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

*LGR Version:* 3.0

*Date:* 2018-01-19

*Document version:* 0.1

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

# General Information/ Overview/ Abstract

This document lays down the Label Generation Rule Set (LGR) for the Bangla or Bengali script under the general rubric of the Neo-Brahmi Writing System. Three main components of the Bangla 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 "Proposed-LGR-Bengali-yyyymmdd.xml".

Labels for testing can be found in the accompanying text document:

Labels-SSSS-xxx-YYYYMMDD.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 and Principal Languages Using It

## 3.1 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 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 called Bangla or Bengali 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, one could consider Suniti Kumar Chatterji’s (1926: 16) suggestion of dividing the Indo-Aryan Speech into three broad periods considering the main phonetic 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 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 be 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 ‘Apabhramṁ̰śa’ and ‘Avahaṭṭha’.

(iii) The third stage is known as the *New Indo-Aryan* (NIA), starting from 1000 AD, when the total character of the language was altered and the vernaculars of modern Indo Aryan began to spring up. Bangla is said to have been evolved from ‘Māgadhī Apabhramṁ̰śa-Avahaṭṭha’ along with Asamiya (or Assamese), Odia, Magahi, Maithili, and Bhojpuri. Bangla belongs to the earlier group of the Magadhan sub-family along with Asamiya and Odia.

Bangla or 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 Apabhramṁ̰śa or ‘Avahaṭṭha’ except for small inscriptions and the manuscripts of the Tantric Buddhist text titled ‘Caryācaryaviniścaya’ or the Caryā-Pada (Shastri 1916) dating back to the 9th-11th century. As a result, there are not many epigraphical evidences as to the development of the writing system. However, whatever evidences of the genesis of Bangla writing system is available is discussed in the section 3.1.

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

(i) Firstly, Old Bangla 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 *Caryācaryaviniścaya* , the *Dohākōśa* of Saraha and the *Dohā kōśa* of Kanha (mostly in Apabhraṃśa), and the *Ḍākarṇava* (in a variety of Prakrit), (b) Old Bangla specimens of over 300 words in a commentary on the *Amara-kōśa* dated 1159 AD, and finally (c) the *Rāma Carita* of Sandyakara Nandi, of 11th century AD attesting some place names.

(i) Then there is Middle Bangla Period - 1200-1800 AD, again divided into three stages: (a) Transitional Middle Bangla (1200-1300 A.D): No genuine specimens but only he legends of Gopīcanda, Behula-Lakhindar, Khullana-Dhanapati, Phullara-Kālketu, Lausena, etc. (b) Early Middle Bangla (1300- 1500 A.D) with classics such as the *Śrī-Kṛṣṇa-kīrttana* of Caṇḍidāsa or the Bengali *Rāmāyaṇa* of Kṛttivāsa. (c) Late Middle Bangla (1500-1800 A.D): This period is attested by the development of Vaisnava literature with the influence of Sree Chaitanya Deva and his disciples.

(iii) Finally, after 1800 AD, we find the Modern or New Bangla, marked by the introduction of written prose (Mazumdar 1920) where the colloquial variety of Calcutta made its first appearance through the *Hutōm Pēcar Naksā* (1862) and in the books of Fort William College, Christian missionaries, or in the works of Raja Ram Mohan Roy, Ishwarchandra Vidyasagar, Bankimchandra Chattopadhyay, Michael Madhusudan Datta, Rabindranath Tagore, and Sarat Chandra Chattopadhyay. The influence of English in vocabulary, idioms, expressions as well as in the writing styles were significant. The fonts and types developed during this time also spread to all parts of Bangla speech community (Cf. Bandyopadhyay 1981, Sur 1986).

Bengali had developed two literary styles during the 19th-20th Century: The *Sādhubhāshā* (সাধুভাষা - "Elegant Language or Style") and the *Calitabhāshā* (চলিতভাষা "Current Language, or Modern Style"). The former is the traditional literary style based on Middle Bengali of the sixteenth century, while the later is a 20th century creation and is based on the speech of educated people in and around Calcutta or Kolkata Cf. Singh & Maniruzzaman 1983). It is the latter style that is prevalent today. With the latter style came many spelling and script reforms.

