI LETTER II
0989	উ	BENGALI LETTER U
098A	ঊ	BENGALI LETTER UU
098B	ঋ	BENGALI LETTER VOCALIC R
098C	ঌ	BENGALI LETTER VOCALIC L
098D		<reserved>
098E		<reserved>
098F	এ	BENGALI LETTER E
0990	ঐ	BENGALI LETTER AI
0991		<reserved>
0992		<reserved>
0993	ও	BENGALI LETTER O
0994	ঔ	BENGALI LETTER AU
Consonants		
0995	ক	BENGALI LETTER KA
0996	খ	BENGALI LETTER KHA
0997	গ	BENGALI LETTER GA
0998	ঘ	BENGALI LETTER GHA
0999	ঙ	BENGALI LETTER NGA
099A	চ	BENGALI LETTER CA
099B	ছ	BENGALI LETTER CHA
099C	জ	BENGALI LETTER JA
099D	ঝ	BENGALI LETTER JHA
099E	ঞ	BENGALI LETTER NYA
099F	ট	BENGALI LETTER TTA
09A0	ঠ	BENGALI LETTER TTHA
09A1	ড	BENGALI LETTER DDA
09A2	ঢ	BENGALI LETTER DDHA
09A3	ণ	BENGALI LETTER NNA
09A4	ত	BENGALI LETTER TA
09A5	থ	BENGALI LETTER THA
09A6	দ	BENGALI LETTER DA
09A7	ধ	BENGALI LETTER DHA
09A8	ন	BENGALI LETTER NA
09A9		<reserved>
09AA	প	BENGALI LETTER PA
09AB	ফ	BENGALI LETTER PHA
09AC	ব	BENGALI LETTER BA = Bengali va, wa
09AD	ভ	BENGALI LETTER BHA
09AE	ম	BENGALI LETTER MA
09AF	য	BENGALI LETTER YA
09B0	র	BENGALI LETTER RA
09B1	ৰ	BENGALI LETTER RA WITH MIDDLE DIAGONAL • Used in Assamese
09B2	ল	BENGALI LETTER LA
09B3		<reserved>
09B4		<reserved>
09B5	ৱ	BENGALI LETTER VA WITH LOWER DIAGONAL • Used in Assamese & Manipuri
09B6	শ	BENGALI LETTER SHA
09B7	ষ	BENGALI LETTER SSA
09B8	স	BENGALI LETTER SA
09B9	হ	BENGALI LETTER HA
09BA	়	BENGALI INVISIBLE LETTER
Various signs		
09BC	়	BENGALI SIGN NUKTA • For extending the alphabet to new letters Used as abbreviation delimiter
09BD	ঽ	BENGALI SIGN AVAGRAH

Dependent vowel signs

09BE	া	BENGALI VOWEL SIGN AA
09BF	ি	BENGALI VOWEL SIGN I • stands to the left of the consonant
09C0	ী	BENGALI VOWEL SIGN II
09C1	ু	BENGALI VOWEL SIGN U
09C2	ূ	BENGALI VOWEL SIGN UU
09C3	ৃ	BENGALI VOWEL SIGN VOCALIC R
09C4	ৄ	BENGALI VOWEL SIGN VOCALIC RR
09C5		<reserved>
09C6		<reserved>
09C7	ে	BENGALI VOWEL SIGN E • stands to the left of the consonant
09C8	ৈ	BENGALI VOWEL SIGN AI • stands to the left of the consonant
09C9		<reserved>
09CA		<reserved>
09CB	ো	BENGALI VOWEL SIGN O • pieces on both sides of the consonant = 09C7 ে 09BE া
09CC	ৌ	BENGALI VOWEL SIGN AU • pieces on both sides of the consonant = 09C7 ে 09D7 ৗ

Various signs

09CD	্	BENGALI SIGN HALANT
09CE		<reserved>
09CF		<reserved>
09D0		<reserved>
09D1		<reserved>
09D2		<reserved>
09D3		<reserved>
09D4		<reserved>
09D5		<reserved>
09D6		<reserved>
09D7	ৗ	BENGALI AU LENGTH MARK

Additional consonants

09DC	ড়	BENGALI LETTER RRA
09DD	ঢ়	BENGALI LETTER RHA
09DE		<reserved>
09DF	য়	BENGALI LETTER YYA

Generic additions

09E0	ৠ	BENGALI LETTER VOCALIC RR • Not in current usage
09E1	ৡ	BENGALI LETTER VOCALIC LL • Not in current usage
09E2	ৢ	BENGALI VOWEL SIGN VOCALIC L • Not in current usage
09E3	ৣ	BENGALI VOWEL SIGN VOCALIC LL • Not in current usage
09E4	।	BENGALI SIGN URNACHHED
09E5	॥	BENGALI SIGN DEERGH VIRAM

Digits

09E6	০	BENGALI DIGIT ZERO
09E7	১	BENGALI DIGIT ONE
09E8	২	BENGALI DIGIT TWO
09E9	৩	BENGALI DIGIT THREE
09EA	৪	BENGALI DIGIT FOUR
09EB	৫	BENGALI DIGIT FIVE
09EC	৬	BENGALI DIGIT SIX
09ED	৭	BENGALI DIGIT SEVEN
09EE	৮	BENGALI DIGIT EIGHT
09EF	৯	BENGALI DIGIT NINE

Bengali-specific additions

09F0	ৰ	BENGALI LETTER RA WITH MIDDLE DIAGONAL • Not in use

09F1	ৰ	BENGALI LETTER RA WITH LOWER DIAGONAL = BENGALI LETTER VA WITH LOWER DIAGONAL • Not in use
09F2	৲	BENGALI RUPEE MARK
09F3	৳	BENGALI RUPEE SIGN
09F4	৴	BENGALI CURRENCY NUMERATOR ONE • Not in current usage
09F5	৵	BENGALI CURRENCY NUMERATOR TWO • Not in current usage
09F6	৶	BENGALI CURRENCY NUMERATOR THREE • Not in current usage
09F7	৷	BENGALI CURRENCY NUMERATOR FOUR
09F8	৸	BENGALI CURRENCY NUMERATOR ONE LESS THAN THE DENOMINATOR
09F9	৹	BENGALI CURRENCY DENOMINATOR SIXTEEN
09FA	৺	BENGALI ISSHAR
09FB	৻	BENGALI SIGN YA PHALLA
09FC	ৼ	BENGALI SIGN KHAND TA
09FD	৽	BENGALI LETTER KHYA • Used in Assamese

9.1.2 Bangla Script Details

Introduction

Historically the Bangla language is also known as Bengali, Bangal-Bhasha, Banga-Bhasha.

Demographic Information

Bangla is fourth most popular language in the world. It is one of the major Scheduled languages in India and is the State language of Bangladesh. It is the official language of West-Bengal and Tripura. As first language 70,561,000 in India (1997 IMA), 100,000,000 in Bangladesh (1994 UBS), and 36439000 in other countries (1999 WA); including second language, globally 211,000,000 speakers (1999 WA).

