Proposal for a Bengali Script Root Zone Label Generation Ruleset (LGR)

*LGR Version:* 3.0

*Current Date:* 2018-06-26

*Document version:* 3.10

*Authors:* Neo-Brahmi Generation Panel [NBGP]

# General Information

This document lays down the Label Generation Rule Set (LGR) for the Bengali[[1]](#footnote-1) (or ‘Bangla’) script under the general rubric of the Neo-Brahmi Writing System. Three main components of the Bengali Script LGR i.e. Code point repertoire, Variants and Whole Label Evaluation Rules have been described in detail here, having given the historical background of the Script under Section 3.

All these components will be incorporated in a machine-readable format in an XML file named "proposal-lgr-beng-20180626.xml". Labels for testing can be found in the accompanying text document “test-label-beng-20180526.txt”.

# Script for Which the LGR Is Proposed

ISO 15924 Code: Beng

ISO 15924 Key N°: 325

ISO 15924 English Name: Bengali (Bangla)

Latin transliteration of native script name: bɑːŋlɑː

Native name of the script: বাংলা

Maximal Starting Repertoire (MSR) version: MSR-3

# Background on Script & Principal Languages Using It

## 3.0. Introduction

Bengali (or ‘Bangla’) is an eastern Indo-Aryan language with around 178.2 million speakers in Bangladesh, and 83.4 million speakers in the Indian states of West Bengal, Tripura and South Assam as well as in the Andaman and Nicobar Islands. It is a major language in Jharkhand, too and a language with a sizable population in Bihar. Apart from these there are a huge number of Bengali diasporas spread all over the world. It is the seventh largest spoken and written language in the world. Bengali is the national and official language of Bangladesh, and one of the 22 Official languages in India. The script is also called Bengali [102] which is an eastern variety of the ‘Brāhmī-Devanāgarī’ Writing System, written from left to right. Historically it derives from the Brāhmī alphabet of the Ashokan inscriptions (269-232 BC).

In order to understand the genesis of Bengali (or Bangla) language, one could consider Suniti Kumar Chatterji’s [103, pp 16] suggestion of dividing the Indo-Aryan Speech into three broad periods considering the main phonetic changes and morphological trends. They are as follows:

(i) The *Old Indo Aryan* (OIA), when the language was most copious in both its sound and forms. The OIA period begins from the composition of the Vedic Hymns, i.e. from around 1500/1200 B.C. to the 557-477 B.C., the time immediately preceding Gautam Buddha.

(ii) The *Middle Indo-Aryan* (MIA), when there was a movement towards simplification of older consonant groups, and a general curtailment or simplification of grammatical forms. The MIA period (600 BC-1000 AD) is further subdivided into an early, a second and a late stage, with a transitional stage between the early and the second stage. The early stage is attested by inscriptional ‘Prakrit’ and ‘Pali’, the second MIA stage by literary Prakrits, and the late MIA stage by ‘Apabhraṅśa’ and ‘Avahaṭṭha’.

(iii) It is the third stage that branches off into the *New Indo-Aryan* (NIA) languages, starting from roughly 1000 AD, when the total character of the language was altered, and the vernaculars of modern Indo Aryan began to spring up. Bengali is said to have evolved from ‘Māgadhī Apabhraṅśa-Avahaṭṭha’ along with Assamese (or Asamiyā), Odia, Magahi, Maithili, and Bhojpuri. Bengali belongs to the earlier group of the Magadhan sub-family along with Assamese and Odia [104].

Bengali and its cognate languages, as mentioned above, together form a linguistic group known as the Eastern New Indo-Aryan (NIA). There is a gross inadequacy of the inscriptions and manuscripts in the Eastern Apabhraṅśa or ‘Avahaṭṭha’ except for small inscriptions and the manuscripts of the Tantric Buddhist text titled ‘Caryyācaryyaviniścaya’ or the Caryā-Pada [114] dating back to the 9th-11th century. As a result, there is not much epigraphic evidence for the development of its writing system. However, what evidence is available of the genesis of Bengali writing system is discussed in the section 3.1 [109].

Historically, the Bengali language has been divided into three periods as evident from various sources:

(i) Firstly, the Old Bengali Period (roughly 950/1000 to A.D.1200/1350) when three specimens are found: (a) 47 Caryā songs composed by the Sahajiya Buddhists (Cf. Shastri 1916) - the *Caryyācaryyaviniścaya* , the *Dohākōṣa* of Saraha and the *Dohākōṣa* of Kānha (mostly in Apabhraṅśa), and the *Ḍākārṇava* (in a variety of Prakrit), (b) Old Bengali specimens of over 300 words in a commentary on the *Amara-kōśa* dated 1159 AD, and finally (c) the *Rāma Carita* of Sandyakara Nandi (1084-1155 AD), attesting some place names [141].

(ii) Then there is the Middle Bengali Period - 1200-1800 AD, again divided into three stages: (a) Transitional Middle Bengali (1200-1300 A.D): No genuine specimens are found, except for the legends of Gopīcanda, Behula-Lakhindar, Khullana-Dhanapati, Phullara-Kālaketu, Lausena, etc. which were dealt with in great poems in subsequent centuries (e.g. *Gopīcandrer Gāna* by an anonymous poet, *Gopīcandrer Pā̃cāli* by Bhavani Das, *Gopīcandrer Sanyās* by Sakur Mamud, *Caṇḍīmaṅgalas* of Manik Datta/Dvija Madhav/Kvikankana Mukunda, etc., *Manasāmaṅgalas* of Vipradas Piplai/Vijay Gupta, etc.) [149], (b) Early Middle Bengali (1300- 1500 A.D) with classics such as the *Śrī-Kṛṣṇa-kīrttana* of Caṇḍidāsa (born 1408 AD) or various Bengali translations of *Bhāgavata*, the *Ramayana* (e.g. Bengali *Rāmāyaṇa* of Kṛttivāsa Ojhā ,1381-1461) and the *Mahabharata* (Bengali translation by Parameshwar Das aka *Parāgalī Mahābhārata)*. (c) Late Middle Bengali (1500-1800 A.D) which is attested by the development of Vaisnava literature under the influence of Śrī Chaitanya Deva (1486-1534 AD) and his disciples.

(iii) Finally, after 1800 AD, we find the Modern or New Bengali, marked by the introduction of written prose [109] in the books of Fort William College (established in 1800; books published—*Rājā Pratāpāditya Caritra* by Ramram Basu, *Batriś Siṅhāsana* by Mrityunjaya Vidyalankara, etc.) Christian missionaries, or in the works of Raja Ram Mohan Roy (1772-1833), Ishwar Chandra Vidyasagar (1820-1891), Bankimchandra Chattopadhyay (1838-1894), Michael Madhusudan Datta (1824-1873), Rabindranath Tagore (1861-1941), and Sarat Chandra Chattopadhyay (1876-1938) [150]. The colloquial variety of Calcutta (called ‘Kolkata’ now) made its first appearance through the *Hutōm Pẽcār Naksā* (1862) by Peari Chand Mitra. The influence of English in vocabulary, idioms, expressions as well as in the writing styles were significant. The fonts and types for Bengali developed during this time also spread to all parts of the Bengali speech community [101, 120].

Bengali prose had developed two literary styles during the 19th-20th Century: The *Sādhubhāşā* (সাধুভাষা - "Elegant Language or Style") and the *Calitabhāşā* (চলিতভাষা "Current Language, or Modern Style"). The former is the traditional literary style based on Middle Bengali of the sixteenth century, while the latter is a basically 20th century creation and is based on the speech of educated people in and around Calcutta or Kolkata [115]. It is the latter style that is prevalent today in written prose. With the latter style came many spelling and script reforms [118]. In the dialogues of plays and fiction, as also in certain experiments as the sole descriptive medium, the dialects of Dhaka-Mymensingh and other places are being extensively used.

## 3.1. Written Bengali

The ‘Bengali alphabet’ (বাংলা লিপি - Bānglā lipi, ISO 15924) is derived from the Brāhmī writing system, which is closely related to the Devanagari script [108]. Considered to be fifth most widely used writing system in the world, this combined Bengali-Assamese-Manipuri Script (showing some variations for Assamese and Meitei or Bishnupriya Manipuri) ( [131]), was used in the eastern Indian Sanskrit manuscripts too. It was once used also for Bodo and Santali as well both of which officially use Devanāgarī now. For Chakma in India and Bangladesh and for Kokborok in Tripura, it was and still is one of the scripts used. A close variant, called *Tirhutā* (now available also in UNICODE 10.0 as 11480 114DF) or *Mithilākşara* was used for Maithili from the 14th Century until the early-20th century [106]. Some varieties of Bānglā lipi were also written in a system that derived from ‘Nāgarī’ lipi but showed a difference from both Devanagari and Bengali-Assamese-Manipuri scripts. A case in point is ‘Sylheti Nagari lipi’ or ‘Siloţi’ (added to the Unicode Standard in March, 2005 with the release of version 4.1) the details of which could be of interest to historians and historical linguists (See 137 and 145), but Sylheti Bengali is generally written by many in the modern-day Bangla script now for all practical purposes. Originally, during the reign of the Pāla dynasty (750-1154 AD) in the eastern India, and even earlier, perhaps during the Malla period (694 AD onwards), the present-day Bengali writing system got a shape comparable to the modern-day ones [111, 119]. A pictorial description of Brāhmī to Modern Bengali could be presented here in a tabular form:



|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Modern | ক | জ | ম | র | স | অ |
|   | k | j | m | r | s | a |

Table 1: Pictorial depiction of Evolution of Brāhmī to Devanagari & Bengali

The inscriptional evidence in Brāhmī is found in the Archaic Brāhmī from the 3rd century B.C. to the 1st century B.C, and in Middle Brāhmī – soon after (1st-3rd Century A.D.) and then on in the Late Brāhmī (4th-6th Century A.D.). As R.C. Majumdar [108] shows, in his *History of Ancient Bengal*, this evidence could be seen in both West Bengal and Bangladesh by 1) The Mahasthangarh (Bogra district, Bangladesh – the ancient name being Pundranagara or Paundravardhanapura) inscriptions, 2) Brāhmī (and Kharoṣṭhī) inscriptions from the lower Gangetic Bengal and (3) Copper plate inscriptions of the Imperial Guptas from Northern part of West Bengal and North-West Bangladesh - in the areas under Dharmaditya, Gopachandra and Samāchāradeva (about whom one only knows from five Copper-plates found in Kotalipara in the Faridpur district in Bangladesh, one in Mallasarul in the Burdwan district (West Bengal), and one in Jayramapur (Balleshvara district, now in Odisha).

These epigraphs from the eastern part of Undivided India (dating back to the 4th-6th Centuries A.D.) showed some characteristic features of letters (especially in ম ‘ma’, ল ‘la’, শ ‘sha’, স ‘sa’ and হ ‘ha’), which led to the development of eastern variety of Gupta script. Epigraphic records from Bangladesh demonstrate remarkable developments in Eastern Brāhmī. In this context, the Tippera copper plate inscription of the ‘Samataţa’ rulers (139, pp 265) such as Lokanātha (dated 7th Century A.D., during the later half), the Kailan inscription of Sridharana Rāta as well as the Astafpur copper plates. The letters seem to hang down from wedge shaped solid triangles with right hand verticals bending down at the bottom, because of which it was described by Prinsep and Fleet as *Kuṭila-lipi* (literally, ‘Cursive writing style’), whereas the term *Siddhamātrikā* was used by Al Biruni (973-1048) to designate the script of Northern India. The next stage of development is illustrated by the 9th Century copper plate inscriptions from Khalimpur of the reign of Dharmapāla, from Monghyr and Nalanda of the time of Devapāla in Bihar, and from Jagjibanpur (Malda) of the reign of Mahendrapāla. The Siddhamātrikā (mentioned as ‘Siddham’ in Chinese sources) is said to have been prevalent also in Bengal up to the end of the tenth century. Also called the Gauri (i.e. Gandi) in Pūrvadeśā or the Eastern country, it was regarded as the same script to which is given the appellative Proto-Bengali characteristics in rudimentary forms, in the period between A.D. 875 and A.D. 1025. In some epigraphs it is considered as belonging to the second quarter of the eleventh century A.D. Flattening of head-marks becomes prominent in comparison to the wedge-shaped serifs. An important landmark in the development of the Bengali script is the Ramganj copper plate inscription of Mahāmāṇḍalika in the last quarter of the eleventh century A.D. It is the earliest document from Bengal which bears the letter m, a with a tick rising upwards. The full vowel i develops a tick at the right end of the upper horizontal bar above and a curved hook below. Initial e approaches the modern Bengali character. A mature form of Proto-Bengali, the immediate precursor of Bengali script, is illustrated in the inscriptions of the Varman Sena and Deva rulers of the twelfth and thirteenth centuries [104].