## 3.2 Written Bengali

The Bengali alphabet (বাংলা লিপি - Bangla lipi) is derived from the Brāhmī writing system, which is closely related to the Devanagari script. Considered to be fifth most widely used writing system in the world, Bangla Script with some variation for Assamese and Meitei or Bishnupuriya Manipuri was used in the eastern Indian Sanskrit manuscripts too. It was once used also for Bodo and Santali as well. For Chakma in India and Bangladesh and for Kokborok in Tripura, it was one of the scripts used. A close variant, called *Tirhutā* was used for Maithili until 1900. 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), this writing system got a shape comparable to the modern-day ones (Pal, 1981; Sripantha 1986). A pictorial description of Brāhmī to Modern Bengali could be presented here in a tabular form:



|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Modern | **ক** | **জ** | **ম** | **র** | **স**  | **অ** |

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 (1971) shows, in his *History of Ancient Bengal*, these evidences could be seen in both West Bengal and Bangladesh by 1) The Mahasthangarh (Bogra district, Bangladesh) 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 (7th Century A.D., later half, especially the Tippera copper plate of the reign of Lokanātha, the Kailan inscription of Sridharana Rāta and the Astafpur copper plates) demonstrate remarkable developments in Eastern Brāhmī. 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 *Kutila-lipi* (literally, ‘Cursive writing style’), whereas the term *Siddhamātika* was used by Al Beruni 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 Pūrvadesā 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 Mahāmānḍalika 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 develop a tick at the right end of the upper horizontal bar above and ā curved hook below. Initial e approaches the modern Bengali character. A Mature forms 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 (Chatterji 1939).

The current printed form of Bengali alphabet first appeared in 1778 when Charles Wilkins developed printing in Bengali. A few archaic letters were modernised during the 19th century. It was standardized by Pandit Ishwar Chandra Vidyasagar (1820-91) when the Bengali type fonts were to be used to publish on a large scale under the Calcutta School Book Society. All these could be presented in a table.

|  |  |  |
| --- | --- | --- |
| **PERIOD** | **DESCRIPTION** | **NAMES** |
| 3rd Millenium 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 Kharoshtī scripts begin in the subcontinent. Brāhmī was widely used during the Mauryan King, Ashoka. 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 *Kutila-lipi* | Kutila-lipi |
| 8th Century AD | Some copper plate inscriptions are found in the Khalimpur, Bangladesh during the reign of Dharmapāla, from Monghyr and Nalanda in Bihar, of the time of Devapāla, and from Jagjibanpur in West Bengal of the reign of Mahendrapāla. | *Siddhamātika* |
| 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 Mahāmānḍalika 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 Vijayamānikya-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’, Anuswara 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 or bangla Writing System

The overall development of Bengali Script from the Kutila-lipi period to Modern Bengali could be seen here in Table 3 (Cf. Banerjee, R.D. 1919. *The Origin of Bangla Script*. Calcutta: University of Calcutta and also see the web-page http://www.bengaliandsylheti.com/bnscriptevol.htm#.WhwWAEqWbIU)



Table 3: Bengali Script in Different Centuries

## 3.3 Languages Considered

Below is the tabular representation of the languages using Bangla script that are placed on EGDIS Scale 1-6 (See Singh et al 2017 for details). Some languages under EFDIS 5 and 6 have also developed their own scripts for printing and publishing. Some had used Bangla 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 Santali in both Devanagari and *Ol-chiki*. For the purposes of the Bengali LGR, as per the requirement of the LGR procedure, 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** | **EGDIS****Scale 5** | **EGDIS 6** |
| Bengali |  |  |  | Santali, Bodo,Riang, Khumi, Mru(ng), Asho | LepchaPnar, Koda/ Kora, Chak |
|  | Assamese  |   |   | Koch or Rajbangshi | Malto or Malpahariya |
|  | Manipuri or Meitei |   | Bishnupuriya (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 Bangla Script on the EGDIS Scale