Genetic affiliation and history

Bangla along with two other cognate languages, Assamese and Oriya, as well as *Magadhi*, *Maithili* and *Bhojpuri* in southeast zone forms a linguistic group. Their immediate source can be traced back to the *Magadhi Prakrit* or *Eastern Prakrit* which was brought to this area from Magadh (or Bihar) and the language of *Gauda-Banga* with other eastern languages developed through *Magadhi Apabhramsa.* Genetically Bangla is derived from Indo-Aryan (IA) or the Indic sub-branch of the Indo-Iranian branch of the Indo-European (IE) family of languages.

The literary documents of IA language in Indian Peninsula can be classified into three periods according to their linguistic changes. (i) Old Indo-Aryan (OIA) (1500 BC/1200 BC - 600 BC) (ii) Middle Indo-Aryan (MIA) (600 BC -1000 AD) (iii) New Indo-Aryan (NIA) (1000 AD - Present Time)

The inadequacy of written documents of immediate Pre-Bengali period is one of the most important limitations to find out the gradual change from Apabhramsa, to the historic period of Bangla (16th century AD). There is no document of Magadhi Apabhramsa except a small inscription. Till the 16th century, all the documents are copies of the original with varying degrees of correctness. After the 16th Century AD, the documents have more or less survived. Based on these documents, Bangla has three distinct periods:

1. Old Bangla: AD 950/1000 - AD 1200/1350
2. Middle Bangla : AD 1350 - AD 1800
   (i) Early Middle Bangla AD 1350 - AD 1450/1600
   (ii) Late Middle Bangla AD 1600 - AD 1800
3. Modern Bangla: AD 1800 - today.

The domain of use

The Bangla language is formally taught as first language in all public schools and most of the private schools in West Bengal and Tripura and as second or third language in many premier schools in other parts of India. Due to its rich literature, it enjoys departmental status in most colleges and universities not only in West Bengal but also in other major states of India. In Bangladesh, it is the principal language used in all formal and educational activities. In many European universities and US, there are full professors on Bangla language. In technical and professional courses, Bangla is also used in some parts of West Bengal and Tripura and several textbooks are available and published regularly by Governmental and private publications. In lower level of administrations, it is popularly used in West Bengal and Tripura but in higher level of administrations, Bangla is rarely used though government encourages using it. Bangla is the common language in the transactions of the assembly of West Bengal and Tripura. As a medium of journalism, it flourishes in newspapers, magazine, radio-broadcasts, and TV-telecasts. It is also one of the most popular mediums of feature films produced in eastern part of Indian subcontinent.

Bangla Script

Origin and development

The word *lipi* (script/alphabet) in Bangla came from the Sanskrit *lip* meaning 'to plaster' or 'to apply'. In ancient India, writings were normally done by scribing on palm leaves with a stylus and then applying ink on it. The *word lipi* probably originated from this method of applying layers of ink on leaves. Till recently, early scholars are of the opinion that the art of writing in India dates back to the period of Asoka (3rd century BC), when inscriptions were engraved in two different scripts, which are known as Brahmi *nit* Kharoshti. These scripts are mainly of Sumerian origin. However, the recent discovery of a number of seals bearing inscriptions in an unknown script, however, has brought to light that the art of writing in India is as old as the third or fourth millennium BC to which these inscriptions are referred to, on the basis of their similarity with the Sumerian.

Bangla script has been derived from the eastern variety of Brahmi script, known as *Kutilalipi,* which took a distinctive form around the 7th century. The script evolved over the centuries, acquired the cursive form.

The evolution of the Bangla script with the advent of printing technology in Bengal gives an interesting picture. The first Bangla scripts (movable type) were used in the printing of N. B. Halhed's book, 'A Grammar of the Bengal Language' (1778). In the year 1785, Warren Hastings requested another civilian, Charles Wilkins to cut punches for Bangla printing characters. Wilkins is 'the father of movable type in Bangla'. He also taught Panchanan Karmakar, a renowned artist in Bengal, the technique of cutting punches for printing characters. Karmakar and his family subsequently became famous in the light of the sart-evolution in Bangla printing technology. The printing types ultimately became more angular with sharper turns and edges. The movable types, first developed in Korea, were introduced into Bengal from Europe, where it evolved independently. The movable type technique continued in the Bengal printing industry for a long time. The Linotype technique, invented by M. Thellar in 1886, was introduced into Bangla printing in 1935 by S. C. Majumdar, R. Basu and others. Within a few years the more advanced monotype technology came to Bangla printing. A significant contribution to Bangla publication was also made by Ishwarchandra Vidyasagar, a great educationist and social reformer.

Initially in Bangla handmade typing, more than five hundred typing characters were required in each font, but the number had been gradually reduced. The existing Bangla code of signs in the foundry type consists of 448 to 536 characters. Linotype provides for 292 characters of which 260 are good enough for the ordinary job. Monotype composition provides for 319 characters.

In the area of computerised composing popularly known as Desk Top Publishing (DTP), a large number of Bangla type fonts have appeared in the market. The most significant contribution to Bangla computerised font designing came from the Institute of Typographical Research (ITR) in Pune. The Bangla fonts, Dilip, Devasree, Rabindra, Suvarna, Uttama and Vivek, introduced by ITR are in widespread use today. C-DAC, Pune introduced several Bangla TTF such as BN-TT Durga, BN-TT Bidisha and BN-TT Satyajit. The fonts of Modular Info Tech in the Shreelipi series are also popular in particular, for web designing purpose.

Technical Characteristics

Like other Indian scripts that evolved from the Brahmi script, Bangla is also written from left to right and consists of sequence of simple and complex characters.

Bangla Alphabets

Basic characters

There are 11 vowel and 39 consonant Live characters in modern Bangla alphabet. They are called *basic characters*. As each of these basic characters has a stand-alone code (character) in the common core, this set is also referred to as *primary* or *independent character set*. These live characters along with some {\it obsolete} or {\it very rarely used characters} are mapped to Hexadecimal code in ISCII Standard, from 131-144 (for vowels) and from 146-232 (for consonants).

In Unicode independent vowels are placed in the range U0985 to U0994. Some of the places are reserved (098D, 098E, 0991, 0992). Consonants ranges from U0995 to U09B9 and again from U09DC to U09DF.

The basic characters and their standard representations in Roman characters are shown in the following two tables. The concept of upper/ lower case is absent in Bangla.

অ (A) আ (AA) ই (I) ঈ (II) উ (U) ঊ (UU) ঋ (R) এ (E) ঐ (AI) ও (O) ঔ (AU)

Table I : Live Vowel set

ক (KA) খ (KHA) গ (GA) ঘ (GHA) ঙ (NGA) চ (CA) ছ (CHA) জ (JA) ঝ (JHA) ঞ (NYA) ট (TTA) ঠ (TTHA) ড (DDA) ঢ (DDHA) ণ (NNA) ত (TA) থ (THA) দ (DA) ধ (DHA) ন (NA) প (PA) ফ (PHA) ব (BA) ভ (BHA) ম (MA) য (YA) র (RA) ল (LA) শ (SHA) ষ (SSA) স (SA) হ (HA) ড় (RRA) ঢ় (RHA) য় (YYA)

Table II : Live Consonant set

Modifiers and compound characters

A vowel other than অ (A) following a consonant takes a modified shape, depending on the position of the vowel: to the left, right (or both) or bottom of the consonant. These are called *dependent vowel set*, *vowel modifiers*, *or vowel allographs.*

া (AA) ি (I) ী (II) ু (U) ূ (UU) ৃ (R) ে (E)

ৈ (AI) ো (O) ৌ (AU)

Table III : Vowel Modifier

'◌' indicates the consonant character position. For example, for the first consonant character ক (KA), the vowel allographs are supposed to be attached with the consonant in the following way: কা, কি, কী, কু কূ, কে, কৈ, কো and কৌ. The vowels উ (U) ঊ (UU) ঋ (R) may take different modified shapes when attached to some consonant characters. They also change the shape of some consonant characters to which they are attached. In Unicode, modifiers (dependent vowel signs) are ranged between U09BE to U09CC.