The evolution of the Bengali script is aligned with the story of advancement of printing technology. The first “Movable type” scripts technically created and used while printing Nathaniel Brassey Halhed's (1751-1830) 1778-book titled, 'A Grammar of the Bengal Language'. In 1785, Governor-General Warren Hastings (1732-1818) requested another civilian, Charles Wilkins (1749-1836) to cut punches for Bengali printing characters. The current printed form of Bengali script appeared soon after. It is generally agreed that Wilkins developed Bengali print script [111]. He passed on this knowledge to Panchanan Karmakar (?-1804), a renowned artist in Bengal. Later it was Karmakar and his family that became famous in Bengali printing technology. Shepherd was another assistant of Wilkins in this designing of script, which became more angular with sharper turns and edges [133]. A few archaic letters were modernised during the 19th century. It was standardized by Pandit Ishwar Chandra Vidyasagar when the Bengali type fonts were to be used to publish on a large scale under the Calcutta School Book Society [116 for several references]. Much later, in 1935, the Linotype technique, invented by Ottmar Mergenthaler (1854-1899) in 1886, was introduced into Bengali printing in 1935, by the efforts of Suresh Chandra Majumdar (1888-1954), Rajsekhar Basu (1880-1960), Jatindra Kumar Sen (1882-1966) and his disciple, Sushil Kumar Bhattacharya and had begun being used by the Anandabazar Patrika group, later followed by others. Within a few years the more advanced monotype technology came to be used Bengali printing. However, in Bengali printing culture, monotype has a very limited acceptance and linotype held stage till, eventually, the digital technology came in to replace all earlier techniques.

All these could be presented in a table:

| **PERIOD** | **DESCRIPTION** | **NAMES** |
| --- | --- | --- |
| 3rd Millennium B.C. | During the Harappan civilization, the script was developed which was partly pictographic, and perhaps written from right to left, and also in a manner of ‘boustrophedon’, i.e. bi-directionally, where every other line is reversed. The attempts are still on to unravel the mystery of this script and its characters. | Indus Valley Script |
| 3rd Century B.C. | Use of Brāhmī and Kharoşţhī scripts begin in the subcontinent. Brāhmī was widely used during the Mauryan King, Aśoka. In one theory, Brāhmī is based on North Semitic alphabet but suitably modified to fit the need of local languages. It is currently believed to have been an independent development. | Brāhmī |
| 1st-3rd Century AD | The Kushan script, named after the Kushan royal dynasty. | Kushan script |
| 4th-5th Century AD | The next stage of its evolution was into the Gupta script, named after the Gupta royal dynasty. | Gupta script |
| 7th Century AD | Epigraphic records from Bangladesh demonstrate remarkable developments in Eastern Brāhmī, giving rise to the *Kuţila-lipi* | Kuţila-lipi |
| 8th Century AD | Some copper plate inscriptions are found in the Khalimpur, Bangladesh during the reign of Dharmapāla, from Monghyr and Nālandā in Bihar, of the time of Devapāla, and from Jagjibanpur in West Bengal of the reign of Mahendrapāla. | *Siddhamātikā* |
| 9th Century AD until 1025 AD | Proto-Bengali characteristics in rudimentary forms develops. An important landmark in the development of the Bengali script is the Ramganj copper plate inscription of Mahāmāᶇȡalika found in the last quarter of the eleventh century A.D. | Proto-Bengali Script & Language |
| 12th-13th Century AD | A mature form of Proto-Bengali, the immediate precursor of Bengali script, is found in the inscriptions of the Varman Sena and Deva rulers of the twelfth and thirteenth centuries. | Matured Proto-Bengali |
| 14th-15th Century AD | The characteristics of typical Bengali script began to develop, as could be seen in the copper plate inscription of Vijayamāᶇikya-I of Tripura dated 1478 AD - also Illustrates forms of Bengali letters in the fifteenth century A.D. | Modern Bengali Script era begins (See Ross 1999) |
| 16th-17th Century AD | The chart of the Bengali alphabet, appended to the China Monuments, published from Amsterdam in 1667 and The code of Gentoo law, published from London in 1776, both show a chart of the Bengali alphabet. They show 16 Vowel letters, including the Long ‘Li’, Anusvāra and Visarga, and 34 Consonants. | Printed Charts of Bengali |
| 18th-19th Century AD | Charles Wilkins develops printing in Bengali in 1778 & Vidyasagar reforms it. | Bangla Type Fonts |

Table 2: Development of the Bengali Writing System

The overall development of Bengali Script from the Kutila-lipi period to Modern Bengali could be seen here in Table 3 ([102 and 147] and also see the web-page in 148).



Table 3: Bengali Script in Different Centuries

## 3.2. Languages Considered

Below is the tabular representation of the languages using Bengali script that are placed on EGIDS Scale 1-6 (See 117 for details). Some languages under EGIDS 5 and 6 have also developed their own scripts for printing and publishing. Some had used Bengali script earlier (such as Bodo), or used it in West Bengal at some point of time (Santali) but have later shifted to another writing system. Bodo is now written in Devanagari and for Santali one uses both Devanagari and *Ol-chiki* (146). For the purposes of the Bengali LGR, languages belonging to the EGIDS scale 1 to 4 only have been considered - just as it was done in the case of Devanagari LGR. Consider the following table:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **EGIDS Scale 1** | **EGIDS Scale 2** | **EGIDS****Scale 3** | **EGIDS Scale 4** | **EGIDS****Scale 5** | **EGIDS 6** |
| Bengali(Bangla) |  |  |  | Santali, Bodo,Riang, Khumi, Mru(ng), Asho | LepchaPnar, Koda/ Kora, Chak |
|  | Assamese |   |   | Koch or Rajbangshi | Malto or Malpahariya |
|  | Manipuri or Meitei |   | Bishnupriya Manipuri,Kok-Borok (Tripura & Bangladesh) | Chakma, Hajong, Mundari &Kurux ( of Bangladesh) | Toto,Rohingya,Tippera,Megam,Tanchangya |
|   |   |   | Usoi | Limbu, Sadri or Oraon | Bhumij or Mundari, Bawm, Chin |

Table 4: Main languages in India and Bangladesh

that use Bengali Script on the EGIDS Scale

## 3.3. Notable Features [151]

* The Bengali script is an alpha-syllabic writing system in which writing of all consonants are assumed to contain an accompanying ‘inherent’ vowel (theoretically after each consonant). It varies between /ɔ/ and /o/ depending on the position of the consonant in the word. At times, these assumed or ‘inherent’ vowels are not pronounced at all [142].
* Vowels can be written as independent letters, or by using a variety of diacritical marks which are written above, below, before, after or both of the last two positions the consonant they follow in pronunciation [105].
* All Bengali consonants when pronounced in isolation are uttered with an inherent vowel - / ɔ/; hence ক ‘k’, খ ‘kh’ or গ ‘g’ are usually pronounced as [kɔ], [khɔ], or [gɔ], etc. Phonologically, Bengali vowel - / ɔ/ corresponds to the Hindi schwa /ə/
* When consonants occur together in clusters, special conjunct letters are formed. In printed Bengali, a large number of these consonantal clusters or conjoined consonants are in use. The letters for the consonants other than the final one in the group are generally reduced. But there are a few special conjunct characters which are compounds of the consonant characters, e.g. ক্+ষ=ক্ষ, ঞ্+জ=ঞ্জ, জ্+ঞ=জ্ঞ, হ্‌+ম=হ্ম. There are other issues also—র as the second member of a cluster is reduced to a secondary symbol, e.g. প্+র=প্র, ষ্+ট্+র=ষ্ট্র (as in উষ্ট্র); য, when used as a primary symbol, represents /jɔ/ in Bengali. But its secondary symbol (allograph) jɔ-phalā has two phonetic values. When added to the initial consonant in a word, it is a vowel /æ/ (as in শ্যামল, র‌্যাপার, etc.). But after a non-initial consonant, it just doubles it in pronunciation (as in কার্য, ধার্য, etc.). The র্+য combination has two physical manifestations—র‌্য and র্য. In case of দ্+ধ, গ্+ধ, ন্+ধ the shape of the second member is changed—e.g. দ্ধ, গ্ধ, and ন্ধ respectively. The solitary example of র্‌+ঋ=র্ঋ (as in নৈর্ঋত) shows the use of secondary symbol of a consonant followed by the primary symbol of a vowel. The inherent vowel only applies to the final consonant of the cluster.
* In consonant clusters, many consonants took a completely different form. Some typical examples are ক্ত (kt), ক্র (kr), ক্ষ (kṣ), গ্ধ (gdh), জ্ঞ (jñ), ঞ্চ (ñc), ঞ্জ (ñj), ট্ (ṭṭ), ন্ত (nt), ন্ধ (ndh), ব্ধ (bdh), ভ্র (bhr), ম্ব (mb), স্ত (st) etc. র has two allographs, apart from this full shape : one is ‘reph’, as found in র্ক (rk), র্প (rp); and another is ra-phalā, as in প্র (pr), ক্র (kr). ষ্ণ (ṣ+ṇ) is another one, where the cerebral nasal consonant sign takes a queer shape. [152]
* The Bengali script has at least fifty-two primary symbols and quite a few allographs (positional variants of them), corresponding to forty-four (7 oral and 7 nasal vowels and 30 consonants) phonemes (151) or functional speech sounds, with some obvious redundancies, although in one of the first phonemic analysis, the number was thought to be thirty-five phonemes [140].
* As mentioned above, in Bengali, several graphemic symbols have secondary shapes, technically called ‘allographs’ with a complementary distribution in each case. These graphs or markings are generally added to the following positions of the primary symbol [113] in the following manner:

1) Below (e.g. কু, ন্ত, কূ, হ্র, etc.)

2) Above (e.g. চঁ, র্ক, etc.)

3) Right side (e.g. কা, কং, etc.)

4) Left side (e.g. কে)

5) Left Side and above simultaneously (e.g. কৈ, কি etc.)

6) Right side and above simultaneously (e.g. কী)

7) Right side and left side simultaneously (e.g. কো)

8) Right side, left side and above simultaneously (e.g. কৌ).

* Besides some simple Vowel Modifiers or ‘Matra’s there are some combinatory modifiers of Bengali Vowels with certain consonants. For example, whereas

আ U+0986 BENGALI LETTER AA is substituted by

া U+09BE BENGALI VOWEL SIGN AA,

ই U+0987 BENGALI LETTER I is substituted by

pre-posed ি U+09BF BENGALI VOWEL SIGN I,

ঈ U+0988 BENGALI LETTER II is substituted by

 ী U+09C0 BENGALI VOWEL SIGN II or

উ U+0989 BENGALI LETTER U is substituted by

 ু U+09C1 BENGALI VOWEL SIGN U by marking below the primary grapheme, there are some special vowel modifiers of উ as in the following combined letters:

 গু gu, rather than writing as গ্ + ু

 রু ru, rather than writing as র্ + ু

 শু śu, rather than writing as শ্ + ু

 হু hu, rather than writing as হ্ + ু

 ন্তু/ন্তু ntu, rather than writing as ন্+ত্+ু

 Similarly, there could be vowel modifiers of ঊ or ‘(Long) ū’ as well; e.g.

 ভ্+র্ (ভ্রূ), শ্+র্ (শ্রূ), ঋ after হ্ (হৃ), etc.