## 3.4 Notable features

* The Bengali alphabet is a syllabic alphabet in which consonants all have an inherent vowel which has two different pronunciations, the choice of which is not always easy to determine and which is sometimes not pronounced at all.
* Vowels can be written as independent letters, or by using a variety of diacritical marks which are written above, below, before or after the consonant they belong to.
* All Bengali consonants are pronounced with an inherent vowel - / ɔ/ that corresponds to the Devanagari Schwa /ə/
* When consonants occur together in clusters, special conjunct letters are used. The letters for the consonants other than the final one in the group are reduced. The inherent vowel only applies to the final consonant.
* The Bangla script has forty-four symbols or graphemes representing thirty-five phonemes or functional speech sounds, with some obvious redundancies..
* In Bangla, several symbols or graphemes have secondary shapes, technically called the ‘allographs’ with a complementary distribution in each case. These variations or allographs are generally added to the following positions of the primary graphemes.

 1) Below
 2) Above
 3) Right side
 4) Left side
 5) Left Side and above simultaneously
 6) Right side and above simultaneously
 7) Right side and left side simultaneously
 8) Right side, left side and above simultaneously.

* Besides some simple Vowel Modifiers or ‘Matra’s there are some combinatory modifiers of Bengali Vowels with certain consonants. For example, whereas আ is substituted by া, ই is substituted by pre-posed ি, ঈ is substituted by ী or উ is substituted by ু 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 গ + ু
 রু ru, rather than writing as র + ু
 শু śu, rather than writing as শ + ু

 হু hu, rather than writing as হ + ু

* The global Bengali 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.

### 3.4.1 The Consonants

As per traditional classification Bengali Consonants are categorized according to their phonetic properties (especially in terms of place plus manner of articulation). There are Five ‘Barga’ or Groups (classes) in terms of Place of Articulation, and one Non-Barga group (Hai 1964). Each Barga, which corresponds to Stops, contains a series of five consonants classified as per their phonetic qualities, beginning from Unvoiced and Unaspirated to Voiced Aspirates (in the fourth column), finally ending with a Homo-organic or Corresponding nasal (Kostic & Das 1981). Consider the following table:

|  |  |  |  |
| --- | --- | --- | --- |
| **Barga** | **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 6: Barga classification of Bengali consonants

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Non-****Barga** | য ‘y’U+09AF | র ‘r’ U+09B0 | ল ‘l’U+09B2 | श ‘Sh’U+09B6 | ष ‘Ss’U+09B7 | स ‘s’U+09B8 | ह ‘h’U+0939 |

Table 7: Non-Barga consonants

### 3.4.2 The Implicit Vowel Killer: Hasanta (=’Halant’ in Devanagari)

As stated earlier, all consonants have an implicit vowel (central back /-ɔ/ in Bengali as the neutral vowel, corresponding to Devanagari schwa) within them. Unicode (cf. Unicode 3.0 and above) prefers the term ‘*Virāma*’. In this report both these terms have been used to denote the character that suppresses the inherent vowel. Thus, a special sign is needed whenever this implicit vowel is stripped off. This symbol is known as the *Hasanta* (=Halant) "्" (U+09CD). By placing the Hasanta under the first consonant of a combination or cluster, one could 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 it 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 Bangla words as seen in print till date. Given the confluence of languages happening to interact in the Internet age, 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 LGR work, this limit will not be enforced.

### 3.4.3 Vowels

Separate symbols exist for all ‘*Swara*’ or Vowels, 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 (*Mātrā*) 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 under:

|  |  |
| --- | --- |
| **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 (Hasanta) |
| - | ঽ U+09BD Abagraha |

Table 8: Vowels with corresponding Mātrās

### 3.4.4 The Anusvara (ং - U+0982)

The Anusvara represents a homorganic nasal. It replaces a conjunct group of a Nasal Consonant+Halant+Consonant belonging to that particular barga. Before a non-barga consonant the anusvara represents a nasal sound. 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.

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

Candrabindu denotes nasalization of the preceding vowel as in চাঁদ /cãd/ ‘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.