Like in other Indic scripts, some of these vowel modifiers use two-part vowel signs. In those vowels one-half of the vowel is placed on each side of a consonant letter or cluster – for example, ো (O, U+09CB) and ৌ (AU, U+09CC). It may be noted that in Unicode, these vowel signs are coded in each case in the position in the charts isomorphic with the corresponding vowel in Devanagari. Hence Bangla vowel sign ৌ (AU, U+09CC) is isomorphic with Devanagari vowel sign ौ (AU, U+094C). To provide compatibility with existing implementations of the scripts that use two-part vowel signs, the Unicode standard explicitly encodes the right-half of these vowel signs; for example, Bangla length mark of AU, ৗ (U+09D7) represents the right-half glyph component of Bangla vowel sign ৌ (AU, U+09CC).

A consonant following (preceding) a consonant is represented by a modifier called *consonant modifier* if the shape of other character to which it is attached, remains unaltered. Among consonant modifiers, ্য (YA-phalaa) is the most frequent one. This character is a presentation form of the character য (YA, U+09AF). It takes this shape depending upon the context. It could be applied to nominal form of any consonant or conjunct or even to the vowel. Application of this character may either bring in repetitive sound or 'ya'-ness of the character it is applied to. Though it is a presentation form of an existing character, it should be treated like a separate character because of the two reasons. This character may appear along with a vowel where the conjunct formation rules cannot be applied. As for example:

অ + ্য + া = অ্যা

Apart from that, র (RA, U+09B0) and য (YA, U+09AF) combination has two different conjunct forms:

র + য = র্য

র + য = র‍্য

Using the same conjunct rule both of them cannot be formed. If it is treated as a combining character, rather than a original form of YA, both of these issues could be solved.

If the shape/size of all involved characters are changed from that of their respective basic characters, then the cluster of these attached consonants is called a compound character or *yuktakshar*. In Bangla, consonant conjuncts can be formed as vertical conjuncts, where the components are placed vertically unlike the normal conjuncts where components are placed side-by-side. To construct these conjuncts, ligatures and consonants might take upper-half and lower-half forms. Compounding of two consonants is most abundant although three consonants can also be compounded. There are about 250 compound characters, of which a subset of frequently used characters is shown in Table below. The total number of basic, modified and compound characters is about 300.

Two characters are conjunct: ক্ক (KA+KA) ঙ্ক (NGA+KA) ল্ক (LA+KA) ষ্ক (SSA+KA) স্ক (SA+KA) ঙ্খ (NGA+KHA) স্খ (SA+KHA) ঙ্গ (NGA+GA) ল্গ (LA+GA) গ্গ (GA+GA) চ্চ (CA+CA) ঞ্চ (NYA+CA) শ্চ (SHA+CA) চ্ছ (CA+CHA) ঞ্ছ (NYA+CHA) শ্ছ (SHA+CHA) L (JA+JA) ঞ্জ (NYA+JA) জ্ঞ (BA+JA) জ্ঝ (JA+JHA) ঞ্ঝ (NYA+JHA) জ্ঞ (JA+NYA) চ্ঞ (CA+NYA) ক্ট (KA+TTA) ট্ট (TTA+TTA) ণ্ট (NNA+TTA) ন্ট (NA+TTA) প্ট (PA+TTA) ল্ট (LA+TTA) ষ্ট (SSA+TTA) স্ট (SA+TTA) ণ্ঠ (NNA+TTHA) ন্ঠ (NA+TTHA) ষ্ঠ (SSA+TTHA) ড্ড (DDA+DDA) ণ্ড (NNA+DDA) ন্ড (NA+DDA) ল্ড (LA+DDA) ণ্ঢ (NNA+DDHA) ন্ঢ (NA+DDHA) ণ্ণ (NNA+NNA) ষ্ণ (SSA+NNA) হ্ণ (HA+NNA) ক্ত (KA+TA) ত্ত (TA+TA) ন্ত (NA+TA) প্ত (PA+TA) স্ত (SA+TA) থ্ত (TA+THA) ন্থ (NA+THA) স্থ (SA+THA) দ্দ (DA+DA) ন্দ (NA+DA) ব্দ (BA+DA) গ্ধ (GA+DHA) দ্ধ (DA+DHA) ন্ধ (NA+DHA) ব্ধ (BA+DHA) গ্ন (GA+NA) ঘ্ন (GHA+NA) ত্ন (TA+NA) ধ্ন (DHA+NA) ন্ন (NA+NA) প্ন (PA+NA) ম্ন (MA+NA) শ্ন (SHA+NA) স্ন (SA+NA) হ্ন (HA+NA) প্প (PA+PA) M (MA+PA) ল্প (LA+PA) ষ্প (SSA+PA) স্প (SA+PA) ম্ফ (MA+PHA) ল্ফ (LA+PHA) ষ্ফ (SSA+PHA) স্ফ (SA+PHA) ক্ব (KA+VA) গ্ব (GA+VA) জ্ব (JA+VA) ট্ব (TTA+VA) ণ্ব (NNA+VA) ত্ব (TA+VA) থ্ব (THA+BA) দ্ব (DA+BA) ধ্ব (DHA+BA) ন্ব (NA+BA) ব্ব (BA+BA) ম্ব (MA+BA) ল্ব (LA+BA) শ্ব (SHA+BA) স্ব (SA+BA) হ্ব (HA+BA) চ্ব (CA+BA) ড্ব (DDA+BA) ভ্ব (MA+BHA) ড্ভ (DA+BHA) ক্ম (KA+MA) গ্ম (GA+MA) ঙ্ম (NGA+MA) ণ্ম (NNA+MA) ত্ম (TA+MA) দ্ম (DA+MA) ন্ম (NA+MA) ম্ম (MA+MA) ল্ম (LA+MA) শ্ম (SHA+MA) ষ্ম (SSA+MA) স্ম (SA+MA) হ্ম (HA+MA) ক্র (KA+RA) খ্র (KHA+RA) গ্র (GA+RA) ঘ্র (GHA+RA) প্র (PA+RA) দ্র (DA+RA) ধ্র (DHA+RA) জ্র (JA+RA) ত্র (TA+RA) থ্র (THA+RA) ব্র (BA+RA) ভ্র (BHA+RA) ফ্র (PHA+RA) ট্র (TTA+RA) ড্র (DDA+RA) হ্র (HA+RA) স্র (SA+RA) শ্র (SHA+RA) ম্র (MA+RA) ক্ল (KA+LA) গ্ল (GA+LA) প্ল (PA+LA) ফ্ল (PHA+LA) ব্ল (BA+LA) ম্ল (MA+LA) ল্ল (LA+LA) শ্ল (SHA+LA) স্ল (SA+LA) হ্ল (HA+LA) ক্ষ (KA+SSA) ক্স (KA+SA) ন্স (NA+SA) প্স (PA+SA)