* The global Bengali-speaking diaspora using Bangla script (and language) live in a number of countries, including in the UK, USA, Canada, the Middle East, Japan, South Korea, Malaysia, Pakistan, Singapore, and Italy and some other countries of Europe.
* There have been many notable contributions in simplifying and modifying Bengali spellings and combinatory techniques, especially by scholars such as Pabitra Sarkar (1992) [134]. In this there has been an attempt to reduce the number of allographs of both vowels and consonants in clusters, and it has been widely accepted in the printing of school texts in both Bangladesh and West Bengal [152, 153]. But in preparation of this LGR document, the aim has been to consider the widely used and usable sequences and combinations and their variations across other scripts belonging to the basket of Devanāgarī writing systems.
* After the establishment of Bangla Akademi of West Bengal in 1986, its first President, Annadasankar Ray (1904-2002), in his inaugural address, gave a direction for standardization of Bangla alphabet, script, the spelling system and clearly argued that they would not blindly follow the Sanskritic model of conventional grammar. A broad list of proposals was sent to experts on Bangla, and a broad agreement was reached for ‘homogenization of Bangla spelling’ by 1988. Based on opinions received from different quarters, a unanimous list of ‘rules’ was agreed upon. This was published by a ‘Spelling Dictionary’ titled, Akademi Banan Abhidhan (1997), which was obviously more comprehensive than ‘The University of Calcutta proposals’, made in 1936. Along with the ‘rationalization’ of spellings, another step was taken make the writing system easier to read, by making the symbols used, both single and combined ones, more ‘transparent’.

There were, in fact, two types of proposals.  One concerned the shape of the letters, those of consonant + vowel (CV) combinations and conjuncts, that is consonant + consonant combinations.  There were further complex shapes, i.e. those of consonant + consonant+ (consonant+) vowel (CC(CV) signs, as in প্রু, or স্ক্রু. Some decisions in this area were necessary because a few of the CC(C) symbols represented complexities that made learning them difficult for the children. The other dealt with the spellings of words only, without any reference to the shapes of letters in which they were written.  The basic objective here was ‘one word, one spelling’, to the greatest extent that was possible. [152]

### 3.3.1 The Consonants

As per traditional classification Bangla Consonants are categorized according to their phonetic properties, especially in terms of place and manner of articulation [107]. There are Five ‘Barga’ or Groups (sets or classes) distinguished by Place of Articulation, and one Non-’Barga’ group [105]. Each Barga, which corresponds to Stops at a certain place of articulation, contains a series of five consonants classified as per their phonetic qualities (i.e. manner of articulation), beginning from Unvoiced and Unaspirated to Voiced and Aspirated (in the fourth column), finally ending with a Homorganic or Corresponding nasal [107]. Consider the following table:

|  |  |  |  |
| --- | --- | --- | --- |
| **‘Barga’ or Sets** | **Unvoiced** | **Voiced** | **Nasal** |
|  | -Asp | +Asp | -Asp | +Asp |  |
| **Velar** | ক ‘k’U+0995 | খ ‘kh’U+0996 | গ ‘g’U+0997 | ঘ ‘gh’U+0998 | ঙ ‘Ng’U+0999 |
| **Palatal** | চ ‘c’U+099A | ছ ‘ch’U+099B | জ ‘j’U+099C | ঝ ‘jh’U+099D | ঞ ‘Ny’U+099E |
| **Retroflex** | ট ‘Tt’U+099F | ঠ ‘Tth’U+09A0 | ড ‘Dd’U+09A1 | ঢ ‘Ddh’U+09A2 | ণ ‘Nn’U+09A3 |
| **Dental** | ত ‘t’U+09A4 | থ ‘th’U+09A5 | দ ‘d’U+09A6 | ধ ‘dh’U+09A7 | ন ‘n’U+09A8 |
| **Bilabial** | প ‘p’U+09AA | ফ ‘ph’U+09AB | ব ‘b’U+09AC | ভ ‘bh’U+09AD | ম ‘m’U+09AE |

Table 5: Barga classification of Bengali consonants

(Falling into a Pattern of Five Sets of Unvoiced Unaspirated, Unvoiced Aspirated, Voiced Unaspirated, Voiced Aspirated and Nasals, called five ‘Barga’)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Non-****Barga** | য ‘y’U+09AF | র ‘r’U+09B0 | ল ‘l’U+09B2 | শ ‘Sh’U+09B6 | ষ ‘Ss’U+09B7 | স ‘s’U+09B8 | হ ‘h’U+0939 |

Table 6: Non-Barga consonants (Not falling into any of the five categories)

### 3.3.2 The Implicit Vowel Killer: Halant (=’Hasanta’ in Bengali)

As stated earlier, all consonants are pronounced in isolation with an implicit vowel (central back /-ɔ/ in Bengali as the neutral vowel) assumed to be associated with them [121]. The ‘Halant’ (=’Hasanta in Bengali) or the term ‘*Virāma*’ as preferred in UNICODE (cf. Unicode 3.0 and above) have been used in this report as terms that have been used to denote the character that mark the absence of this inherent vowel. Because a special sign is needed whenever this implicit vowel is stripped off, the symbol is known as the *Halant* (=*Hasanta*) "्" (U+09CD). By placing the Halant under the first consonant of a combination or cluster, one could – in common parlance, “kill” its vowel, and create conjuncts. In this manner, conjunct characters can be generally done by joining two to four consonant combinations. In rare cases, this process can join up to five consonants. However, the notion of maximum number of consonants joining to form one *akshara* is to be empirically seen. It is an observation based on the CIIL-Emille Corpora of Bengali words [132] as seen in print till date. Given the mixture of scripts and languages happening on the web, the possibility that one may want a generic Top Level Domain [gTLD] which may have more than the observed maximum cannot be ruled out. This can be the case when a foreign language word, which admits a large number of consonants, is transliterated into Bengali. Hence, in the Bangla LGR work, this limit will not be enforced.

### 3.3.3 Vowels

Separate symbols exist for all ‘*Swara*’ or Vowels in Bengali, which are pronounced independently either at the beginning of the word or after another vowel or consonant sound. To indicate a Vowel sound other than the implicit one, a Vowel sign, called *Mātrā* in Deva-nagari (although the term Mātrā in Bengali stands for an altogether different concept, viz.the top bar placed over a letter – typically available in Hindi and Bengali but missing in Gujarati) is attached to the consonant. Since the consonant has this built in neutral vowel at the end, there are equivalent Mātrās for all vowels except the অ (pronounced /-ɔ/). The correlation is shown as follows:

| **Vowel** | **Corresponding vowel sign (Mātrās)** |
| --- | --- |
| অ ‘a’ U+0985 |   |
| আ ‘aa’ U+0986 |  া U+09BE |
| ই ‘i’ U+0987 | ি U+09BF |
| ঈ ‘ii’ U+0988 | ী U+09C0 |
| উ ‘u’ U+0989 |  ু◌ U+09C1 |
| ঊ ‘uu’ U+098A | ূ◌ U+09C2 |
| ঋ Vocalic ’r’ U+098B | ৃ◌ U+09C3 |
| ৠ Vocalic ‘rr’ U+09E0 |  ৄ◌ U+09C4 |
|  ঌ Vocalic ‘l’ U+098C |  ৢ◌ U+09E2 |
|  ৡ Vocalic ‘ll’ U+09E1 |  ৣ◌ U+09E3 |
| এ ‘e’ U+098F |  ে U+09C7 |
| ঐ ‘ai’ U+0990 | ৈ U+09C8 |
| ও ‘o’ U+0993 | ো U+09CB |
| ঔ ‘au’ U+0994 | ৌ U+09CC |
| - | ৗ U+09D7 |
| Could appear on top of অ ‘a’ U+0985 or any other vowel | ঁ U+0981 Candrabindu |
| Could appear after অ ‘a’ U+0985 or any other vowel | ং U+0982 Anusvara |
| Could appear after অ ‘a’ U+0985 or any other vowel | ঃ U+0983 Visarga |
| - | ় U+09BC Nukta |
| After any consonant | ্ U+09CD (Halant) |
| - | ঽ U+09BD Avagraha |

Table 7: Bengali Vowels with corresponding Mātrās

### 3.3.4 The Anusvara (ং - U+0982)

The Anusvara in Bengali at times represents a homorganic nasal but not always. It replaces a conjunct group of a ‘Nasal Consonant+Halant+Consonant’ where the second consonant belongs to the Velar *barga* or set as in লংকা. But it often appears also for such combinations involving non-velars appearing as the last member of the combination as in ল্যাংটা, or ল্যাংচা. Before a non-*barga* consonant, the Anusvara represents a nasal sound that may have an alternative conjoined writing symbol representing the corresponding nasal consonant of the particular set. Although Modern Hindi, Marathi and Konkani prefer the anusvara to the corresponding Half-nasal, in Bengali it is clearly demarcated as to where one must use the Anusvara and where it has to be a conjunct cluster with a nasal as the first or the second component.

### 3.3.5 Nasalization: Candrabindu (ँ - U+0981)

Candrabindu denotes nasalization of the preceding vowel as in চাঁদ /cā̃d/ ‘moon’ (U+099A U+09BE U+0981 U+09A6). This sign with a dot inside the half-moon mark is used as nasalization marker in many Indian scripts. [143]

### 3.3.6 Nukta (़ - U+09BC)

The nukta sign is placed below a certain number of consonants to represent sounds found only in words borrowed from Perso-Arabic. It is predominantly used in this manner in Bodo, Hindi, Kashmiri, Maithili, Santali and Sindhi. In Bengali, its use is further restricted. It can be optionally adjoined to ক KA (U+0995), খ KHA (U+0996), গ GA (U+0997), জ JA (U+099C) and ফ PHA (U+09AB) to show that words having these consonants with a nukta are to be pronounced in the Perso-Arabic style. e.g. ফিরোজ় /firoz/ (U+09AB U+09BC U+09BF U+09B0 U+09CB U+090C U+09BC)

It is also placed under "ড" DDA (U+09A1) and "ঢ" DDHA (U+09A2) to indicate flapped sounds বড় /bədh/ (U+09AC U+09A1 U+09BC). Of course, Bengali Unicode points already account for these two letters separately as under ড় RRA (U+09DC) and ঢ় RRHA (U+09DD).

Normally a Nukta is appended to Consonants. However, Santali language uses Nukta in a unique way, also under certain vowels and vowel signs, especially when it uses Devanagari script (and not when Santali is printed in Bengali script):

a. आ LETTER AA (U+0906)

b. ओ LETTER O (U+0913)

c. ा VOWEL SIGN AA (U+093E)

d. ो VOWEL SIGN O (U+094B)

### 3.3.7 Visarga (ঃ - U+0983) and Avagraha (ঽ - U+09BD)

The Visarga U+0983 is frequently used in Bengali loanwords borrowed from Sanskrit and represents a sound very close to /h/. One could quote, as an example: দু:খ /duhkho/ sorrow, unhappiness (U+0926 U+0941 U+0983 U+0916).

The Avagraha "ঽ" (U+09BD) is mainly used in Sanskrit or Maithili texts written in Bengali. It is gradually being replaced by an upper comma (e.g. নরো’পরাণি). It is now rarely used in other languages using Bengali script. In case of LGR, the Avagraha is not part of the repertoire as it is barred in the Maximal Starting Repertoire.

Please see the Annex II for a complete list of Bengali consonants and their allographs.

# Overall Development Process and Methodology

The Neo-Brahmi Generation Panel (NBGP) has been formed by members having experience in Linguistics (especially in NLP/Computational linguistics), Literature, Language History and Epigraphy. Under the Neo-Brahmi Generation Panel, Bengali and eight other scripts belonging to separate Unicode blocks are being taken up to assign a separate LGR for each. However, an attempt is made to ensure that the fundamental philosophy behind building those LGRs consistent with all other Brāhmī-derived scripts, especially with the Devanagari writing system. The present LGR will cater to multiple languages belonging to EGIDS scale 1 to 4 (see Table 4) that use Bengali script.

The following guiding principles are used in making decisions about Bengali LGR Code-points:

## Guiding Principles

The NBGP adopts following broad principles for selection of code-points in the code-point repertoire across the board for all the Neo-Brahmi scripts within its ambit.

### Inclusion Principles

#### Modern Usage

Every character proposed should be in the everyday usage of a particular linguistic community. The characters which have been encoded in the Unicode for transcription purposes only or for archival purposes will not be considered for inclusion in the code-point repertoire.