### 3.4.6 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, Santhali and Sindhi. In Bengali, its use is further restricted. It can be optionally adjoined to ক (U+0995), খ (U+0996), গ (U+0997), জ (U+099C) and ফ (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" (U+09A1) and "ঢ" (U+09A2) to indicate flapped sounds বড় /bədh/ (U+09AC U+09A1 U+09BC). Of course, Bangla Unicode points already account for these two letters separately as under ড় (U+09DC) and ঢ় (U+09DD).

Normally a Nukta is appended to Consonants. However, Santhali 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 Bangla script):

a. आ (U+0906)

b. ओ (U+0913)

c. ा (U+093E)

d. ो (U+094B)

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

The Visarga 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 /du:kho/ sorrow, unhappiness (U+0926 U+0941 U+0903 U+0916).

The Avagraha "ऽ" (U+093D) creates an extra stress on the preceding vowel and is used in Sanskrit or Maithili texts written in Bangla. It is rarely used in other languages using Bangla script In case of LGR, the Avagraha is not part of the repertoire as it is barred in the Maximal Starting Repertoire.

# Overall Development Process and Methodology

The Neo-Brahmi Generation Panel (NBGP) has been formed by members having experience in Linguistics (especially in the 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 are all in sync with all other Brāhmī-derived scripts, especially with the Devanagari writing system. The present LGR will caters to multiple languages belonging to EGIDS scale 1 to 4 (see Table 4) that use Bangla or Bengali script..

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

##  4.1 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.

### 4.1.1 Inclusion principles:

#### 4.1.1.1 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.

#### 4.1.1.2 Unambiguous use:

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

### 4.1.2 Exclusion principles:

The main exclusion principle is that of Acknowledgement to Environmental Limitations. 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.

#### 4.1.2.1 Acknowledgement to Environment Limitations:

The code point repertoire for root zone being a very special case, up the ladder in the 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, especially in Bengali-Asamiya-Manipuri Script, given the elaborate and exhaustive character inclusion efforts made by 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 introduces exclusion of some characters out of Unicode repertoire from being part of the domain names.

Here are some examples: Bangla first consonant Letter with “nukta” corresponding to Devanagari Qa "क़" (U+0958) is not allowed to be a part of domain name. Its decomposed form, i.e. Bangla Letter “ক” Ka followed by Devanagari Sign Nukta, i.e. "ক" (U+0995) + "़" (U+09BC) can be used instead.

*iii. Maximal Starting Repertoire* (MSR)*:*

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

Example: Bangla 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. This is further narrowed down by the IDNA Protocol and finally an additional filter in the form of Maximal Starting Repertoire restricts the character set associated with the given language even more.

#### 4.1.2.2 No Punctuation Marks:

The TLDs being identifiers, punctuation markers present in brahmi based languages such as Danda "।" (U+09F7) and double Danda of Devanagari, i.e. "॥" (U+0965) will not be included.

#### 4.1.2.3 No Symbols and Abbreviations:

Abbreviations, weights and measures and other such iconic characters like Isshar "৺" (U+09FA), Abbreviation sign "॰" (U+09F9) etc. will also not be included.

#### 4.1.2.4 No Rare and Obsolete Characters:

There are characters which have been added to Unicode to accommodate rare forms especially like Sanskritic VOCALIC RR "ৠ" (U+09E0) and VOCALIC L “ঌ” as well as LL "ৡ" (U+09E1) and the matra forms of the latter two symbols - "ৢ" (U+09E2) and “ৣ" (U+09E3). All such characters will not be included. This is in consonance with the Conservatism principle as laid down in the Root Zone LGR procedure. However, in Bengali, the vowel matra corresponding to "ৠ" (U+09E0) which is “ৄ” (U+09C4) is still in active use in certain borrowed or Sanskritic words, and could therefore be retained.

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

Stress markers for classical Sanskrit e.g. DEVANAGARI STRESS SIGN UDATTA "॑" (U+0951) and DEVANAGARI STRESS SIGN ANUDATTA "॒" (U+0952) will not be included. This is also in consonance with the Letter principle as laid down in the Root Zone LGR procedure.