Three Characters are Conjunct: ক্ষ্ণ (KA+SSA+NNA) ক্ষ্ম (KA+SSA+MA) চ্ছ্র

(CA+CHA+BA) চ্ছ (CA+CHA+RA) জ্জ (JA+JA+VA) ঙ্ (NNA+DDA+RA) ত্ (TA+TA+BA) দ্ভ (DA+BHA+RA) ন্ত (NA+TA+BA) ন্ত্র (NA+TA+RA) ন্দ্র (NA+DA+RA) ন্ধ (NA+DA+BA) ম্প্র (MA+PA+RA) ম্ভ (MA+BHA+RA) স্প্র (SA+PA+RA) স্ত্র (SA+TA+RA) ন্ধ (NA+DHA+RA) ষ্ট্ (SA+TTA+RA) ষ্ত্ (SA+TA+BA) ল্ড (LA+DDA+RA) ক্ষ্ (SSA+KA+RA) ভ্ (DA+BHA+RA) দ্ধ (DA+BHA+RA) ন্ড্র (NA+DDA+RA) ঙ্ (NNA+DDA+RA) ক্ত্র (SA+KA+RA) ক্ট্ (KA+TTA+RA) ষ্প্র (SSA+PA+RA) দ্দ্ব (DA+DA+RA) ণ্ট (NNA+TTA+RA)

Table V : Conjunct Character set in Bangla

In the following, technical issues related to few characters (including some special characters) are separately elaborated:

Khanda-ta (ৎ)

It is a combining character, which is widely used in Bangla script. For some transliterated words it might appear at the beginning. Examples: ৎকম্পন মহৎ ৎজিলার্ড and নাৎসী.

Visarga (ঃ)

This character is used for writing Sanskrit words in Bangla script.

Nukta

Bangla script does not use '়', the Nukta sign (U+09BC) explicitly. Internally ড়, (RRA, U+09DC), ঢ় (RHA, U+09DD) and য় (YYA, U+09DF) are formed with the help of Nukta sign.

Avagraha

The *avagraha* is used for indicating the presence of the vowels অ (A) and আ (AA) that are sometimes elided in enphonic (*sandhi*) combination, which is a pervasive feature of Sanskrit and its cognate languages like Bangla. Such elision occurs when the A or B at the beginning of a word follows a word having আ (AA) or ঃ (visarga) at the end.

Thus, সঃ + অহম yields the form সো হম, where the mark ऽ indicates that the expression সোঽহম formed by *sandhi*, and the second word that has been combined here contains 'A' at the beginning. In the absence of *avagraha* such disjoining of the enphonic combination becomes difficult, and their meaning cannot be easily understood. Even during these days, thousands of copies of holy works like Raamaayana, Mahaabhaarata, Bhaagavad-Gita, Durgaasaptasati and other scripts dealing with rituals like marriage, funeral, worship of Shiva, Vishnu, Durgaa, Kaali etc. that are written in Sanskrit are printed in Bangla script. In the absence of *avagraha*, correct and dependable rendering of Sanskrit works in the Bangla script will be impossible.

Numerals

In Bangla compound numbers from 11 to 19 and the components involving them in higher numbers are pronounced in such a way that it is somehow impossible to understand where it belongs. For example, 11 is read as 'egaro', from this it is not possible to understand that this number is after 10 or 20. Another interesting fact is that the numbers from 21 to 99 are written from left to right but their number names are counted or read from right to left. However, after 100 or 1000 the digits at 100th or 1000th position are read first then the rest. Hundred is called 'shata', Thousand is called 'sahasra', Ten Thousand is called 'azut', Hundred Thousand is called 'lakh' or 'lakhkha', One Million/ Ten Lakhs is called 'nizut' and Ten Million is called 'koti'. The numerals in Bangla are as follows :

০	১	২	৩	৪	৫	৬	৭	৮	৯
0	1	2	3	4	5	6	7	8	9

Table VI : Bangla numeral set

In Unicode, numerals are ranged between U09E6 to U09EF.

Punctuation Marks

Modern Bangla uses punctuation marks, which are borrowed from English except the end-of-sentence marker. Old bangla books contain single and double vertical bars to indicate a fullstop, but the modern bangla only uses the single vertical bar '।'.

Ancient/Obsolete glyphs

U+09F2 to U+09F9 are a series of Bangla addition for writing currency and fractions. Among these, ৃ ৠ ৰ ৱ ৢ ৣ ৲ ৳ ৴ ৵ ৶ ৷ ৸ are the ancient Bangla glyphs frequently used on or before mid thirties of the last century. After the recommendation of the spelling reform committee of Calcutta University in 1936, the use of all these

glyphs becomes infrequent. Still we find their usage in some documents and if any one wants to get the digital copy of ancient documents one has to have these glyph supports in font files as well as in Unicode. Hence, these glyphs are not discarded from Unicode.

Character Statistics

Corpus based statistical analysis of any language is very useful in various applications including OCR, cryptography, linguistics, speech analysis and recognition, spelling error correction, electronic dictionary and machine aids to visually handicapped. The following are frequencies of characters in Bangla corpus provided by the MIT database, which consists of more than 3 million words (total Characters : 1,43,18,761) of running text covering a wide range of genre viz., modern fiction, short stories, newspapers and journals.

The global grapheme statistics of characters in the corpus are presented in the following Table.

Character	Percentage
Vowel	36.39
Consonant	63.61

Table VII : Character statistics summary table

The table shows that consonant percentage is much higher than vowel percentage and compound character percentage (7.34 %) is very small compared to other consonant and vowel. In commonly used language, there is a tendency of using words containing maximum number of consonants followed by vowel containing minimum number of compound character. The global character occurrence statistics would be useful for optical character recognition (OCR) development, spell-checker design and other problems.

The next table represents the consonants and vowels (vowels and their modified forms) according to their percentage of occurrence in the said corpus.

Char.	% of occur.	Char.	% of occur.	Char.	% of occur.
া	10.58635	ই	1.28592	্	0.30554
ে	9.12345	ব	1.27980	ফ	0.27616
র	9.07098	এ	1.16347	ড	0.26520
ি	5.79748	ী	1.15866	ঙ	0.25636
ন	5.28963	ছ	1.07006	ঠ	0.25339
ক	5.14978	আ	1.00475	ঘ	0.19234
ত	4.63765	চ	0.97830	ঞ	0.14776
ব	3.99412	ঐ	0.90798	ৈ	0.09725
এ	3.09293	থ	0.85059	ঝ	0.08548
ল	3.01622	ভ	0.84297	ী	0.08430
ম	2.96961	ধ	0.82305	ঃ	0.05097
প	2.70545	খ	0.82027	ঐ	0.03451
দ	2.29465	অ	0.76769	ঢ	0.02161
য়	2.25143	ও	0.72528	ঋ	0.00990
য	2.11691	ণ	0.70735	ঈ	0.00836
ু	2.05253	ং	0.46951	ঢ়	0.00666
ও	1.56969	ড়	0.45512	ঔ	0.00569
া	1.40914	উ	0.44692	ঊ	0.00376
ট	1.37966	ব	0.33243	অ্যা	0.00005
গ	1.34470	ঁ	0.32727		
জ	1.31026	ূ	0.31469		

Table VIII: Character-wise percentage of occurrence In the next Table, we show the most frequently used compound characters found in the corpus. The cluster প্র (PA+RA) is maximally used followed by ক্ষ (KA+SSA), ন্ত (NA+TA) and ত্র (TA+RA).

