3. Bengali

Bengali (ethnonym: Bangla) language is categorized within the Bengali-Assamese branch of Eastern Zone of Indo Aryan languages. It is spoken by more than 200 million people across the world out of which about 100 million speakers reside in Bangladesh and 70 million speakers reside in India [12]. Bengali is the national language of Bangladesh while it is also the state language of the Indian state of West Bengal.

Bengali is an Indic language which uses Bengali script, closely related to Devanagri script, both deriving from Brahmi script. Bengali script is also used to write other languages, including Assamese, Daphla, Garo, Hallam, Khasi, Manipuri, Mizo, Munda, Naga, Rian and Santali [4, 13].

3.1. Writing System

3.1.1. Character Set

Bengali character set is divided into 21 vowels, 36 consonants and modifiers [15]. The vowels themselves can be divided into dependent and independent vowels, shown in Figure 3.1 below.

অ আ ই ঈ উ ঊ ঋ এ ঐ ও ঔ

Independent Vowels

া ি ী ু ূ ৃ ে ৈ ো ৌ

Dependent Vowels

Figure 3.1. Bengali Vowels

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 ড় ঢ় য় ৎ

Figure 3.2. Bengali Consonants

Along with consonants and vowels there are some special modifiers, called Virama, Visarga, Anusvara, Candrabindu and Ishar, shown in Table 3.1. Anusvara is used for final velar nasal

sound, Visarga adds voiceless breath after vowel and Candrabindu is used to nasalize vowels [13, 14]. Virama, also called Halanta is discussed in the next section.

Table 3.1. Bengali Special Characters

Name	Glyph	Usage with a Consonant 'k'
Virama	◌্	ক্
Candrabindu	◌ঁ	কঁ
Anusvara	◌ং	কং
Visarga	◌ঃ	কঃ

Bengali also has its own numerals, shown in Figure 3.3.

০ ১ ২ ৩ ৪ ৫ ৬ ৭ ৮ ৯

Figure 3.3. Bengali Numerals

There are some additional characters for punctuation, etc. in the Bengali character set, which are ignorable for collation. The complete encoded character set in Unicode is given in [4].

3.1.2. Script Details

3.1.2.1. Consonants and Vowels

Bengali is written from left to right. Space is used to mark word boundaries. Letters are uncased and are grouped together based on place and manner of articulation. The characters in Bengali script hang from a horizontal line, called the head stroke. When writing, these characters within a word head strokes merge to form single base line, as shown for the word BAABAA (father) in Figure 3.4.

ব (Letter Ba) + া (Vowel AA) + ব (Letter Ba) + া (Vowel AA) = বাবা

Figure 3.4. Merging of Head Strokes of Bengali Characters

The consonants in Bengali have an inherent [ɔ][1] sound by default. For example Bengali letter ক represents [kɔ] and not [k] sound. Virama is placed below delete the vowel sound and get the pure consonantal sound. However, the use of Virama is often implied and optionally written by Bangla speakers.

The vowels take the independent vowel shape if they are in a syllable without an onset consonant. In case they are in a syllable with an onset consonant, they attach with the consonant taking the dependent shape. Thus, all vowels have an independent and dependent shape, except the vowel [ɔ] which only has an independent shape অ. It does not have a dependent shape as it is inherently present with each consonant by default if not explicity deleted by Virama or over-ridden by another dependent vowel. The dependent vowels attach at the front, back, top or bottom of a consonant. These are illustrated in Table 3.2. In some cases the vowel splits into two halves and is placed across consonant such that one half is at right while other is at left.

Table 3.2. Dependent Vowels with Consonant [k]

Consonant + Dependent Vowel	Joined Form	Comment
ক+ি	কি	Connects to left of consonant
ক+ী	কী	Connects to right of consonant
ক+ু	কু	Connects to base of consonant
ক+ৌ	কৌ	Wraps around the consonant

As is shown in Table 3.2, the vowel is typed after the consonant no matter where it attaches. Also, only one vowel can connect to a consonant at a time. The dependent vowels can not occur with independent vowels or by themselves.