## 4.2 The Basis of Present IDN

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

### 4.2.1 The ABNF Variables

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

Dash → Hyphen -

Digit → Indo-Arabic digits [0-9]

C → Consonant

V → Vowel

M → Matra

D → Anusvara

B → Chandrabindu

X → Visarga

H → Halant/ Hasant /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 9: The ABNF Formalism

# Repertoire

Bengali Writing System is represented in UNICODE as under ISO 15924. There are 93 entries in this list. This section details the code-point repertoire that the Neo-Brahmi Generation Panel [NBGP] proposes to be included in the Bengali LGR.

For each of the code points, language references have been given in the last column titled "Reference" under Table 11 titled the “Code Point Repertoire”. For entire coverage of Bengali code points, references of Bangla, Asamiya, Manipuri (Meitei), and Bishnupuriya have been given. Kokborok, written in Bangla 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 in Bangla Unicode Points (as given in UNICODE 10).

However, before the details are presented, it is ideal to take a look at the Bengali Code Point Chart from U0980 file to have a fair idea of Code Points. 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:



Table 10: Code-Point Table for Bengali-Asamiya-Manipuri

Given the Bengali Unicode Block as in Table 11, the following symbols will need a separate treatment:

ৎ U+09CE Bengali Letter Khanda-Ta

ৰ U+09F0 Asamiya-Bengali Letter Ra With Middle Diagonal
ৱ U+09F1 Asamiya-Bengali Letter Ra With Lower Diagonal
৺ U+09FA Bengali Isshar ৻ U+09FB Bengali Ganda Mark
০ U+09E6 Bengali Digit Zero ১ U+09E7 Bengali Digit One
২ U+09E8 Bengali Digit Two ৩ U+09E9 Bengali Digit Three
৪ U+09EA Bengali Digit Four ৫ U+09EB Bengali Digit Five৬ U+09EC Bengali Digit Six ৭ U+09ED Bengali Digit Seven
৮ U+09EE Bengali Digit Eight ৯ U+09EF Bengali Digit Nine
৲ U+09F2 Bengali Rupee Mark ৳ U+09F3 Bengali Rupee Sign

৴ U+09F4 Bengali Currency Numerator One
৵ U+09F5 Bengali Currency Numerator Two
৶ U+09F6 Bengali Currency Numerator Three
৷ U+09F7 Bengali Currency Numerator Four
৸ U+09F8 Bengali Currency Numerator One Less Than The Denominator
৹ U+09F9 Bengali Currency Denominator Sixteen

The following is Bengali Unicode Block:

|  |
| --- |
| **Bengali** |
|   | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | A | B | C | D | E | F |
| U+098x | ঀ | ঁ | ং | ঃ |  | অ | আ | ই | ঈ | উ | ঊ | ঋ | ঌ |  |  | এ |
| U+099x | ঐ |  |  | ও | ঔ | ক | খ | গ | ঘ | ঙ | চ | ছ | জ | ঝ | ঞ | ট |
| U+09Ax | ঠ | ড | ঢ | ণ | ত | থ | দ | ধ | ন |  | প | ফ | ব | ভ | ম | য |
| U+09Bx | র |  | ল |  |  |  | শ | ষ | স | হ |  |  | ় | ঽ | া | ি |
| U+09Cx | ী | ু | ূ | ৃ | ৄ |  |  | ে | ৈ |  |  | ো | ৌ | ্ | ৎ |  |
| U+09Dx |  |  |  |  |  |  |  | ৗ |  |  |  |  | ড় | ঢ় |  | য় |
| U+09Ex | ৠ | ৡ | ৢ | ৣ |  |  | ০ | ১ | ২ | ৩ | ৪ | ৫ | ৬ | ৭ | ৮ | ৯ |
| U+09Fx | ৰ | ৱ | ৲ | ৳ | ৴ | ৵ | ৶ | ৷ | ৸ | ৹ | ৺ | ৻ | ৼ | ৽ |  |  |
| Table 10: Bengali UNICODE Block |