Comp. Char.	% of Occur.	Comp. Char.	% of Occur.
প্র	9.950	স্ব	1.768
ক্ষ	4.590	ত্য	1.638
ন্ত	4.213	ক্ত	1.618
ত্র	3.328	র্ত	1.536
ব্য	3.035	ক্র	1.534
ব	2.804	ত্ত	1.526
ন্দ	2.402	চ্ছ	1.521
স্ত	2.254	র‍্যা	1.449
হ্	2.251	স্প	1.413
ধ্য	2.184	দ্র	1.376
গ্র	1.921	দ্ধ	1.337
ষ্ট	1.892	ন্য	1.324

Table IX : Percentage of occurrence of some compound characters.

Positional Character Occurrence Statistics in Bangla is as follows : The percentages of occurrence of each consonant and vowel at each position of the words in the said corpus are tabulated in the Table IX. The character ক (KA) occurs in maximum percentage (9.98%) at the first position of the words. The occurrence of character ব (BA), স (SA) and প (PA) at the first position of words is quite high compared to other characters. The character র (RA) occurs in highest percentage at all but the first positions of the words.

Char.	Pos 1		Pos 2		Pos 3		Pos 4		Pos 5		Pos 6	
	% of occur.	Rank	% of occur.	Rank	% of occur.	Rank	% of occur.	Rank	% of occur.	Rank	% of occur.	Rank
অ	3.57	12	0.01	37	0.82	30	0.02	39	0.02	40	0.03	38
আ	4.69	7	0.01	36	1.05	24	0.03	38	0.03	38	0.04	37
ই	0.71	28	4.16	9	2.21	16	1.81	18	2.16	14	1.82	17
ঈ	0.04	37	0.00	40	0.01	43	0.00	44	0.00	43	0.00	43
উ	1.72	20	0.20	32	0.66	31	0.08	36	0.15	35	0.09	34
ঊ	0.02	40	0.00	46	0.00	46	0.00	52	0.00	53	0.00	46
ঋ	0.04	36	0.00	43	0.01	44	0.00	42	0.00	42	0.00	42
এ	5.39	5	0.02	35	1.14	23	0.10	34	0.05	37	0.09	35
ঐ	0.17	33	0.00	44	0.02	41	0.00	43	0.00	44	0.00	44
ও	1.58	21	0.31	30	1.04	25	0.98	25	0.89	24	1.01	25
ঔ	0.03	39	0.00	45	0.01	45	0.00	45	0.00	45	0.00	45
ক	9.98	1	6.69	6	7.32	3	6.96	5	5.15	5	6.69	5
খ	0.97	25	1.76	15	2.50	15	0.62	27	0.89	26	0.61	27
গ	2.31	15	1.21	20	1.92	19	2.60	13	1.65	18	1.93	16
ঘ	0.66	29	0.07	34	0.24	37	0.09	35	0.23	33	0.07	36
ঙ	0.00	41	1.63	18	0.29	36	0.35	29	0.35	31	0.22	32
চ	1.94	18	1.08	22	1.22	21	2.44	14	1.07	23	1.22	22
ছ	1.20	23	1.18	21	1.64	20	3.50	10	2.08	15	1.53	19
জ	2.43	14	0.89	26	2.62	12	1.77	19	1.64	19	1.77	18
ঝ	0.13	34	0.00	38	0.37	33	0.03	37	0.10	36	0.02	39
ঞ	0.00	45	0.48	28	0.15	39	0.21	33	0.45	30	0.24	31
ট	0.54	30	1.27	19	2.74	11	2.22	15	3.45	9	2.48	13
ঠ	0.21	32	1.01	23	0.34	34	0.29	31	0.25	32	0.33	30
ড	0.40	31	0.08	33	0.22	38	0.46	28	0.49	29	0.40	28
ঢ	0.10	35	0.00	39	0.02	42	0.00	41	0.00	41	0.00	41
ণ	0.00	42	0.83	27	0.42	32	1.67	20	1.71	17	2.75	10
ত	4.56	9	5.32	8	5.19	6	9.16	3	8.48	4	9.56	2
থ	1.23	22	1.74	16	0.96	28	1.54	21	1.13	22	1.29	20
দ	4.49	11	2.53	12	2.51	14	3.66	9	3.02	10	3.88	8
ধ	0.74	27	1.65	17	0.95	29	1.87	16	1.57	20	1.08	23
ন	4.79	6	7.54	3	8.75	2	7.95	4	9.95	2	9.02	3
প	7.72	4	2.60	11	2.88	10	3.03	11	2.63	13	2.51	12
ফ	0.83	26	0.30	31	0.32	35	0.25	32	0.21	34	0.19	33
ব	8.68	2	6.97	5	4.78	7	6.08	6	4.02	7	4.33	7
ভ	1.77	19	0.99	24	1.03	26	1.52	22	0.89	25	1.01	24
ম	4.55	10	7.18	4	3.88	9	4.48	8	4.01	8	3.53	9
য়	3.38	13	5.87	7	6.59	4	9.42	2	8.62	3	7.76	4
র	2.07	17	18.79	1	13.87	1	11.49	1	18.41	1	17.57	1
ল	1.18	24	7.72	2	6.51	5	4.77	7	4.71	6	5.12	6
ব	0.00	44	0.00	41	0.98	27	0.31	30	0.75	27	0.80	26
শ	2.12	16	0.90	25	2.57	13	1.32	23	1.72	16	2.39	14
ষ	0.03	38	0.32	29	2.03	18	1.83	17	2.72	11	2.21	15
স	8.42	3	2.94	10	4.06	8	2.90	12	2.64	12	2.72	11
হ	4.64	8	1.92	13	2.05	17	1.25	24	1.14	21	1.27	21
ড়	0.00	43	1.81	14	1.14	22	0.92	26	0.57	28	0.40	29
ঢ়	0.00	46	0.00	42	0.03	40	0.00	40	0.02	39	0.01	40

Table X : Positional Character Occurrence Statistics in Bangla

Fonts

Glyphs to be supported in Bangla Fonts

There are separate glyphs for numerals (১, ২, ৩, ৪, ৫, ৬, ৭, ৮, ৯, ০). Since the number of consonant cluster exceeds from the available spaces in a true type font file so some unique glyphs has been designed. For displaying all the clusters, combination of those glyphs are used. Different attempts to reform the Bangla script have been taken. All of them basically aimed at reducing the variations in the number of glyphs. Some attempts to reform the conjuncts as well so that a transparency will be observed in the conjunct clusters. As for example the conjuncts formed by clustering ক and ত changes its shape to ক্ত. One cannot understand from the cluster ক্ত that it is formed by ক and ত. In these types of cases the conjunct glyphs varies a lot from the basic glyphs. There are consonant graphic variants like RA-phalla, YA-phalla are placed after the conjunct, conjunct clusters. As for example ক্র this is formed by clustering ক, ত, র the graphic variant of র is ্র . Consonant variants ঁ, ঃ, ্, and ় are placed after consonants or consonant clusters তঁ, তঃ, ত্, র্ত.