3.1.2.2. Conjunct Consonants

In Bengali two or more consonants may join together to form complex conjuncts with alternate shapes. In Unicode, Virama is placed on the first consonant in a pair to enforce the conjoined shape of the consonants [4]. Some conjuncts and non-conjunct shapes are given in Table 3.3.

[1] Square brackets [ ] are conventionally used to represent a phone or a sound.

Like other Indic languages, র (or /r/) also forms different shapes in consonant clusters. When in initial position it is displayed as a mark to top, and when at the end it appears as a wavy line below the consonant to which it connects [4], as shown in last two rows of Table 3.3.

Table 3.3. Conjunct Consonants

Consonants C1 C2	Clustered Form C1 + C2	Conjunct Form C1 + ্ + C2
ক ষ	কষ	ক্ষ
স ক	সক	স্ক
ব দ	বদ	ব্দ
র য	রয	র্য
ব র	বর	ব্র

A more comprehensive list of conjunct consonants can be viewed at [14].

3.2. Collation

Bengali collation sequence has been defined by Bangla Academy, the language authority of Bangladesh. This section elaborates on this collation sequence for Bengali and an algorithmic implementation using UCA [2] for Bengali collation.

In Bengali all characters have primary level significance for collation purposes. Numerals and currency symbols are given smallest weight; these are followed by independent vowels, modifiers, consonants and dependent vowels. However, before collation is applied some text processing is required. Details of the text processing are also presented.

3.2.1. Text Processing

3.2.1.1. Reordering

As mentioned above the independent vowels combine with consonants in different manners i.e. joining to right, left, above or below. In hand-written orthography, old type-writers and non-standard Bengali encodings, the vowels that attach to the left are written first followed by a consonant. Others are written after the consonant. Thus, the typing order for কে is ে + ক and for

কী is ক + ী. For collation, the logical comparison order is consonant and then the dependent vowel, wherever the vowels attaches to the consonant. The typing sequence just discussed is inconsistent and thus the logical comparison between two combinations is not possible. Thus, the preceding vowel needs to be re-ordered, after the consonant, if a comparison has to be enabled. This is true for all the encodings which require dependent vowels to be typed before the consonant. However, the Unicode standard for Bengali requires consonant + vowel typing order whether the vowel visually appears after or before the consonant. The visual placement is separately handled in the rendering process. Therefore, if the Unicode encoding is followed, no reordering is required.

3.2.1.2. Normalization

There are different ways some Bengali characters, both consonants and vowels, can be encoded in Unicode. Thus, normalization is required before collation can be done. As discussed during the general discussion on collation in the second chapter, both composed or decomposed forms may be taken to do the collation, as long as it is consistently done. This section lists some of the equivalent forms for Bengali.

The first set of equivalents in Bengali are formed due encoding of Nukta as a combining character ◌়. (U+09BC). Nukta combines with consonants to give additional consonants, which are also separately encoded. Examples are given in Table 3.4. below.

Table 3.4. Normalization Due to Nukta in Bengali [4]

Decomopsed Form	Unicodes of Decomposed Form	Equivalent Composed Form	Unicode of Composed Form
ড ◌়	09A1 09BC	ড়	09DC
ঢ ◌়	09A2 09BC	ঢ়	09DD
য ◌়	09AF 09BC	য়	09DF

Similarly, dependent vowels which have two parts and surround the consonant also have equivalent encodings, equivalent to the case where a single vowel is split into the parts which come before and after the consonant respectively. The equivalents are given in Table 3.5. As can be seen in the table, both forms render in the same way when combined with a consonant are equivalent in terms of collation.

Table 3.5. Normalization Due to Glyph Splitting of Two-Part Dependent Vowels

Decomposed Form	Unicodes of Decomposed Form	Use with a Consonant	Equivalent Composed Form	Unicode of Composed Form	Use with a Consonant
ে া	09CB 09BE	ক ে া = কো	ো	09CB	ক ো = কো
ে ৗ	09C7 09D7	ক ে ৗ = কৌ	ৌ	09CC	ক ৌ = কৌ

One can form half shape of consonants in Indic scripts. Unicode enables that by typing Virama after the consonant. In a special case, Bengali conjunct character 'tta' can be encoded in multiple ways, but must show the same behavior for collation. Thus, the variations must be normalized to represent the same collation weight.