#### Unambiguous Use

Every character proposed should have unambiguous understanding among linguists about its usage in the language.

### Exclusion Principles

The main exclusion principle is that of External Limits on Scope. These consist of protocols or standards which are prerequisites to the Label Generation Rule-sets. All further principles are in fact subsumed under these limitations but have been spelt out separately for the sake of clarity.

#### External Limits of Scope

The code point repertoire for root zone being a very special case, at the top of protocol hierarchies, the canvas of available characters for selection as a part of the Root Zone code point repertoire is already constrained by various protocol layers beneath it. The following three main protocols/standards act as successive filters:

*i. The Unicode Chart*

Out of all the characters that are needed by the script in question, if a particular character is not encoded in Unicode, it cannot be incorporated in the code point repertoire. Such cases are quite rare, and especially so in Bengali-Assamese-Manipuri Script, given the elaborate and exhaustive character inclusion efforts made by the Unicode consortium.

*ii. IDNA Protocol*

Unicode being the character-encoding standard for providing the maximum possible representation of a given script/language, it has encoded as far as possible all the possible characters needed by the script. However the Domain name being a specialized case, it is governed by an additional protocol known as IDNA (Internationalized Domain Names in Applications). The IDNA protocol excludes some characters out of Unicode repertoire from being part of the domain names.

Here are some examples: Bengali Letter RRA “ড়" (09DC), a consonant letter with “nukta” corresponding to Bengali Letter DDA “ড” (09A1), is not allowed to be a part of domain name. Bengali Letter DDA “ড” (09A1) followed by Bengali Sign NUKTA “়” (09BC) can be used instead.

*iii. Maximal Starting Repertoire (MSR)*

Since the Root-zone LGR being is the repertoire of characters which are going to be used for creation of the Root-zone TLDs, which in turn constitute an even more specialized case of domain names, the ROOT LGR procedure introduces additional exclusions on the IDNA’s allowed set of characters.

Example: Bengali Sign Avagraha "ঽ" (U+093D) even if allowed by IDNA protocol, is not permitted in the Root Zone Repertoire as per the MSR.

To sum up, the restrictions start off with admitting only such characters as are part of the code-block of the given script/language. The IDNA Protocol further narrows this down and finally an additional filter in the form of Maximal Starting Repertoire restricts the character set associated with the given language even more.

#### No Punctuation Marks

The TLDs being identifiers, punctuation markers present in Brahmi-based languages will not be included.

#### No Symbols and Abbreviations

Abbreviations, weights and measures and other such iconic characters like BENGALI ISSHAR "৺" (U+09FA), BENGALI CURRENCY DENOMINATOR SIXTEEN "৹" (U+09F9) etc. will also not be included.

#### No Rare and Obsolete Characters

There are characters which have been added to Unicode to accommodate rare forms such as Sanskritic VOCALIC RR "ৠ" (U+09E0) and VOCALIC L “ঌ”(U+098C) as well as VOCALIC LL "ৡ" (U+09E1) and the matra forms of the latter two symbols - VOWEL SIGN VOCALIC L "ৢ" (U+09E2) and VOWEL SIGN VOCALIC LL “ৣ" (U+09E3). All such characters will be excluded. This is in compliance with the Conservatism principle as laid down in the Root Zone LGR procedure. However, in Bengali, the vowel matra corresponding to VOCALIC RR "ৠ" (U+09E0) which is VOWEL SIGN VOCALIC RR “ৄ” (U+09C4) is still in active use in certain borrowed or Sanskritic words, and could therefore be retained.

#### No Stress Markers of Classical Sanskrit and Vedic

Stress markers for classical Sanskrit will not be included. This is also in consonance with the Letter principle as laid down in the Root Zone LGR procedure.

# Repertoire

The Bengali Writing System is represented in UNICODE using the same script name as ISO 15924 corresponding to languages such as Assamese, Bengali (Bangla) and Manipuri. The Bengali block used for Bengali-Assamese-Manipuri in the UNICODE has 93 entries. This section details the code-point repertoire that the Neo-Brahmi Generation Panel [NBGP] proposes to be included in the Bengali LGR. It could be mentioned here that the Government of Assam has submitted a proposal to Bureau of Indian Standards (BIS) on 26th February, 2016 for dis-unification of Bengali and Assamese Script, and the BIS in its 8th Meeting of Indian Language Technologies and Products Sectional Committee, LITD 20, held on 23rd Aug 2017, and decided to refer the proposal for recognition of Assamese script in ISO/IEC 10646 to ISO. Until the UNICODE Consortium takes any further action, it will be assumed that the Code Point Repertoire under Table 11 will be valid for all the three languages as above.

For each of the code points, language references have been given in the last column titled "Reference" under Table 8 titled the “Code Point Repertoire”. For entire coverage of Bengali code points, references of Bengali, Assamese, Manipuri (Meitei), and Bishnupriya have been given. Kokborok, written in Bengali script, is not known to have introduced many new complications. Though only a few representative languages under EGIDS Scale 1-4 have been chosen for referencing, they together cover all the code-points required for all the languages that NBGP has considered as given under Bengali Unicode Points (as given in UNICODE 6.3).

However, before the details are presented, it is ideal to take a look at the Bengali Code Point Chart from Maximal Starting Repertoire [MSR] Version 3. It may be noted that the shapes of the reference glyphs given below in the code charts are based on one of the many fonts designed, and are not prescriptive, because there could be some variations in actual fonts. Consider the following Code point table:

|  |  |
| --- | --- |
| Figure 1: Bengali Code Page from [MSR] for Bengali-Assamese-Manipuri | **Color convention[[2]](#footnote-2):**All characters that are included in the [MSR] - Yellow backgroundPVALID in IDNA2008 but excluded from the [MSR] - Pinkish backgroundNot PVALID in IDNA2008, or are ineligible for the root zone (digits, hyphen) - White background |

Given the Bengali Unicode Block as in Figure 1, for the code points those are included in the MSR, the following symbols will need a separate treatment:

ৎ U+09CE Bengali Letter Khanda-Ta

ৰ U+09F0 Assamese-Bengali Letter Ra With Middle Diagonal

ৱ U+09F1 Assamese-Bengali Letter Ra With Lower Diagonal

## Code Point Repertoire Inclusion

| **No.** | **Unicode Code Point** | **Glyph** | **Character Name** | **Indic Syllabic Category** | **Language(s), with EGIDS Value** | **References** |
| --- | --- | --- | --- | --- | --- | --- |
| 1. | U+0981 | ঁ | BENGALI SIGN CANDRABINDU | Chandra- bindu | 1 Bengali,2 Manipuri, 2 Assamese | [101], [102],[103], [111], [112], [113], [119],[120],[121], [122], [125], [127], [128] |
| 2. | U+0982 | ং | BENGALI SIGN ANUSVARA | Anusvara | 1 Bengali,2 Manipuri, 2 Assamese | [101], [102],[103], [111], [112], [113], [119],[120],[121], [122], [125], [127], [128] |
| 3. | U+0983 | ঃ | BENGALI SIGN VISARGA | Visarga | 1 Bengali,2 Manipuri, 2 Assamese | [101], [102],[103], [111], [112], [113], [119],[120],[121], [122], [125], [127], [128] |
| 4. | U+0985 | অ | BENGALI LETTER A | Vowel | 1 Bengali,2 Manipuri, 2 Assamese | [102], [103], [104], [105], [107], [111], [112], [113], [114], [121], [122], [125], [127], [128] |
|  5. | U+0986 | আ | BENGALI LETTER AA | Vowel | 1 Bangla,2 Manipuri, 2 Assamese | [102], [103], [104], [105], [107], [111], [112], [113], [114], [121], [122], [125], [127], [128] |
|  6. | U+0987 | ই | BENGALI LETTER I | Vowel | 1 Bengali,2 Manipuri, 2 Assamese | [102], [103], [104], [105], [107], [111], [112], [113], [114], [121], [122], [125], [127], [128] |
|  7. | U+0988 | ঈ | BENGALI LETTER II | Vowel | 1 Bengali,2 Manipuri, 2 Assamese | [101], [102], [103], [104], [105], [107], [111], [112], [113], [114], [121], [122], [125], [127], [128] |
| 8. | U+0989 | উ | BENGALI LETTER U | Vowel | 1 Bengali,2 Manipuri, 2 Assamese | [102], [103], [104], [105], [107], [111], [112], [113], [114], [121], [122], [125], [127], [128] |
| 9. | U+098A | ঊ | BENGALI LETTER UU | Vowel | 1 Bengali,2 Manipuri, 2 Assamese | [101], [102], [103], [104], [105], [107], [111], [112], [113], [114], [121], [122], [125], [127], [128] |
| 10. | U+098B | ঋ | BENGALI LETTER VOCALIC R | Vowel | 1 Bengali,2 Manipuri, 2 Assamese | [101], [102], [103], [104], [105], [107], [111], [112], [113], [114], [121], [122], [125], [127], [128] |
| 11. | U+098F | এ | BENGALI LETTER E | Vowel | 1 Bengali,2 Manipuri, 2 Assamese | [102], [103], [104], [105], [107], [111], [112], [113], [114], [121], [122], [125], [127], [128] |
| 12. | U+0990 | ঐ | BANGLA LETTER AI | Vowel | 1 Bengali,2 Manipuri, 2 Assamese | [102], [103], [104], [105], [107], [111], [112], [113], [114], [121], [122], [125], [127], [128] |
| 13. | U+0993 | ও | BENGALI LETTER O | Vowel | 1 Bengali,2 Manipuri, 2 Assamese | [102], [103], [104], [105], [107], [111], [112], [113], [114], [121], [122], [125], [127], [128] |
| 14. | U+0994 | ঔ | BENGALI LETTER AU | Vowel | 1 Bengali,2 Manipuri, 2 Assamese | [102], [103], [104], [105], [107], [111], [112], [113], [114], [121], [122], [125], [127], [128] |
| 15. | U+0995 | ক | BENGALI LETTER KA | Consonant | 1 Bengali,2 Manipuri, 2 Assamese | [102], [103], [104], [105], [107], [111], [112], [113], [114], [121], [122], [125], [127], [128] |
| 16. | U+0996 | খ | BENGALI LETTER KHA | Consonant | 1 Bengali,2 Manipuri, 2 Assamese | [102], [103], [104], [105], [107], [111], [112], [113], [114], [121], [122], [125], [127], [128] |
| 17. | U+0997 | গ | BENGALI LETTER GA | Consonant | 1 Bengali,2 Manipuri, 2 Assamese | [102], [103], [104], [105], [107], [111], [112], [113], [114], [121], [122], [125], [127], [128] |
| 18. | U+0998 | ঘ | BANGLA LETTER GHA | Consonant | 1 Bengali,2 Manipuri, 2 Assamese | [102], [103], [104], [105], [107], [111], [112], [113], [114], [121], [122], [125], [127], [128] |
| 19. | U+0999 | ঙ | BENGALI LETTER NGA | Consonant | 1 Bengali,2 Manipuri, 2 Assamese | [101], [102], [103], [104], [105], [107], [111], [112], [113], [114], [121], [122], [125], [127], [128] |
| 20. | U+099A | চ | BENGALI LETTER CA | Consonant | 1 Bengali,2 Manipuri, 2 Assamese | [102], [103], [104], [105], [107], [111], [112], [113], [114], [121], [122], [125], [127], [128] |
| 21. | U+099B | ছ | BENGALI LETTER CHA | Consonant | 1 Bengali,2 Manipuri, 2 Assamese | [102], [103], [104], [105], [107], [111], [112], [113], [114], [121], [122], [125], [127], [128] |
| 22. | U+099C | জ | BENGALI LETTER JA | Consonant | 1 Bengali,2 Manipuri, 2 Assamese | [102], [103], [104], [105], [107], [111], [112], [113], [114], [121], [122], [125], [127], [128] |
| 23. | U+099D | ঝ | BENGALI LETTER JHA | Consonant | 1 Bengali,2 Manipuri, 2 Assamese | [102], [103], [104], [105], [107], [111], [112], [113], [114], [121], [122], [125], [127], [128] |
| 24. | U+099E | ঞ | BENGALI LETTER NYA | Consonant | 1 Bengali,2 Manipuri, 2 Assamese | [101], [102], [103], [104], [105], [107], [111], [112], [113], [114], [121], [122], [125], [127], [128] |
| 25. | U+099F | ট | BENGALI LETTER TTA | Consonant | 1 Bengali,2 Manipuri, 2 Assamese | [102], [103], [104], [105], [107], [111], [112], [113], [114], [121], [122], [125], [127], [128] |
| 26. | U+09A0 | ঠ | BENGALI LETTER TTHA | Consonant | 1 Bengali,2 Manipuri, 2 Assamese | [102], [103], [104], [105], [107], [111], [112], [113], [114], [121], [122], [125], [127], [128] |
| 27. | U+09A1 | ড | BENGALI LETTER DDA | Consonant | 1 Bengali,2 Manipuri, 2 Assamese | [102], [103], [104], [105], [107], [111], [112], [113], [114], [121], [122], [125], [127], [128] |
| 28. | U+09A2 | ঢ | BENGALI LETTER DDHA | Consonant | 1 Bengali,2 Manipuri, 2 Assamese | [102], [103], [104], [105], [107], [111], [112], [113], [114], [121], [122], [125], [127], [128] |
| 29. | U+09A3 | ণ | BENGALI LETTER NNA | Consonant | 1 Bengali,2 Manipuri, 2 Assamese | [101], [102], [103], [104], [105], [107], [111], [112], [113], [114], [119], [120], [121], [122], [125], [127], [128] |
| 30. | U+09A4 | ত | BENGALI LETTER TA | Consonant | 1 Bengali,2 Manipuri, 2 Assamese | [102], [103], [104], [105], [107], [111], [112], [113], [114], [121], [122], [125], [127], [128] |
| 31. | U+09A5 | থ | BENGALI LETTER THA | Consonant | 1 Bengali,2 Manipuri, 2 Assamese | [102], [103], [104], [105], [107], [111], [112], [113], [114], [121], [122], [125], [127], [128] |
| 32. | U+09A6 | দ | BENGALI LETTER DA | Consonant | 1 Bengali,2 Manipuri, 2 Assamese | [102], [103], [104], [105], [107], [111], [112], [113], [114], [121], [122], [125], [127], [128] |
| 33. | U+09A7 | ধ | BENGALI LETTER DHA | Consonant | 1 Bengali,2 Manipuri, 2 Assamese | [102], [103], [104], [105], [107], [111], [112], [113], [114], [121], [122], [125], [127], [128] |
| 34. | U+09A8 | ন | BENGALI LETTER NA | Consonant | 1 Bengali,2 Manipuri, 2 Assamese | [102], [103], [104], [105], [107], [111], [112], [113], [114], [119], [120], [121], [122], [125], [127], [128] |
| 35. | U+09AA | প | BENGALI LETTER PA | Consonant | 1 Bengali,2 Manipuri, 2 Assamese | [102], [103], [104], [105], [107], [111], [112], [113], [114], [121], [122], [125], [127], [128] |
| 36. | U+09AB | ফ | BENGALI LETTER PHA | Consonant | 1 Bengali,2 Manipuri, 2 Assamese | [102], [103], [104], [105], [107], [111], [112], [113], [114], [121] |
| 37. | U+09AC | ব | BENGALI LETTER BA | Consonant | 1 Bengali,2 Manipuri, 2 Assamese | [102], [103], [104], [105], [107], [111], [112], [113], [114], [121], [122], [125], [127], [128] |
| 38. | U+09AD | ভ | BENGALI LETTER BHA | Consonant | 1 Bengali,2 Manipuri, 2 Assamese | [102], [103], [104], [105], [107], [111], [112], [113], [114], [121], [122], [125], [127], [128] |
| 39. | U+09AE | ম | BENGALI LETTER MA | Consonant | 1 Bengali,2 Manipuri, 2 Assamese | [102], [103], [104], [105], [107], [111], [112], [113], [114], [121], [122], [125], [127], [128] |
| 40. | U+09AF | য | BENGALI LETTER YA | Consonant | 1 Bengali,2 Manipuri, 2 Assamese | [102], [103], [104], [105], [107], [111], [112], [113], [114], [119], [120], [121] |
| 41. | U+09B0 | র | BENGALI LETTER RA | Consonant | 1 Bengali,2 Manipuri | [102], [103], [104], [105], [107], [111], [112], [113], [114], [121] |
| 42. | U+09B2 | ল | BENGALI LETTER LA | Consonant | 1 Bengali,2 Manipuri, 2 Assamese | [102], [103], [104], [105], [107], [111], [112], [113], [114], [121] [122], [125], [127], [128] |
| 43. | U+09B6 | শ | BENGALI LETTER SHA | Consonant | 1 Bengali,2 Manipuri, 2 Assamese | [102], [103], [104], [105], [107], [111], [112], [113], [114], [118], [119], [120], [121], [122], [125], [127], [128] |
| 44. | U+09B7 | ষ | BENGALI LETTER SSA | Consonant | 1 Bengali,2 Manipuri, 2 Assamese | [102], [103], [104], [105], [107], [111], [112], [113], [114], [118], [119], [120], [121], [122], [125], [127], [128] |
| 45. | U+09B8 | স | BENGALI LETTER SA | Consonant | 1 Bengali,2 Manipuri, 2 Assamese | [102], [103], [104], [105], [107], [111], [112], [113], [114], [118], [119], [120], [121], [122], [125], [127], [128] |
| 46. | U+09B9 | হ | BENGALI LETTER HA | Consonant | 1 Bengali,2 Manipuri, 2 Assamese | [102], [103], [104], [105], [107], [111], [112], [113], [114], [119], [120], [121] |
| 47. | U+09BC | ় | BENGALI SIGN NUKTA | Nukta | 1 Bengali,2 Assamese2 Manipuri | [102], [103], [104], [105], [107], [111], [112], [113], [114], [119], [120], [121] [122], [128] |
| 48. | U+09BE | া | BENGALI VOWEL SIGN AA | Matra | 1 Bengali,2 Manipuri, 2 Assamese | [102], [103], [104], [105], [107], [111], [112], [113], [114], [121], [122], [125], [127], [128] |
| 49. | U+09BF | ি | BENGALI VOWEL SIGN I | Matra | 1 Bengali,2 Manipuri, 2 Assamese | [102], [103], [104], [105], [107], [111], [112], [113], [114], [121], [122], [125], [127], [128] |
| 50. | U+09C0 | ী | BENGALI VOWEL SIGN II | Matra | 1 Bengali,2 Manipuri, 2 Assamese | [102], [103], [104], [105], [107], [111], [112], [113], [114], [119], [120], [121], [122], [125], [127], [128] |
| 51. | U+09C1 | ু | BENGALI VOWEL SIGN U | Matra | 1 Bengali,2 Manipuri, 2 Assamese | [102], [103], [104], [105], [107], [111], [112], [113], [114], [119], [120], [121], [122], [125], [127], [128] |
| 52. | U+09C2 | ূ | BENGALI VOWEL SIGN UU | Matra | 1 Bengali,2 Manipuri, 2 Assamese | [102], [103], [104], [105], [107], [111], [112], [113], [114], [119], [120], [121], [122], [125], [127], [128] |
| 53. | U+09C3 | ৃ | BENGALI VOWEL SIGN VOCALIC R | Matra | 1 Bengali,2 Manipuri, 2 Assamese | [101], [102], [103], [104], [105], [107], [111], [112], [113], [114], [119], [120], [121], [122], [125], [127], [128] |
| 54. | U+09C4 | ৄ | BENGALI VOWEL SIGN VOCALIC RR | Matra | 1 Bengali,2 Assamese | [101], [102], [103], [104], [105], [107], [111], [112], [113], [114], [119], [120], [121], [122], [128] |
| 55. | U+09C7 | ে | BENGALI VOWEL SIGN E | Matra | 1 Bengali,2 Manipuri, 2 Assamese | [102], [103], [104], [105], [107], [111], [112], [113], [114], [121], [122], [125], [127], [128] |
| 56. | U+09C8 | ৈ | BENGALI VOWEL SIGN AI | Matra | 1 Bengali,2 Manipuri, 2 Assamese | [102], [103], [104], [105], [107], [111], [112], [113], [114], [121], [122], [125], [127], [128] |
| 57. | U+09CB | ো | BENGALI VOWEL SIGN O | Matra | 1 Bengali,2 Manipuri, 2 Assamese | [102], [103], [104], [105], [107], [111], [112], [113], [114], [121], [122], [125], [127], [128] |
| 58. | U+09CC | ৌ | BENGALI VOWEL SIGN AU | Matra | 1 Bengali,2 Manipuri, 2 Assamese | [102], [103], [104], [105], [107], [111], [112], [113], [114], [121], [122], [125], [127], [128] |
| 59. | U+09CD | ্ | BENGALI SIGN BIRAMA | Halant/ Birama | 1 Bengali,2 Assamese2 Manipuri | [101], [102], [103], [104], [105], [107], [108], [109], [111], [112], [113], [114], [119], [120], [121], [122], [126], [128] |
| 60. | U+09CE | ৎ | BENGALI LETTER KHANDA TA | Consonant | 1 Bengali,2 Manipuri, 2 Assamese | [101], [102], [103], [104], [105], [107], [111], [112], [113], [114], [119], [120], [121], [125], [127] |
| 61. | 09F0 | ৰ | BENGALI LETTER RA WITH MIDDLE DIAGONAL | Consonant | 2 Assamese | [102], [103], [111], [121], [122], [124], [126], [128] |
| 62. | 09F1 | ৱ | BENGALI LETTER RA WITH LOWER DIAGONAL | Consonant | 2 Assamese2 Manipuri | [102], [103], [111], [121], [122], [124], [125], [126], [127], [128] |

Table 8: Bengali Code-Point Repertoire

Apart from the above individual code-points, the Neo-Brahmi Generation Panel also proposes some specific sequences which enable conditional inclusion of the "Bengali LETTER A and E" followed by Bengali SIGN VIRAMA and Bengali LETTER YA again followed by Bengali VOWEL SIGN AA in the repertoire for enabling inclusion of /æ/ sound as in English ‘bat’, ‘cat’ etc.

| **Sr. No.** | **Unicode Code Points** | **Sequence** | **Character Names** | **Example languages using the code-point (Not exhaustive list)** | **Reference** |
| --- | --- | --- | --- | --- | --- |
| S1. | 098509CD09AF09BE | অ্যা | BENGALI LETTER ABENGALI SIGN VIRAMABENGALI LETTER YABENGALI VOWEL SIGN AA | Bengali, Assamese | [102], [103], [104], [105], [107], [111], [112], [113], [114], [121], [122], [125], [127], [128] |
| S2. | 098F09CD09AF09BE | এ্য়া | BENGALI LETTER EBENGALI SIGN VIRAMABENGALI LETTER YABENGALI VOWEL SIGN AA | Bengali | [102], [103], [104], [105], [107], [111], [112], [113], [114], [121], [122], [125], [127], [128] |

Table 9: Sequences

## Code Point Repertoire Exclusion

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Sr. No.** | **Code Points** | **Glyph** | **Character Names** | **Note** |
| 1. | U+098C | ঌ | BENGALI LETTER VOCALIC L | Limited or declining use  |
| 2. | U+09D7 | ৗ | BENGALI AU LENGTH MARK | Limited or declining use  |

Table 10: Excluded Code Points

## The Basis of Present IDN

The present LGR has also benefited from the earlier work on IDN for Bengali (different versions) done for .भारत or .ভারত drafted between 20.11.2009 and 18.07.2013.

### The ABNF Variables

The Augmented Backus-Naur Formalism (ABNF) began with the following variables:

C → Consonant

V → Vowel

M → Matra

B → Anusvara

D → Chandrabindu

X → Visarga

H → Halant /Virama

N → Nukta

Y → Avagraha

Z → Khanda Ta

The Augmented Backus-Naur Formalism (ABNF) will use the following Operators:

|  |  |  |
| --- | --- | --- |
| Sr. Number | Operator | Function |
| 1 | “ | “ | Alternative |
| 2 | “[ ]” | Optional |
| 3 | “ \* ” | Variable Repetition |
| 4 | “( )” | Sequence Group |

Table 11: The ABNF Formalism

## In what follows, the Vowel Sequence and the Consonant Sequence pertinent to Bengali are given to facilitate understanding.