## 5.1 Code Point Repertoire

| **No.** | **Unicode Code Point** | **Glyph** | **Character Name** | **Unicode General Category (gc)** | **Indic Syllabic Category** | **Language(s), with EGIDS Value** | **References** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 1. | U+0981 | ঁ | BENGALI SIGN CANDRABINDU | Mn | Chandra- bindu | TBD | TBD |
| 2. | U+0982 | ং | BENGALI SIGN ANUSVARA | Mn | Anuswara | TBD | TBD |
| 3. | U+0983 | ঃ | BENGALI SIGN VISARGA | Lo | Vowel | TBD | TBD |
| 4. | U+0985 | অ | BENGALI LETTER A | Lo | Vowel | TBD | TBD |
|  5. | U+0986 | আ | BENGALI LETTER AA | Lo | Vowel | TBD | TBD |
|  6. | U+0987 | ই | BENGALI LETTER I | Lo | Vowel | TBD | TBD |
|  7. | U+0988 | ঈ | BENGALI LETTER II | Lo | Vowel | TBD | TBD |
| 8. | U+0989 | উ | BENGALI LETTER U | Lo | Vowel | TBD | TBD |
| 9. | U+098A | ঊ | BENGALI LETTER UU | Lo | Vowel | TBD | TBD |
| 10. | U+098B | ঋ | BENGALI LETTER VOCALIC R | Lo | Vowel | TBD | TBD |
| 11. | U+098C | ঌ | BENGALI LETTER VOCALIC L | Lo | Vowel | TBD | TBD |
| 12. | U+098F | এ | BENGALI LETTER E | Lo | Vowel | TBD | TBD |
| 14. | U+0990 | ঐ | BENGALI LETTER AI | Lo | Vowel | TBD | TBD |
| 15. | U+0993 | ও | BENGALI LETTER O | Lo | Vowel | TBD | TBD |
| 16. | U+0994 | ঔ | BENGALI LETTER AU | Lo | Consonant | TBD | TBD |
| 17. | U+0995 | ক | BENGALI LETTER KA | Lo | Consonant | TBD | TBD |
| 18. | U+0996 | খ | BENGALI LETTER KHA | Lo | Consonant | TBD | TBD |
| 19. | U+0997 | গ | BENGALI LETTER GA | Lo | Consonant | TBD | TBD |
| 20. | U+0998 | ঘ | BENGALI LETTER GHA | Lo | Consonant | TBD | TBD |
| 21. | U+0999 | ঙ | BENGALI LETTER NGA | Lo | Consonant | TBD | TBD |
| 22. | U+099A | চ | BENGALI LETTER CA | Lo | Consonant | TBD | TBD |
| 23. | U+099B | ছ | BENGALI LETTER CHA | Lo | Consonant | TBD | TBD |
| 24. | U+099C | জ | BENGALI LETTER JA | Lo | Consonant | TBD | TBD |
| 25. | U+099D | ঝ | BENGALI LETTER JHA | Lo | Consonant | TBD | TBD |
| 26. | U+099E | ঞ | BENGALI LETTER NYA | Lo | Consonant | TBD | TBD |
| 27. | U+099F | ট | BENGALI LETTER TTA | Lo | Consonant | TBD | TBD |
| 28. | U+09A0 | ঠ | BENGALI LETTER TTHA | Lo | Consonant | TBD | TBD |
| 29. | U+09A1 | ড | BENGALI LETTER DDA | Lo | Consonant | TBD | TBD |
| 30. | U+09A2 | ঢ | BENGALI LETTER DDHA | Lo | Consonant | TBD | TBD |
| 31. | U+09A3 | ণ | BENGALI LETTER NNA | Lo | Consonant | TBD | TBD |
| 32. | U+09A4 | ত | BENGALI LETTER TA | Lo | Consonant | TBD | TBD |
| 33. | U+09A5 | থ | BENGALI LETTER THA | Lo | Consonant | TBD | TBD |
| 34. | U+09A6 | দ | BENGALI LETTER DA | Lo | Consonant | TBD | TBD |
| 35. | U+09A7 | ধ | BENGALI LETTER DHA | Lo | Consonant | TBD | TBD |
| 36. | U+09A8 | ন | BENGALI LETTER NA | Lo | Consonant | TBD | TBD |
| 37. | U+09AA | প | BENGALI LETTER PA | Lo | Consonant | TBD | TBD |
| 38. | U+09AB | ফ | BENGALI LETTER PHA | Lo | Consonant | TBD | TBD |
| 39. | U+09AC | ব | BENGALI LETTER BA | Lo | Consonant | TBD | TBD |