Table 3.6. Encoding and Rendering Variations of 'tta' Conjunct with Khanda Ta Character

Constituent Characters	Unicode Sequence	Rendered Variant Form
ত ্ ত	09A4 09CD 09A4	ত্ত
ত ্ ZWJ ত	09A4 09CD 200D 09A4	ৎত
ত ্ ZWNJ ত	09A4 09CD 200C 09A4	ত্‌ত
ৎ ত	09CE 09A4	ৎত

The normalization with Khanda Ta is different from the first two cases discussed because the final conjunct form is not encoded. Thus, the sequence can only be equated in decomposed forms and cannot be mapped onto a single composed form.

3.2.1.3. Contraction

In case the encoding is being translated into decomposed form, contraction is needed for assigning the collation elements, i.e. multiple character codes would map onto a single collation element. This contraction for consonants and vowels, presented in Tables 3.4 and 3.5, is illustrated in Table 3.7.

Table 3.7. Contraction to Single Collation Element from Multiple Encoded Characters

Glyph	Unicodes of Decomposed Form	Unicode of Composed Form	Collation Element	Unicode Name
ড়	09A1 09BC	09DC	15BD 0020 0002	LETTER RRA
ঢ়	09A2 09BC	09DD	15BF 0020 0002	LETTER RHA
য়	09AF 09BC	09DF	15CC 0020 0002	LETTER YYA
ৌ	09C7 09D7	09CC	15E3 0020 0002	VOWEL SIGN AU
ো	09C7 09BE	09CB	15E2 0020 0002	VOWEL SIGN O

3.2.1.4. Conjunct Consonants

The formation of alternate glyphs for conjuncts does not change input sequence logically but only visually. Collation is dependent on the logical sequence and thus is not affected by the change in shape. The Zero Width Joiner and Zero Width Non-Joiner are ignored in the process. However, ambiguity occurs in case of the combination of Ra and Ya, where Zero Width Non-Joiner plays a significant role. See [26] for further details.

3.2.2. Collation Elements

In order to realize Bengali collation as defined by Bangla Academy [15], following collation element table may be used. The table gives multiple entries in relevant columns if required. The table is further divided into sub-sections for various families of characters, including signs, numerals, dependent vowels, characters and dependent vowels.

Table 3.8. Collation Elements for Bengali Language

Glyph	Unicode	Collation Elements	Unicode Name
← Various Signs →			

◌়	09BC	13A0 0020 0002	BENGALI SIGN NUKTA
◌ং	0982	13A2 0020 0002	BENGALI SIGN ANUSVARA
◌ঃ	0983	13A3 0020 0002	BENGALI SIGN VISARGA
◌ঁ	0981	13A4 0020 0002	BENGALI SIGN CANDRABINDU
← Numerals & Currency Symbols →			
৺	09F8	0DC7 0020 0002	BENGALI CURRENCY NUMERATOR ONE LESS THAN THE DENOMINATOR
৹	09F9	0DC8 0020 0002	BENGALI CURRENCY DENOMINATOR SIXTEEN
ঀ	09FA	0350 0020 0002	BENGALI ISHAR
৲	09F2	0E12 0020 0002	BENGALI RUPEE MARK
৳	09F3	0E13 0020 0002	BENGALI RUPEE SIGN
০	09E6	0E29 0020 0002	BENGALI DIGIT ZERO
১	09E7	0E2A 0020 0002	BENGALI DIGIT ONE
৴	09F4	0E2A 0020 0002	BENGALI CURRENCY NUMERATOR ONE
২	09E8	0E2B 0020 0002	BENGALI DIGIT TWO
৵	09F5	0E2B 0020 0002	BENGALI CURRENCY NUMERATOR TWO
৩	09E9	0E2C 0020 0002	BENGALI DIGIT THREE
৶	09F6	0E2C 0020 0002	BENGALI CURRENCY NUMERATOR THREE
৪	09EA	0E2D 0020 0002	BENGALI DIGIT FOUR
৷	09F7	0E2D 0020 0002	BENGALI CURRENCY NUMERATOR FOUR
৫	09EB	0E2E 0020 0002	BENGALI DIGIT FIVE