### The Vowel Sequence

In what follows, the Vowel Sequence and the Consonant Sequence pertinent to Bengali are given. To facilitate understanding, equivalents in Devanagari are provided.

A vowel sequence is made up of a single vowel. It may be followed but not necessarily (optionally) by an Anusvara (B), Candrabindu (D) or a Visarga (X). The number of D, B or X which can follow a V in Bengali may not be restricted to one.

The possibility of a Visarga or Anusvara following a Candrabindu exists in Bengali. Vowel can optionally be followed by a combination of Halant/ Virama [H], Consonant [C] to form a Ya-phala. “Ya-phala is a presentation

form of U+09AF Bengali letter য or ‘ya’. Represented by the sequence < U+09CD, i.e. ্ Bengali SIGN VIRAMA, U+09AF -Bengali LETTER য ya>, ya-phala has a special form: য়. Again, when combined with U+09BE া , i.e. Bengali VOWEL SIGN for ‘aa’(ā), it is used for transcribing [æ] as in the “a” in the English word “bat” written in Bengali as ব্যাট.

A Vowel-sequence admits the following combinations:

### A Single Vowel

Examples: V অ अ

### A Vowel with Conditions

A Vowel can optionally be followed by Anusvara [B] or Candrabindu [D] or Visarga [X] or Candrabindu+Anusvara [DB] or Candrabindu+Visarga [DX] or combination of Halant (also known as Virama) [H] followed by Consonant [C] followed by Matra [M].

Examples:

VB অং अं

 VD অঁ अँ

VX অঃ अः

VDB অঁং अँंं

VDX অঁঃ अँंः

VHCM অ্যা /এ্যা

### VHCM Sequence

A VHCM sequence can optionally be followed by Anusvara [B] or Candrabindu [D] or Visarga [X] or Chandrabindu+Anusvara [DB] or Candrabindu+Visarga [DX].

Examples:

 VHCMB অ্য়াং/এ্যাং

 VHCMD অ্য়াঁ/এ্যাঁ

 VHCMX অ্য়াঃ/এ্যাঃ

 VHCMDB অ্য়াঁং/এ্যাঁং

 VHCMDX অ্য়াঁঃ/এ্যাঁঃ

### The Consonant Sequence

### A Single Consonant (C)

 Example: C ক क

### A Consonant with Conditions

A Consonant optionally followed by dependent vowel sign / Matra [M] or Anusvara [B] or Candrabindu [D] or Visarga [X] or Halant (also known as Virama) [H] or Candrabindu+Anusvara [DB] or Candrabindu+Visarga [DX]

Example:

 CM কি/ কৃ कि/ कृ

 CB কং कं

 CD কঁ कँ

 CX কঃ कः

 CH ক্ क् (Pure consonant)

 CDB কঁং कँं ं

 CDX কঁঃ कँः

### CM Sequence

A CM sequence can be optionally followed by B, D, X, DB or DX.

Example:

 CMB কীং/ কৃং कीं/ कृं

 CMD কাঁ काँ

 CMX বীঃ वीः

 CMDB কাঁং काँ ं

 CMDX কাঁঃ काँः

### Sequence of Consonants

A sequence of consonants (up to 4) joined by Halant (also known as Virama).

\*3(CH)C

Example:

 CHC ন্ত → ন্ + ত न् + त

 CHCHC ন্ত্র → ন্ + ত্ + র न् + त् + र

 CHCHCHC ন্ত্র্য় → ন্ + ত্ + র্ + য় न् + त् + र् + य

### Subsets:

While considering its subsets, as a representative example, we will consider the combination CHC only, however the same is equally applicable to CHCHC and CHCHCHC.

**[A].** The combination may be followed by M, B, D, X, DB or DX.

Example:

 CHCM ক্কী →ক ্ ক ী क्की → क ् क ী

 CHCB ক্কং →ক ্ ক ং क्कं → क ् क ंं

CHCD ক্কঁ →ক ্ ক ঁ क्कँ→ क ् क ंँ

CHCX ক্কঃ →ক ্ ক ঃ क्कः → क ् क ঃ

CHCDB ক্কঁ ং →ক ্ ক ঁ ং क्कँंं→ क ् क ंँ ं

CHCDX ক্কঁঃ →ক ্ ক ঁ ঃ क्कँंः→ क ् क ंँ ः

**[B].** \*3(CH)CM may further be followed by a B, D, X, DB or DX

Example:

CHCMB ক্কীং → ক ্ ক ী ং क्कीं → क ् क ी ं

 ক্কৃং → ক ্ ক ৃ ং क्कृं → क ् क ृ ं

CHCMD ক্কাঁ → ক ্ ক া ঁ क्काँ → क ् क ा ंँ

CHCMX ক্কীঃ → ক ্ ক ী ঃ क्कीः → क ् क ी ः

CHCMDB ক্কাঁং→ ক ্ ক া ঁ ং क्काँ→ क ् क ा ँ ं

CHCMDX ক্কাঁঃ → ক ্ ক া ঁ ঃ क्काँः → क ् क ा ंँ ः

### A single ‘Khanda’-Ta (Z)

 Example: Z ৎ = ত্

### A Khanda Ta Combination

A Khanda Ta can be preceded by a consonant and Halant (also known as Virama)

[CH]Z

Example:

 র + ্ + ৎ = র্ৎ as in ভর্ৎসনা

### Special Cases S and P

Two special cases involving Sequences (referred to as S and P in Table 16 under Section 7) could be described briefly here. Let us take up S at the first instance. It is noteworthy that there are two instances in Bengali where halant is preceded by a full vowel (U+0985 অ - BENGALI LETTER A and U+098F এ - BENGALI LETTER E). For rendering *ya-phala* followed by অ and এ , it is necessary to type U+09CD halant plus U+09AF *ja* preceded by the said vowels. This is a purely ligatural entity and the addition of *ya-phala* and ā matra is used to elicit the /æ/ sound as in English ‘bat’, ‘fat’ etc. The Brahmi script, by nature does not have halant after a vowel. Halant is generally described as ‘vowel killer’, although it actually indicates absence of a vowel after the marked consonant. Only the consonants can have the halant marked. But as we see here, Bengali ends up with a deviant feature in the orthography here in which halant comes immediately after a vowel in ligatures অ্যা and এ্যা (Cf Unicode 10.0 p. 473 [100]).

Another case refers to the formation of *repha* and *ra-phala* in the said script and mentioned in the table above as P. Owing to co-occurrence with HALANT, RA either loses its own implicit vowel (REPHA), or suppresses the implicit vowel of the preceding consonant (RA-PHALA). For instance, ref = *ra* + halant+ C (e.g. র্ক i.e. *ra* + halant+ *ka*, as in অর্ক *arko*); *ra-phala*= C + halant+ *ra* (e.g. ক্র i.e. *ka* + halant+ *ra*, as in চক্র *chakro*). The point is in both the cases the slot for *ra* could be Bengali *ra* র (U+09B0) or the Assamese *ra* ৰ (U+09F0), followed/ preceded by the common halant (U+09CD), whereas the shapes of *repha* and *ra-phala* in both the cases remain the same.

# Variants

This section talks about the variants in the Bengali script. The NBGP categorizes these confusingly variants in two groups.

**Group 1:** Confusing due to pure visual similarity

**Group 2:** Confusing due to deviation from normally perceived character formations by larger linguistic community.

For Group 1, the identical code points are defined as variants. The confusable, but not identical, cases are not proposed, as there is another panel (String similarity assessment panel) entrusted to deal with such cases. However, cases which belong to Group 2 are proposed to be considered as variants. These cases are not of mere visual similarity as they involve some deviations from the widely accepted norms of Bengali Akshar formations. These can cause confusion even to a careful observer and hence being proposed as variants.

The variants are generated in a script when two or more forms are formed with different storage or code points. In Bengali the *e*-matra, *ā*-matra and the *o*-matra have different code points. One can type o with a consonant at one go and the same by typing *e*-matra and *ā*-matra as two separate keys getting the same results. A reader cannot differentiate between the two *ko* (কো), one typed with a single key and the other one typed with two different keys. Moreover, this will not be considered as a case of variant because a matra followed by a matra is not allowed.

On the other side, typing the character U+09B0 র one could be achieved either with the help of the single key (র) or by typing ব followed by nukta (ব+়) resulting in a similar shape as ব়. This could be mistaken for a variant because the র achieved with a single key has a different code point assigned to it in relation to the latter i.e. ব +nukta. This sequence of typing a nukta after ব could be blocked. A direct *ra* has the code value U+09B0. The nukta one is assigned to the code point U+09AC followed by U+09BC. Hence, it does not stand as an example of a variant.

Moreover, the other three most frequent nukta characters (U+09F0) ড়, (U+09F1) ঢ় and (U+09DF) য় do not fall under the category of variants although each of these can be typed in two different ways producing similar looking characters (ড়, ঢ়, য়), because they have been marked white in the MSR chart for Bengali.

## In Script Variants

**CASE I:**

As far as true variants in Bengali are concerned, we may draw our attention to cases wherein halant with (U+09A5) থ (*tha*) appears as conjunct with (U+09B8) স (*sa*) and (U+09A8) ন (*na*).

1. স + halant + থ (U+09B8 + U+09CD + U+09A5) versus
 স + halant + হ (U+09B8 + U+09CD + U+09B9)

2. ন + halant + থ (U+09A8 + U+09CD + U+09A5) versus
 ন + halant + হ (U+09A8 + U+09CD + U+09B9)

The above combinations, if written in traditional orthography, could be little confusing, where the থ (*tha*) in conjunct appears like a হ (*ha*). The conjunct could be in the initial, medial or final positions (as shown below in e.g. no 1). It could be typed wrong as well, thinking it was a হ (*ha)* U+09B9, increasing the chances of risks in label writing and identification.

Examples:

1. স্থ and স্হ (as in স্থান sthān, স্থূল sthulo, স্বাস্থ্য shāshtho, অস্থায়ি asthāyi)

2. ন্থ and ন্হ (as in গ্রন্থ grontho )

The fonts which represent traditional Bengali writing system could tend to create this problem. Therefore, these may be taken as cases of variants in Bengali.

**CASE II:**

Another interesting example of variant is encountered in *ra* + *halant* and *halant* + *ra* combinations in writing labels in the Bengali script ( for languages such as Bengali, Assamese and Manipuri). The variant cases arise in typing ‘***ref***’ (involving *ra* + *halant*) and ‘***ra-phala***’(involving *halant* + *ra*).

‘Ref’ could be formed by two sequences (mainly because both Assamese and Bengali find place in the same UNICODE points, and ‘B\_RA’ as well as ‘A\_RA’ refer to the same phonetic element). Here, the final ligatures look the same which will be as follows:

1. B\_RA + H + C
2. A\_RA + H + C

 Where

B\_RA = U+09B0 BENGALI LETTER RA (র) or

A\_RA = U+09F0 BENGALI LETTER RA WITH MIDDLE DIAGONAL (ৰ)

H = U+09CD BENGALI SIGN VIRAMA (্ )

C = any consonant (theoretically)

Example:

|  |  |  |  |
| --- | --- | --- | --- |
| Sequence1(Using Bengali RA) | Ligature1 | Sequence2(Using Assamese RA) | Ligature2 |
| U+09B0 (র) U+09CD (্)U+0995 (ক) | র্ক | U+09F0 (ৰ) U+09CD (্) U+0995 (ক) | ৰ্ক |
| U+09B0 (র) U+09CD (্)U+09A0 (ঠ) | র্ঠ | U+09F0 (ৰ) U+09CD (্) U+09A0 (ঠ) | ৰ্ঠ |

Table 12: Example of Ref

Note: As Bengali and Assamese ক and ঠ look exactly the same, the resultant combinations with 'Ref' look identical. Addition of 'Ref' does not make any difference.