Font encoding anomaly

Presently in Bangla there exists hundreds of font encoding systems. Most of the large publishing houses have their own fonts and editors. Others use different font encoding commercial software. These fonts have differences in glyph position and even some do not follow any standard encoding schema. For this reason data processing in Bangla from diversified documents is a difficult task.

Keyboard

Some points about the proposed Bangla keyboard are mentioned below.

1. This keyboard layout is based on the pair of letter/ symbol to be placed on a particular key, as decided by the computer subcommittee of the Bangla Academy.
2. Only two types of key operation will necessary—normal and shift
3. Symbols, such as +, -, [,], =, %, have been added.
4. 'O-Kars' and 'AU-Kars' will require single key strokes.
5. The normal practice of keeping a consonant and its aspirate on the same key has been retained.
6. ^ is to obtained by pressing ^- link- ^. This has been done to keep uniformity in writing the conjuncts.
7. This layout is more or less similar to the keyboard used by about 80% of the Bangla typesetters.
8. In designing the layout, proper stress has been given to ensure that during typing in both the left and right hands remain equally busy. Calculated on the basis of the frequencies of occurrences of the letters and vowel symbols, excluding the characters on the number-keys, punctuation marks, the link-key, the hand-use ratio of the right and the left is 1:0.899, and this is almost evenly balanced, the left hand being a bit more busy than the right If the link key (left hand) and the punctuation-keys (mostly right hand), are taken into account the ratio becomes 1:0.915. In actual situation it is likely to be very near 1:1 as the keys for hyphen, first brackets, equal to sign, quotations marks, enter, backspace, delete, insert and the entire numeric pad require right hand operation.

Presentation and Storage Considerations

As in Devanagari, the Bangla plain text memory representation generally follows phonetic order. That means for any syllable, the vowel qualifier (the dependent vowel sign) will always follow the consonant or conjunct. 'র্' (*Reph*) will come before the consonant or conjunct. While rendering, the glyph sequence may not follow the character sequence. Like for 'র্' (R*eph*), actual glyph may change its position to the end of the orthographic syllable. This process is more complex than Devanagari for two-part vowel signs. In those cases the vowel is split into pre-base and post-base components. Pre-base part is placed before the orthographic syllable to which it is applied. Post-base part is placed after the orthographic syllable.

The following provides more formal and detailed rules for minimal rendering of Bangla as part of plain text sequence. It describes the mapping between Unicode characters and the glyphs in a Bangla font. It also describes the combining and ordering of those glyphs.

These rules provide minimal requirements for legibly rendering interchangeable Bangla text. As

with any script, a more complex procedure can add rendering characteristics, depending on the font and application. For a particular sequence different fonts might apply different rules to get the final shape. As for example, if a ligature is found in the font for a particular sequence of code points, it might not require to create the conjunct using half-forms.

Notation

In the next set of rules, the following notation applies:

C_n	Nominal glyph form of consonant C as it appears in the code charts. Directly mapped to a code point.
C_l	A live consonant, depicted identically to C_n. Directly mapped to a code point.
C_d	Glyph depicting the dead consonant form of consonant C.
C_h	Glyph depicting the half form of consonant C.
C_{uh}	Glyph depicting the upper-half form of consonant C.
C_{lh}	Glyph depicting the non-spacing lower-half form of consonant C.
L_n	Nominal glyph form of a conjunct ligature consisting of two or more component consonants. A conjunct ligature composed of two consonants X and Y is also denoted $X.Y_n$.
RA_{sup}	A nonspacing combining mark glyph form of the U+09B0 BANGLA LETTER RA positioned above or attached to the upper part of a base glyph form. This form is also known as *reph*.
RA_{sub}	A nonspacing combining mark glyph form of the U+09b0 BANGLA LETTER RA positioned below or attached to the lower part of a base glyph form. This form is also known as *ra-phalaaa*.
V_{vs}	Glyph depicting the dependent vowel sign form of vowel V. Directly mapped to a code point.
V_{pre}	Glyph depicting the pre-base part of the two-part dependent vowel sign V. Directly mapped to a code-point.
V_{post}	Glyph depicting the post-base part of the two-part dependent vowel sign V. Directly mapped to a code-point.
$HALANT_n$	The nominal glyph form of the non-spacing combining mark depicting U+09CD HALANT. This character is not always depicted; when it is depicted, it adopts this non-spacing mark form.
$YAPHALAA_n$	The nominal glyph form of the U+09FB YA-PHALAA.

Table XI : Notations used for rule generation

Dead Consonant Rule

The following rule logically precedes the application of any other rule to form a dead consonant. Once formed, a dead consonant may be subject to other rules described next.

R1. When a consonant C_n precedes a $HALANT_n$, it is considered to be dead consonant C_d. A consonant C_n that does not precede $HALANT_n$ is considered to be live consonant C_l.

KA_n	+	$HALANT_n$	→	KA_d
ক	+	্	=	ক্

Conjunct Rules

Like Devanagari, Bangla has a large set of consonant conjunct forms. In Bangla, conjuncts could be formed using the following rules. The application of these rules varies from font to font.

R2. When a conjunct X_n precedes a $HALANT_n$, it is considered to be dead conjunct X_d. A conjunct X_n that does not precede $HALANT_n$ is considered to be live conjunct X_l.

$S.TA_n$	+	$HALANT_n$	→	$S.TA_d$
স্ত	+	্	=	স্ত্

Let, X_d, a dead consonant or conjunct precedes Y_l, a live consonant. Now a conjunct $X.Y_n$ will be formed according to the following rules.

R3. If a ligature L_n is available in the font for the combination X and Y, then X_d and Y_l are replaced by L_n.

KA_d + SSA_l → $K.SSA_n$
ক্ + ষ = ক্ষ

R4. This rule is applicable if R3 fails, i.e., no ligature is available for the combination X and Y. If X_h is available in the font, X_d is replaced by X_h.

$NA_d + MA_l$ → $NA_n + MA_l$ Displayed Output
ন্ + ম = ন্ + ম = ন্ম

R5. This rule is applicable if R4 fails, i.e., X_h is not available in the font. If X_{uh} and Y_{lh} are available in the font then X_d is replaced by X_{uh} and Y_l is replaced by Y_{lh}.

$NA_d + TA_l$ → $NA_{uh} + TA_{lh}$ Displayed Output
ন্ + ত = ন্ + ত = ন্ত

R6. This rule is applicable if R5 fails, i.e., X_{uh} or Y_{lh} is not available in the font. If Y_{lh} is available in the font then X_d is replaced by X_n and Y_l is replaced by Y_{lh}.

$SHA_d + LA_l$ → $SHA_n + LA_{lh}$ Displayed Output
শ্ + ল = শ + ◌্ল = শ্ল

R7. This rule is applicable if R6 fails, i.e., Y_{lh} is not available in the font. X_d will be depicted using the nominal form X_n and visible *halant* sign $HALANT_n$.