৬	09EC	0E2F 0020 0002	BENGALI DIGIT SIX
৭	09ED	0E30 0020 0002	BENGALI DIGIT SEVEN
৮	09EE	0E31 0020 0002	BENGALI DIGIT EIGHT
৯	09EF	0E32 0020 0002	BENGALI DIGIT NINE
← Independent Vowels →			
অ	0985	12A2 0020 0002	BENGALI LETTER A
আ	0986	12A3 0020 0002	BENGALI LETTER AA
ই	0987	12A4 0020 0002	BENGALI LETTER I
ঈ	0988	12A5 0020 0002	BENGALI LETTER II
উ	0989	12A6 0020 0002	BENGALI LETTER U
ঊ	098A	12A7 0020 0002	BENGALI LETTER UU
ঋ	098B	12A8 0020 0002	BENGALI LETTER VOCALIC R
ৠ	09E0	12A9 0020 0002	BENGALI LETTER VOCALIC RR
ঌ	098C	12AA 0020 0002	BENGALI LETTER VOCALIC L
ৡ	09E1	12AB 0020 0002	BENGALI LETTER VOCALIC LL
এ	098F	12AC 0020 0002	BENGALI LETTER E
ঐ	0990	12AD 0020 0002	BENGALI LETTER AI
ও	0993	12AE 0020 0002	BENGALI LETTER O
ঔ	0994	12AF 0020 0002	BENGALI LETTER AU
← Consonants →			
ক	0995	15B0 0020 0002	BENGALI LETTER KA

খ	0996	15B1 0020 0002	BENGALI LETTER KHA
গ	0997	15B2 0020 0002	BENGALI LETTER GA
ঘ	0998	15B3 0020 0002	BENGALI LETTER GHA
ঙ	0999	15B4 0020 0002	BENGALI LETTER NGA
চ	099A	15B5 0020 0002	BENGALI LETTER CA
ছ	099B	15B6 0020 0002	BENGALI LETTER CHA
জ	099C	15B7 0020 0002	BENGALI LETTER JA
ঝ	099D	15B8 0020 0002	BENGALI LETTER JHA
ঞ	099E	15B9 0020 0002	BENGALI LETTER NYA
ট	099F	15BA 0020 0002	BENGALI LETTER TTA
ঠ	09A0	15BB 0020 0002	BENGALI LETTER TTHA
ড	09A1	15BC 0020 0002	BENGALI LETTER DDA
ড়	09DC	15BD 0020 0002	BENGALI LETTER RRA
ড ◌়	09A1 09BC	15BD 0020 0002	BENGALI LETTER RRA
ঢ	09A2	15BE 0020 0002	BENGALI LETTER DDHA
ঢ়	09DD	15BF 0020 0002	BENGALI LETTER RHA
ঢ ◌়	09A2 09BC	15BF 0020 0002	BENGALI LETTER RHA
ণ	09A3	15C0 0020 0002	BENGALI LETTER NNA
ত	09A4	15C1 0020 0002	BENGALI LETTER TA
থ	09A5	15C2 0020 0002	BENGALI LETTER THA
দ	09A6	15C3 0020 0002	BENGALI LETTER DA

ধ	09A7	15C4 0020 0002	BENGALI LETTER DHA
ন	09A8	15C5 0020 0002	BENGALI LETTER NA
প	09AA	15C6 0020 0002	BENGALI LETTER PA
ফ	09AB	15C7 0020 0002	BENGALI LETTER PHA
ব	09AC	15C8 0020 0002	BENGALI LETTER BA
ভ	09AD	15C9 0020 0002	BENGALI LETTER BHA
ম	09AE	15CA 0020 0002	BENGALI LETTER MA
য	09AF	15CB 0020 0002	BENGALI LETTER YA
য়	09DF	15CC 0020 0002	BENGALI LETTER YYA
য ◌়	09AF 09BC	15CC 0020 0002	BENGALI LETTER YYA
র	09B0	15CD 0020 0002	BENGALI LETTER RA
ৰ	09F0	15CE 0020 0002	BENGALI LETTER RA WITH MIDLE DIAGONAL
ল	09B2	15CF 0020 0002	BENGALI LETTER LA
ৱ	09F1	15D0 0020 0002	BENGALI LETTER RA WITH LOWER DIAGONAL
শ	09B6	15D1 0020 0002	BENGALI LETTER SHA
ষ	09B7	15D2 0020 0002	BENGALI LETTER SSA
স	09B8	15D3 0020 0002	BENGALI LETTER SA
হ	09B9	15D4 0020 0002	BENGALI LETTER HA
ঽ	09BD	15D5 0020 0002	BENGALI SIGN AVAGRAHA
ৎ	09CE	[15C1 0020 0002],[15E4 0020 0002]	BENGALI LETTER KHANDA TA