‘Ra-phala’ could be formed by two sequences on similar grounds, and the final ligatures would look the same

1. C1 + H + B\_RA
2. C1 + H + A\_RA

 Where

C1 = any consonants except Khanda-ta

Example:

|  |  |  |  |
| --- | --- | --- | --- |
| Sequence1(Using Bengali RA) | Ligature1 | Sequence2(Using Assamese RA) | Ligature2 |
| U+0995 (ক) U+09CD (্) U+09B0 (র) | ক্র | U+0995 (ক) U+09CD (্) U+09F0 (ৰ) | ক্ৰ |
| U+09A8 (ন) U+09CD (্) U+09B0 (র) | ন্র | U+09A8 (ন) U+09CD (্) U+09F0 (ৰ) | ন্ৰ |

Table 13: Example of Ra-phala

As the Assamese and Bengali Ref and Ra-phala conjunct forms look the same, this could cause confusability to the end-users. Hence, the ref and ra-phala cases need to be defined as variants.

NBGP concluded to define র and ৰ as variant code points, where only one variant set between র and ৰ could cover all cases. But this will create block variant labels, e.g. if someone registers “র র র” the variant label “ৰৰৰ”will be generated as variant and will be blocked and vice versa . However it is only blocked at the label level, if someone else needs to register other labels e.g. ৰৰ or ৰৰৰৰ, it is still possible.

## Cross Script Variants

A crisp cross script study for Bengali has been done with respect to sister scripts such as Devanagari, Gurmukhi and Odia[[3]](#footnote-3) (formerly Oriya) keeping in mind the visual and technical confusions they may cause as labels on the web domain. Moreover, there is no in-script variant in Bengali as far as the orthography is concerned. The following characters are being proposed by the NBPG as variants. Although there are certain characters which are somewhat similar but have not been included here. They have been provided in the Appendix (10.2) for reference.

1. Bengali and Devanagari Script

|  |  |
| --- | --- |
| Bengali | Devanagari |
| মU+09AE | मU+092E |
| িU+09BF | िU+093F |

Table 14 - Bengali and Devanagari cross-script variant code point

1. Bengali and Gurmukhi Script

|  |  |
| --- | --- |
| Bengali | Gurmukhi |
| মU+09AE | ਸU+0A38 |
| িU+09BF | ਿU+0A3F |

Table 15 - Bengali and Devanagari cross-script variant code point

# Whole Label Evaluation Rules (WLE)

This section provides the WLEs that are required by all the languages mentioned in section ‎3.2 when written in Bengali[[4]](#footnote-4) Script. The rules have been drafted in such a way that they can be easily translated into the LGR specifications.

Below are the symbols used in the WLE rules, for each of the "Indic Syllabic Category" as mentioned in the table provided in Code point repertoire (Section 5.1).

|  |  |  |
| --- | --- | --- |
| C | → | Consonant |
| M | → | Matra |
| V | → | Vowel |
| B | → | Anusvara |
| D | → | Candrabindu |
| X | → | Visarga |
| H | → | Halant |
| N | → | Nukta |
| Z | → | Khanda Ta |
| S | → | S1, S2 (from Table 9)or(Aa/Ae) *ya-phala* (V1 H C1 M1)whereV1 is either 0985 (অ - BENGALI LETTER A)or 098F (এ - BENGALI LETTER E)H is 09CD (্ - BENGALI SIGN VIRAMA)C1 is - 09AF (য - BENGALI LETTER YA)M1 is - 09BE (া - BENGALI VOWEL SIGN AA) |
| P | → | Ra-Halant (C2 H)whereC2 is either 09B0 (র - BENGALI LETTER RA) or 09F0 (ৰ - ASSAMESE LETTER RA/ Unicode name: BENGALI LETTER RA WITH MIDDLE DIAGONAL )H is 09CD (্ - BENGALI SIGN VIRAMA)  |

Table 16 - Symbol used in WLE rules

It is also perhaps ideal to mention here that in Bengali, the consonant letters (or graphemes) are physically joined to form “clusters” that could theoretically conjoin from two to four consonants and combine to create new shapes. Dash and Chaudhuri (1998) state that there are “nearly 380 unique consonant...clusters” out of which Bi-consonantal combinations are 290, three-letter combinations account for another 80 and the rarer ones with four letters number 10 more [136, Pg 4]. More details of such combinations could be seen in Pabitra Sarkar (1993) [135].

## Final Set of WLE Rules

The prevalent patterns in Bengali, and various restrictions, below are the specific WLE rules that need to be implemented. The rule wise examples also provided, including both attested and hypothetical examples.

1. N: must be preceded only by one of a specific set of Cs

The specific Cs are:

a. ক (U+0995)

b. খ (U+0996)

c. গ (U+0997)

d. জ (U+099C)

e. ড (U+09A1)

f. ঢ (U+09A2)

g. ফ (U+09AB)

h. য (U+09AF)

 Example: ক়, খ়, গ়, জ়, ড়, ঢ়, ফ়, য়

2. H: must be preceded by C or N

Example: ক্, ড়্

1. M: must be preceded by C or N

Example: **কা, জ়া, ড়া**

1. D: must be preceded by either of V, C, N or M

Example: **আঁ, খঁ, ড়ঁ, খাঁ**

1. X: must be preceded by either of V, C, N, M or D

Example: খঃ, বঃ, ড়ঃ, াঃ, দুঃ

1. B: must be preceded by either of V, C, N, M or D

Example: আং, ইং, কং

1. Z: must be preceded by V, C, N, M, D, B, X, S or P
Example: ইৎ, কৎ, ড়ৎ, াৎ, াঁৎ, ংৎ, ঃৎ, এ্য়াৎ, অ্য়াৎ, র্পৎ, প্রৎ
2. V: CANNOT be preceded by H

Details in 7.1.1 Case of V preceded by H

Now let us elaborate each rule with examples from the script keeping in mind the Bengali, Assamese and Manipuri communities. Some combinations of characters may seem unrealistic or rare in usage but there is no harm is adding such ligatures because they are simply possible but not attested combinations. Others, such as the nukta characters have a mixed acceptance in the linguistic community. Whereas nukta characters such as ড় (U+09DC), ঢ় (U+09DD), য় (U+09DF) is common in Bengali, Assamese and Manipuri; জ (U+099C + 09BC) is mostly found in Bengali texts of Bangladesh and nowadays also being used in West Bengal also, particularly in some magazines and newspapers. On the other hand, characters such as ক়, খ়, and গ় are in use in rendering words of Arabic or Persian origin and of religious importance, mostly attributing to Islam. For example, many of these are found in Muslim names and in loan words written in Bengali in Bangladesh. The idea of this generalization is that these analogical inclusions do not necessarily violate linguistic or orthographic rules of the languages(s) and thus have been incorporated to complete the series (combination with other characters) to help computational and NLP tasks, the ultimate goal of which is to deter phishing and cheating on the net when Indian scripts get adopted for e-commerce and related activities. Hence, the combinations are included in the WLE rules. In short, these combinations are possible but not all are attested in the respective languages.

### Case of V Preceded by H:

There could be cases involving multi-word domains where V may need to be allowed to follow an H

e.g. ব্যাঙ্কঅফ্ইন্ডিয়া /bæŋk ʌv ɪndiə / (U+09AC U+09CD U+09AF U+09BE U+0999 U+09CD U+0995 U+0985 U+09AB U+09CD U+0987 U+09A8 U+09CD U+09A1 U+09BF U+09DF U+09BE ) (meaning: *Bank of India*)

This is the case where two different words are joined together first of which ends with a H (অফ্) and the second word begins with a V (ইন্ডিয়া). Some sections of the linguistic community require the explicit presence of H for full representation of the sound intended. However, by and large, the form of the first word without a H is considered enough for full representation of the sound intended for the first word.

This is a unique situation necessitated by the lack of hyphen, space or the Zero Width Non-joiner character in the permissible set of characters in the Root zone repertoire. Otherwise, V is never required to be allowed to follow an H. Permitting this may create a perceptive similarity among two labels (with and without H) for majority of the linguistic community, hence this is explicitly prohibited by the NBGP.

In future if required, depending on the prevailing requirements by the community, the future NBGP may consider revisiting this rule.

## Additional Examples from Bengali ABNF:

Below are some of the examples which help one understand some of the rules ABNF puts in place. These are just given for reference purposes and are not meant to be comprehensive.

1. H, M, B, D or X cannot occur in the beginning of a Bengali IDN. Example:

्क ্ক

ाक াক

ंंक ংক

ंँक ঁক

ःक ঃক

As can be seen such combination will result automatically in a “golu” marking it as an invalid formation. This is an intrinsic property of the Indian language syllable and is quasi automatically applied wherever supported by the OS.

2. H is not permitted after V, B, D, X or M

Example:

अ् অ্

कं ् অং্

कँ ् কঁ্

कः् কঃ্

कि् কি্

1् 1্

 -् - ্

3. Number of B, D or X permitted after Consonant or Vowel or a Matra is restricted to one thus the following combinations are invalidated.

Example:

कं ंं কংং

कँ ंँ কঁ ঁ

कःः কঃঃ

काँ ँ কাঁ ঁ

कीःः কীঃঃ

अंंं অংং

अँ ँ অঁ ঁ

अःः অঃঃ

4. Number of M permitted after Consonant is restricted to one.

Example:

कीी কীী

5. M is not permitted after V.

 Example:

ईा/ ईौ ইা/ ঈৌ

6. The combinations of Anusvara+Visarga as well as Visarga+Anusvara are not permissible.

Example:

कं ः কংঃ

कः ं কঃং

# Contributors

Professor Udaya Narayana Singh, Chair-Professor & Head, Amity Centre for Linguistic Studies (ACLiS), Amity University Haryana, Gurgaon; Pachgaon, Manesar PIN 122431 (Haryana), India

Professor Pabitra Sarkar, formerly Vice-Chancellor, Rabindra Bharati University, Kolkata

Dr Atiur Rahman Khan, Principal Technical Officer, GIST Group, C-DAC, Pune, PIN 411008 (Maharashtra), India.

Mr Akshat Joshi, Project Engineer, GIST Group, C-DAC, Pune, PIN 411008 (Maharashtra), India.

Ms Moumita Chowdhury, Senior Technical Officer, GIST Group, C-DAC, Pune, PIN 411008 (Maharashtra), India.

Mr. Rajib Chakraborty, Linguist, Society for Natural Language Technology Research (SNLTR), Module 114 & 130, SDF Building, Salt Lake, Sector-V, Kolkata-700091 (West Bengal), India

# References

[100] Unicode Consortium. 2017. Unicode Standard 10.0. Mountain View CA.

[101] Bandyopadhyay, Chittaranjan. 1981. *Dui Shataker Bangla Mudran o Prakashan.* Kolkata: Ananda Publishers.

[102] Banerji, R.D. 1919. *The Origin of the Bengali Script*. Kolkata. New Delhi; Asian Educational Services; 2003 reprint.

[103] Chatterji, S.K. 1926. *The Origin and Development of the Bengali Language*. Calcutta: Calcutta University Press.

[104] -----. 1939. *Bhasha-prakash Bangala Vyakaran (A Grammar of the Bengali Language)*, Calcutta: University of Calcutta

[105] Hai, Muhammad Abdul. 1964. *Dhvani Vijnan O Bangla Dhvani-tattwa (Phonetics and Bengali Phonology)*, Dhaka: Bangla Academy

[106] Jha, Subhadra. 1958. *The Formation of Maithili*. London: Luzac & Co.

[107] Kostic, Djordje; Das, Rhea S. 1972. *A Short Outline of Bengali Phonetics*, Calcutta: Statistical Publishing Company.

[108] Majumdar, R.C. 1971. *History of Ancient Bengal*, Calcutta: G. Bhardwaj.

[109] Mazumdar, Bijaychandra. 1920/2000. *The History of the Bengali Language* (Repr. Calcutta, 1920. ed.). New Delhi: Asian Educational Services.

[110] Pandey, Anshuman. 2001. Proposal to Encode the Tirhuta Script in ISO/IEC 10646.

[111] Pal, Palash Baran. 2001. *Dhwanimala Barnamala.* Kolkata: Papyrus.