$KA_d + KHA_l$ → $KA_n + HALANT_n + KHA_l$ Displayed Output
ক্ + খ = ক + ◌্ + খ = ক্খ

Explicit Halant Rule

Normally, *halant* is used to create dead consonants from their nominal forms. However, for some words it is required to put an explicit *halant* instead of merging it in conjuncts. This could be achieved using the following rule.

R8. If the character, U+0930 ZERO WIDTH NON-JOINER comes between a dead consonant and a live consonant, a *halant* sign is depicted between the nominal form of both the consonants.

KA_d + ZWNJ + SSA_l → $KA_n + HALANT_n + SSA_n$
ক্ + ZWNJ + ষ = ক্ষ

YA-phalaa Rule

R9. If $YAPHALAA_n$ follows the nominal form of any consonant or conjunct X_n then this sequence (X_n $YAPHALAA_n$) collectively will behave as a conjunct (X $YAPHALAA_n$).

Consonant Ra Rule

The character U+09B0 RA takes one of a number of visual forms depending on its context in a consonant cluster. By default, this letter is depicted with its nominal glyph form (as shown in the code charts). In two contexts, it is depicted using a non-spacing glyph form that combines with a base letterform.

R10. If the dead consonant RA_d precedes a consonant and RA_d is the first consonant in the consonant cluster of the orthographic syllable, then it is replaced by superscript nonspacing mark RA_{sup}. If the number of consonants in the cluster is 2 then RA_{sup} will be placed after the other consonant of the cluster. If the number of consonants in the cluster is > 2, then a conjunct is formed applying R3 to R7 with the rest of the consonants and RA_{sup} is placed after the conjunct.

$RA_d + TA_l$ → $TA_n + RA_{sup}$ Displayed Output
র্ + ত = ত + ◌্র = র্ত

$RA_d + KA_d + SSA_l$ → $K.SSA_n + RA_{sup}$ Displayed output
র্ + ক্ + ষ = ক্ষ + ◌্র = র্ক্ষ

$RA_d + MA_d + PA_l$ → $MA_n + PA_n + RA_{sup}$ Displayed Output
র্ + ম্ + প = ম্ + প + ◌্র = ম্প্র

$RA_d + NA_d + TA_l$ → $NA_{uh} + TA_{lh} + RA_{sup}$ Displayed Output
র্ + ন্ + ত = ন্ + ◌্ত + ◌্র = ন্ত্র

R11. Except for the dead consonant RA_d, when a dead consonant C_d precedes the live consonant RA_l, C_d is replaced with its nominal form C_n, and RA is replaced by the subscript nonspacing mark RA_{sub}, which is positioned so that it applies to C_n.

$DHA_d + RA_l$ → $DHA_l + RA_{sub}$ Displayed Output
ধ্ + র = ধ + ◌্র = ধ্র

R12. For certain consonants, the mark RA_{sub} may graphically combine with the consonant to form a conjunct ligature form.

$KA_d + RA_l$ → $KA_l + RA_{sub}$ Displayed Output
ক্ + র = ক + ◌্র = ক্র

Vowel Sign Rules

R13. When the dependent vowel I_{vs} or E_{vs} or AI_{vs} is used to override the inherent vowel of a syllable, it is always written to the extreme left of the orthographic syllable.

$TA_d + RA_l + E_{vs} \rightarrow T.RA_n + E_{vs} \rightarrow E_{vs} + T.RA_n$

ত্ + র + ে = ত্র + ে = ত্রে

$TA_n + YAPHALAA_n + E_{vs} \rightarrow E_{vs} + TA_n + YAPHALAA_n$

ত + ্য + ে = ত্যে

R14. When the dependent vowel O_{vs} or AU_{vs} is used to override the inherent vowel of a syllable, it is always broken into pre-base and post-base parts as specified in the code-chart. Pre-base part is placed at the extreme left of the orthographic syllable and post-base part is placed at the extreme right of the orthographic syllable.

$KA_l + AU_{vs} \rightarrow AU_{pre} + KA_l + AU_{post}$

ক + ৌ = কৌ

$NA_d + TA_l + YAPHALAA_n + O_{vs} \rightarrow N.TA_n + YAPHALAA_n + O_{vs} \rightarrow O_{pre} + N.TAn + YAPHALAA_n + O_{post}$

ন্ + ত + ্য + ো = ন্ত + ্য + ো = ন্ত্যো

R15. In some cases, dependent vowel may combine with the consonant or conjunct resulting in a different shape of the combination.

$NA_d + TA_l + U_{vs} \rightarrow N.TA_n + U_{vs}$ Displayed Output

ন্ + ত + ু = ন্ত + ু = ন্তু

$SHA_d + U_{vs} \rightarrow$ Displayed Output

শ + ু = শু

HAd + VOCALICRvs → Displayed Output

হ + ৃ = হৃ

Collating Sequence and Sorting

The collation of units of textual information unambiguously has always been the source of contention. Generally, the sort order or the alphabetic sorting is the order of the position of characters in the alphabet. It is usually specific to a particular language. Though Indian languages agree in having structural (orthographic) similarity in the organization of characters in the alphabet they do differ in certain minor ways forcing different sort orders. Bangla follows the simple alphabetical order in collation where the vowel graphemes are used first followed by diphthongs. In case of consonants, a slightly complex collation process is followed which can be described as follows:

(1) A consonant without any vowel allograph is first considered.

(2) Next, it collates with other vowel graphemes.

(3) After this it collates with two nasal modifiers: *chandrabindu* and *\it anusvara*.

(4) Gradually, it collates with remaining vowel allographs in a very sequential manner. However, in each case, the two nasal modifiers will be first to collate.

(5) Finally, it follows possible combinations with other consonants, which will follow the sequence already mentioned in (4).

All major dictionaries in Bangla such as *Calantika* by Rajsekhar Basu, *Bangiya Sabdakosh* by Jnanendramohan Das, *Samsad Bangla Abhidhan* by Sailendranath Biswas etc., follow the character order mentioned above.

Localization of Data

Calendars

Most commonly used calendar in Bangla is the Western calendar (Gregorian calendar). There is also a native calendar called baNgAbda. It effectively counts from 594 A.D.

Month Names

The names of months (January, February, …, December) are borrowed from English. Months of the year in banGgAbda are: vaishAkha, jaiSTha, ASADh.a, shrAvaNa, bhAdra, Ashvina, kArttika, agrahAYaNa, pauSa, mAgha, phAlguna, caitra. These are unequal in length and are named after twelve of the naKSatras (stars), the months were originally named by a neighbouring nakSatra when the moon was full (the first listed naKSatra, kRttikA, corresponds to the seventh month because sun and moon are on opposite sides during the full moon).

Weekday Names

Seven days of the week in Bangla is given in the following Table.

Sunday	Rabibaar
Monday	Sombaar
Tuesday	Mangalbaar
Wednesday	Budhbaar
Thursday	Brihaspatibaar
Friday	Shukrabaar
Saturday	Shanibaar

Table XII : Days in the week in Bangla

Time

A day of 24 hours is divided into five divisions:

Morning	Sakaal
Noon	Dupur
Afternoon	Bikaal
Evening	Sandhya
Night	Ratri

Table XIII : Time-periods in a day in Bangla.

There is no equivalent term for AM/PM in Bangla.

Date

The date in Bangla is usually expressed in the order of date, month year. For example, March 20, 2002 may be expressed as ২০শে মার্চ ২০০২.