← Dependant Vowels →			
া	09BE	15D6 0020 0002	BENGALI VOWEL SIGN AA
ি	09BF	15D7 0020 0002	BENGAL VOWEL SIGN I
ী	09C0	15D8 0020 0002	BENGAL VOWEL SIGN II
ু	09C1	15D9 0020 0002	BENGAL VOWEL SIGN U
ূ	09C2	15DA 0020 0002	BENGAL VOWEL SIGN UU
ৃ	09C3	15DB 0020 0002	BENGAL VOWEL SIGN VOCALIC R
ৄ	09C4	15DC 0020 0002	BENGAL VOWEL SIGN VOCALIC RR
ৢ	09E2	15DD 0020 0002	BENGAL VOWEL SIGN VOCALIC L
ৣ	09E3	15DF 0020 0002	BENGAL VOWEL SIGN VOCALIC LL
ে	09C7	15E0 0020 0002	BENGAL VOWEL SIGN E
ৈ	09C8	15E1 0020 0002	BENGAL VOWEL SIGN AI
ো	09CB	15E2 0020 0002	BENGAL VOWEL SIGN O
ে া	09C7 09BE	15E2 0020 0002	BENGAL VOWEL SIGN O
ৌ	09CC	15E3 0020 0002	BENGAL VOWEL SIGN AU
ে ৗ	09C7 09D7	15E3 0020 0002	BENGAL VOWEL SIGN AU
্	09CD	15E4 0020 0002	BENGALI SIGN VIRMA
ৗ	09D7	15E5 0020 0002	BENGALI AU LENGTH MARK

3.2.3. Results

Table 3.9 shows output obtained by sorting a sample input using the collation elements given in Table 3.8.

Table 3.9. Input and Corresponding Sorted Output for Bengali

Input		Output	
ঋতু	এ১	অংশ	এও
কোল৪	ঔৎকষ	অংশাংশ	এঃ
ইঃ	ক১	অংশাংশি	এঁঠড়
অকথিত	কই১	অংশানো	ঐ১
কৌচ	এঁঠড়	অংশী	ও২
অকুঠ	অংশ	অংশে	ওঁ
ইউনিফম	কওম	অকথিত	ঔৎকষ
ইংকার	কতক	অকুঠ	ঔৎসুক্
অংশী	ওঁ	ইউনানি	ক১
ইঁচড়	কেল	ইউনিফম	ক৪
উওল	অংশাংশ	ইংকার	কই১
উদার	কোল১	ইঃ	কওম
ঊঢ়	অংশে	ইঁচড়	কওলা
এও	ক৪	উওল	কত
ঋক্	ঔৎসুক্	উদার	কতক
অংশাংশি	কতকটা	ঊঢ়	কতকটা
অংশাংনো	কত	ঋক্	কেল
এ২	ঐ১	ঋতু	কোল১

এইতো	কওলা	এ১	কোল৪
ইউনানি	ও২	এ২	কৌচ
এঃ	কৌচ	এইতো	কৌচ

3.3. Conclusion

Bengali, like other Indic languages, has single level of collation. All characters are sorted at primary level with numerals and currency symbols, independent vowels, modifiers, consonants and dependent vowels sorted in this order. The sorting requires some text processing to decompose the characters and map multiple characters onto single collation elements. However after the mapping, the collation algorithm discussed in the second chapter is applied in a regular manner for eventual collation.