[112] -----. 2007. ‘Bangla Harapher Panch Parba’*.* In Swapan Chakraborty, ed.  *Mudraner Sanskriti O Bangla Boi.* Kolkata: Ababhas.

[113] Ross, Fiona. 1999. *The Printed Bengali Character and its Evolution.* London: Curzon.

[114] Shastri, Mahamahopadhyay Hara Prasad. 1916. *Hājār Bacharēr Purāṇa Bāṅgālā Bhāṣāy Bauddha Gān ō Dōhā.* Calcutta: Bangiya Sahitya Parishat.

[115] Singh, Udaya Narayana (Jointly Maniruzzaman). 1983. *Diglossia in Bangladesh and language planning*. Calcutta : Gyan Bharati. 214 pp.

[116] -----. 1987. *A Bibliography of Bengali Linguistics*. Mysore : CIIL. xii+316 pp.

[117] -----. 2017. (with Rajib Chakraborty, Bidisha Bhattacharjee & Arimardan Kumar Tripathy) *Languages and Cultures on the Margin: Guidelines for Fieldwork on Endangered Languages*. Mimeo. Centre for Endangered Languages, Visva-Bharati.

[118]-----. 1980. Scriptal choice and spelling reform : An essay in language and planning. *Journal of the M.S. University of Baroda*, Social Science Number, 29.2 : 173-186. A modified ver­sion reprinted E. Annamalai, Bjorn Jernudd and Joan Rubin, eds. *Language Planning: Proceedings of an Institute*. Mysore: CIIL. 405-417.

[119] Sripantha. 1996. *Jakhan Chapakhana Elo.*  Kolkata: Paschim-Banga Bangla Academy.

[120] Sur, Atul. 1986. *Bangla Mudraner Dusho Bachar.* Kolkata: Jijnasa.

[121] Script Behaviour for Bengali, Version 1.1, TDIL and C-DAC Pune.

[122] Bora, Mahendra. 1981. *The Evolution of Assamese Script*. Jorhat: Assam Sahitya Sabha.

[123] <http://www.unicode.org/L2/L2011/11175r-tirhuta.pdf>

[124] <https://www.ethnologue.com/cloud/asm>

[125] <https://www.omniglot.com/writing/manipuri.htm>

[126] <https://en.wikipedia.org/wiki/Bengali_alphabet>

[127] <http://www.iitg.ernet.in/rcilts/phaseI/manipuridesign.pdf>

[128] <http://www.iitg.ernet.in/rcilts/phaseI/newassamesedesign.pdf>

[129] <http://www.omniglot.com/writing/syloti.htm>

[130] <http://www.bengaliandsylheti.com/bnscriptevol.htm#.WhwWAEqWbIU>

[131] <https://en.wikipedia.org/wiki/Bishnupriya_Manipuri_language>

### [132] <http://metashare.elda.org/repository/browse/the-emilleciil-corpus/abdd35c8de6f11e2b1e400259011f6ea6bce74d38dbb42d881da76c64a6adb20/> or, <http://catalog.elra.info/product_info.php?products_id=696>

 [133] <https://www.isical.ac.in/~rc_bangla/bangla.html>

[134] Sarkar, Pabitra. 1992. *Bangla Banan Sanskar: Samasya o Sambhabana*. Kolkata: Chirayata Prakashan.

[135] Sarkar, Pabitra. 1993. Bangla Bhashar Yuktabyanjan. *Bhasha* 1.1: 23-45.

[136] Dash, Niladri Shekhar and B.B. Chaudhuri. 1998. Bangla Script: A Structural Study. *Linguistics Today* 1.2: 1-28. Also available at https://www.academia.edu/9967428/Bangla\_Script\_A\_Structural\_Study

[137] Dani, Ahmed Hasan. (1957) ‘Srīhaţţa-Nāgarī Lipir Utpatti o Bikāś.’ Bangla Academy Patrika (Dhaka), Vol 1.2. (Bhadra-Agrahayan, 1364 Bangabda Number).Pg 1.

[138] [https://en.wikipedia.org/wiki/Sylheti\_Nagari accessed on 19.5.2018](https://en.wikipedia.org/wiki/Sylheti_Nagari%20accessed%20on%2019.5.2018).

[139] Furui, Ryosuke. (2015). ‘*Variegated Adaptations: State Formation in Bengal from the Fifth to Seventh Century’*, in Bhairabi Prasad Sahu & Hermann Kulke, eds. *Interrogating Political Systems: Integrative Processes and States in Pre-Modern India.* Chapter 9. Pp 255-73. New Delhi: Manohar.

[140] Ferguson, Chares A. and Munier Chowdhury. (1960) ‘Phones of Bengali’, *Language*, Vol. 36, No. 1, pp. 22-59.

[141] Shahidullah, Muhammad. (2007) *Buddhist Mystic Songs*. Dhaka: Mowla Brothers.

[142] Ray, Punya Sloka. (1966) *Bengali Language Handbook*. Washington.

[143] Hai, Muhammad Abdul. (1960) *A phonetic and phonological study of nasals and nasalization in Bengali*. Dhaka: University of Dhaka.

# [145] <https://www.unicode.org/L2/L2002/02387r-syloti-form.pdf> accessed on May 21, 2018

# [146] [https://en.wikipedia.org/wiki/Ol\_Chiki\_(Unicode\_block)](https://en.wikipedia.org/wiki/Ol_Chiki_%28Unicode_block%29) accessed on May 21, 2018

# [147] <http://www.bangladesh2000.com/bd/bangla_script.html> accessed on May 21, 2018

[148]<http://www.bengaliandsylheti.com/bnscriptevol.htm#.WhwWAEqWbIU>

[149] Bhattacharya, Ashutosh ed. (1942) *Gopichandrer Gan*, Calcutta: Calcutta University.

[150] Das, Sisir Kumar. (1975) *Sahibs and Munshis: An Account of the College of Fort William*. Calcutta.

[151] Islam, Rafiqul, Pabitra Sarkar, Mahbubul Haq & Rajib Chakraborty (eds.). (2014) *Bangla Academy Promito Bangla Byabaharik Byakaran (A Functional Grammar of Standard Bangla)*. Dhaka: Bangla Academy.

[152] Sarkar, Pabitra. [2013] ‘Bangla Spelling Reform: the Long and Short of It’. *Bangla Journal* 19: 215-232.

[153] Bangla Academy. (2012) *Bangla Academy Promito Bangla Bananer Niyam (Standard Bangla Spelling as adopted by Bangla Academy)*. Dhaka: Bangla Academy.

# Appendix

## Augmented Backus Naur Formalism (ABNF)

The Augmented Backus Naur Formalism (ABNF) is generic in nature and when applied to a specific language/script, certain restriction rules apply. In other words, in a given language some of the Formalism structures do not necessarily apply. To take care of such cases restriction rules are set in place. These restrictions will help to fine-tune the ABNF.

In case of Bengali the following rules apply:

1. *Khanda ta* (ৎ) is NOT allowed at the beginning of an IDN label. The same applies to ঞ and the velar nasal ঙ in the Bengali Scheme of five-fold ‘*barga*’ (as defined under Table 5). Moreover, Bengali does not allow *ya* (য়) in the beginning of a word either (but can have a name, mostly of foreign origin such as Yaqub which may be written with ya (য়) in the beginning as in য়াকুব). In very recent times, while transliterating some Chinese and Japanese names in Bangla, one does come across the possibility of *Khanda* *ta* (ৎ) followed by *sa* (স) in the beginning of a word, for example **ত্সেরিং (Tsering)**.

2. CH can come with Khanda Ta in only the case where C is *ra* (র) (09B0).

র্ৎ as in ভর্ৎসনা

3. Nukta shall be allowed only after following characters: ড (ড়), ঢ (ঢ়) and জ (জ়), are characters which allow nukta after them. Apart from these there are some more characters which may admit nukta for special usage and specific linguistic requirement within the speech community. They are ক (ক়), খ (খ়), গ (গ়) and ফ (ফ়). These graphemic extensions could be used to write, for better pronunciation of, words derived from Persian and Arabic in particular, besides being used for any other borrowed word having similar pronunciation.

4. Only following combinations with VHCM will be allowed.

→ অ্যা (together pronounced as æ) as in অ্যাসিড (acid)

→ এ্যা (together also pronounced as æ) as in এ্যাসিড, এ্যাসোসিয়েশান

 (acid, association)

5. A consonant sequence that is intended to end with Halant [H] can only be followed by an Avagraha [Y]. Thus following combinations are permissible.

 আল্লাহ্- শাহ্

**Cross-script Confusable**

6. The number of consecutive identical consonants joined by a Halant within a label shall not exceed two. Thus (ka+halant+ka) is permitted but not (ka+halant+ka+halant+ka).

 ক্ক is permitted, but not something like this ক্ক্ক

7. A label containing not more than three "akshara", which have got variants shall be permitted. As an example let us consider a, b, c and d as four aksharas in a given label having a', b', c' and d' as variants in which case such a label will be disallowed. (Example of disallowed label - abcd, acdb, cdaba and so on)

##  ‘Sylheti Nagari lipi’ or ‘Siloţi’

This version of Bengali script resembles the ‘Kaithī’ script (ISO 12954) used by the Accountants (perhaps by the Kāyastha community) in Eastern Uttar Pradesh and Bihar – widely in use during the 1880s. There were several other names of Sylheti Nagari or Siloti – such as The script has also been known as ‘Jalalabad Nagari’, ‘Fūl (flower) Nagari’, ‘Muslim Nagari’, or ‘Muhammad Nagari’. It is said that Shah Jalal had brought the script with him in 13th-14th Century in Sylhet (138), although some suggested that it was an invention by the Afghan rulers of Sylhet (137). Some ascribe the credit to the Buddhist Bhikkhus from Nepal. Purely for historical reasons, the details of the script with 32 symbols are reproduced here (138):



Table 17 – The Script Table of Sylheti Nagari or Siloti

## Confusable code points

The following code points were analysed and concluded that they are either (a) distinguishable or (b) confusable but not enough to be defined as variant code points.

### Bengali and Devanagari

|  |  |  |
| --- | --- | --- |
| **Bengali** | **Devanagari** | **NBGP Decision** |
| ও U+0993 | उ U+0909 | Confusable |
| ঘ U+0998 | घ U+0918 | Confusable |
| ঁ U+0981 | ॅ U+0945 | Confusable |

Table 18: Bengali and Devanagari confusable code points

### Bengali and Gurmukhi

|  |  |  |
| --- | --- | --- |
| **Bengali** | **Gurmukhi** | **NBGP decision** |
| ঘ U+0998 | ਬ U+0A2C | Confusable |
| ঁ U+0981 | ੱ U+0A71 | Confusable |

Table 19: Bengali and Gurmukhi confusable code points

|  |  |  |
| --- | --- | --- |
| **Bengali** | **Gurmukhi** | **NBGP decision** |
| ও U+0993 | ਤ U+0A24 | Distinguishable |
| শ, শ U+09B6 | ਅ U+0A05 | Distinguishable |
| ম U+09AE | ਮ U+0A2E | Distinguishable |
| বা U+09AC and U+09BE | ਗ U+0A17 | Distinguishable |

Table 20 – Bengali and Gurmukhi distinguishable code points

### Bengali and Oriya

|  |  |  |
| --- | --- | --- |
| **Bengali** | **Oriya** | **NBGP Decision** |
| ও U+0993 | ଓ U+0B13 | Confusable |

Table 21 – Bengali and Oriya distinguishable code points

|  |  |  |
| --- | --- | --- |
| **Bengali** | **Oriya** | **NBGP Decision** |
| ঘ U+0998 | ସ U+0B38 | Distinguishable |

Table 22 – Bengali and Oriya distinguishable code points

1. The LGR follows the convention in the Unicode Standard for the name of this combined Bengali-Assamese-Manipuri Script and its code points. [↑](#footnote-ref-1)
2. This document needs to be printed in color for this to be read correctly. [↑](#footnote-ref-2)
3. Unicode uses Oriya for the script, although Odia is now the official term used. [↑](#footnote-ref-3)
4. As used by the Unicode, denoting and including both Assamese and Manipuri. [↑](#footnote-ref-4)