Weights and measurements

Earlier British units (FPS) was used along with some local units of measurements such as 'Poa' (for liquid mass), 'sher' (for solid mass) etc. Now-a-days metric units are universally accepted.

Currency

There is no standard currency symbol in Bangla. However, ৳ is often used to indicate Rupees. For example, Rs 10 may be written as ১০৳.

(Courtesy : Prof. B.B Chaudhary (CI) Head,
Computer Vision and Pattern Recognition Unit
Indian Statistical Institute, Kolkata –700035
Tel. 033-5778085, 5777694, 5775502,
E-mail : bbc@isical.ac.in)

9.1.3 Typical Colloquial Sentences in Bangla

GREETING

- Hello
  হ্যালো
  ह्यालो
  Hello
- Good Morning
  সুপ্রভাত
  सुप्रभात
  Suprobhaat
- Good Afternoon
  শুভ অপরাহ্ন
  शुभ अपराह्न
  Subho Aparaanho
- Good Night
  শুভ রাত্রি
  शुभ रात्रि
  Subho Raatri
- Good Bye
  বিদায়
  विदाय
  Bidaay
- Thanks
  ধন্যবাদ
  धन्यबाद
  Dhanyabaad
- How are you
  কেমন আছেন?
  केमन आछेन?
  Kemon Achen
- I am fine thank you
  আমি ভাল আছি, ধন্যবাদ
  आमि भाल आछि, धन्यबाद
  Aami Bhaalo Aachi, Dhanyabaad
- Sorry
  দুঃখিত
  दुःखित
  Dukhita

WEATHER

- It is cold
  এখন ঠাণ্ডা
  एखन ठाण्डा
  Ekhon Thaanda
- It is cool outside
  বাইরে ঠাণ্ডা পড়েছে
  बाइरे ठाण्डा पड़ेछे
  Baire Thaanda Poreche
- It is hot
  এখন গরম
  एखन गरम
  Ekhon Garam
- It is raining
  বৃষ্টি পড়ছে
  बृष्टि पड़छे
  Bristi Porche

GENERAL

- What is your name?
  তোমার নাম কি?
  तोमार नाम कि?
  Tomaar Naam ki
- My name is Ranjan
  আমার নাম রঞ্জন
  आमार नाम रञ्जन
  Aamar Naam Ranjan
- Where do you live?
  তুমি কোথায় থাক?
  तुमि कोथाय थाक?
  Tumi Kothaay Thaako
- I live near Ghantaghar
  আমি ঘণ্টাঘরের কাছে থাকি
  आमि घण्टाघरेर काछे थाकि
  Aami Ghontaaghorer Kaache Thaaki
- How old are you?
  তোমার বয়স কত?
  तोमार बयस कत?
  Tomaar Boyos kato
- That building is tall
  বাড়িটা উঁচু
  बाड़िटा उचुँ
  Baarita Uchu
- She is beautiful
  সে সুন্দরী
  से सुन्दरी
  Se Sundari
- I like Bengali sweets
  আমি বাংলার মিষ্টি পছন্দ করি
  आमि बांलार मिष्टि पछन्द करि
  Aami Banglaar Misti Pachanda Kori
- I love birds
  আমি পাখি ভালোবাসি
  आमि पाखि भालोबासि
  Aami Paakhi Bhaalobaasi
- Where is Railway station?
  রেলওয়ে স্টেশন কোথায়?
  रेलओये स्टेशन कोथाय?
  Railway Station Kothaay?
- How far is the Bus Terminal from here?
  এখান থেকে বাস টার্মিনাল কত দূর?
  एखान थेके बास टार्मिनाल कत दूर?
  Ekhan Theke Baas Terminal Kato Dur?
- How long will it take to reach the Airport?
  বিমান বন্দর পৌঁছাতে কত সময় লাগে?
  बिमान बन्दर पौंछाते कत समय लागे?
  Bimaan Bondar Pouchaate Kato Samay Laage?
- Is Mr. Raghunath there?
  মিঃ রঘুনাথ কি সেখানে?
  मिः रघुनाथ कि सेखाने?
  Mr. Ragunath Ki Sekhaane?
- Please tell him to call back as soon as he is free
  কাজ সারা হলেই তাকে ফোন করতে বলো
  काज सारा हलेइ ताके फोन करते बलो
  Kaaj Saaraa Holei Taake Phone Korte Balo
- How much will it cost?
  এটার দাম কত?
  एटार दाम कत?
  Etaar Daam Kato?

- Excuse me
  মাপ করবেন
  মাপ করবেন
  Maap Korben
- From which Platform can I get the train for Chandigarh?
  কোন্ প্লাটফর্ম থেকে চণ্ডীগড়ের গাড়ী পাবো?
  কোন্ প্লাটফর্ম থেকে চণ্ডীগড়ের গাড়ী পাবো?
  Kon Platform Theke Chandigarh-er Gaari Paabo?
- Does this train stop at Aligarh?
  এই ট্রেন কি আলিগড়ে দাঁড়াবে?
  এই ট্রেন কি আলিগড়ে দাঁড়াবে?
  Ei Train Ki Aligarh-re Daaraabe?
- How many kids do you have?
  তোমার কয়টি সন্তান?
  তোমার কয়টি সন্তান?
  Tomaar Kaiti Santaan?
- This gift is wonderful
  উপহারটি অপূর্ব
  উপহারটি অপূর্ব
  Upahaarti Apurba
- It is really pretty
  এটা সত্যিই সুন্দর
  এটা সত্যিই সুন্দর
  Etaa Sattie Sundar
- Food is delicious
  খাবারটা সুস্বাদু
  খাবারটা সুস্বাদু
  Khaabaartaa Sussadu
- Congratulations
  অভিনন্দন
  অভিনন্দন
  Abhinandan
- You look lovely
  তোমাকে সুন্দর দেখাচ্ছে
  তোমাকে সুন্দর দেখাচ্ছে
  Tomaake Sundar Dekhaache
- Wish you happy new year
  তোমাকে শুভ নববর্ষের অভিনন্দন জানাই
  তোমাকে শুভ নববর্ষের অভিনন্দন জানাই
  Tomaake Subho Nobabarsh-er Abhinandan Jaanaai
- I wish you all the happiness
  তোমার সকল সুখ কামনা করি
  তোমার সকল সুখ কামনা করি
  Tomaar Sakal Sukh Kamonaa Kori
- Congratulations on your marriage
  তোমাকে বিয়ের অভিনন্দন
  তোমাকে বিয়ের অভিনন্দন
  Tomaake Biyer Abhinandan
- Keep your eyes wide open before marriage and half- shut afterwards
  বিয়ের আগে চোখ দুটো খুলে রাখো - বিয়ের পরে রাখো আধবোজা
  বিয়ের আগে চোখ দুটো খুলে রাখো - বিয়ের পরে রাখো আধবোজা
  Biyer Aaage Chokh Duto Khule Raakho - Biyer Pore Raakho Aadhbojaa

(Courtesy : Prof. B.B Chaudhary (CI) Head,
Computer Vision and Pattern Recognition Unit
Indian Statistical Institute, Kolkata –700035
Tel. 033-5778085, 5777694, 5775502,
E-mail : bbc@isical.ac.in)