| 40. | U+09AD | ভ | BENGALI LETTER BHA | Lo | Consonant | TBD | TBD |
| 41. | U+09AE | ম | BENGALI LETTER MA | Lo | Consonant | TBD | TBD |
| 42. | U+09AF | য | BENGALI LETTER YA | Lo | Consonant | TBD | TBD |
| 43. | U+09B0 | র | BENGALI LETTER RA | Lo | Consonant | TBD | TBD |
| 44. | U+09B2 | ল | BENGALI LETTER LA | Lo | Consonant | TBD | TBD |
| 45. | U+09B6 | শ | BENGALI LETTER SHA | Lo | Consonant | TBD | TBD |
| 46. | U+09B7 | ষ | BENGALI LETTER SSA | Lo | Consonant | TBD | TBD |
| 47. | U+09B8 | স | BENGALI LETTER SA | Lo | Consonant | TBD | TBD |
| 48. | U+09B9 | হ | BENGALI LETTER HA | Lo | Consonant | TBD | TBD |
| 49. | U+09BC | ় | BENGALI SIGN NUKTA | Mn | Matra | TBD | TBD |
| 50. | U+09BE | া | BENGALI VOWEL SIGN AA | Mn | Matra | TBD | TBD |
| 51. | U+09BF | ি | BENGALI VOWEL SIGN I | Mn | Matra | TBD | TBD |
| 52. | U+09C0 | ী | BENGALI VOWEL SIGN II | Mn | Matra | TBD | TBD |
| 53. | U+09C1 | ু | BENGALI VOWEL SIGN U | Mn | Matra | TBD | TBD |
| 54. | U+09C2 | ূ | BENGALI VOWEL SIGN UU | Mn | Matra | TBD | TBD |
| 55. | U+09C3 | ৃ | BENGALI VOWEL SIGN VOCALIC R | Mn | Matra | TBD | TBD |
| 56. | U+09C4 | ৄ | BENGALI VOWEL SIGN VOCALIC RR | Mn | Matra | TBD | TBD |
| 57. | U+09C7 | ে | BENGALI VOWEL SIGN E | Mn | Matra | TBD | TBD |
| 58. | U+09C8 | ৈ | BENGALI VOWEL SIGN AI | Mn | Matra | TBD | TBD |
| 59. | U+09CB | ো | BENGALI VOWEL SIGN O | Mn | Matra | TBD | TBD |
| 60. | U+09CC | ৌ | BENGALI VOWEL SIGN AU | Mn | Matra | TBD | TBD |
| 61. | U+09CD | ্ | BENGALI SIGN BIRAMA | Mn | Hasanta/ Birama | TBD | TBD |
| 62. | U+09CE | ৎ | BENGALI LETTER KHANDA TA | Lo | Consonant | TBD | TBD |
| 63. | U+09D7 | ৗ | BENGALI AU LENGTH MARK | Mc | Matra | TBD | TBD |
| 64. | 09F0 | ৰ | BENGALI LETTER RA WITH MIDDLE DIAGONAL | Lo | Consonant | TBD | TBD |
| 65. | 09F1 | ৱ | BENGALI LETTER RA WITH LOWER DIAGONAL | Lo | Consonant | TBD | TBD |

# 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, a 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 *a*-matra as two separate keys getting the same results. A reader cannot differentiate between the two *ko* (কো), one typed with one key and the other one typed with two different keys. But 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 র one can get it by typing ব and then by typing nukta resulting in a shape as this one ব়. This is could be mistaken for a variant as the র has a different code point assigned to it in relation to the ব and the nukta. This sequence of typing a nukta after ব could be blocked. A direct *ra* has the code value U+09B0 on one hand and U+09AB followed U+09BC on the other. Hence, it does not stand as an example of a variant.

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

True examples of variants in Bengali may be the cases where halant থ appeasrs in a conjunct with *sa* and *na*.

Bangla variant:

স + halanta + থ versus স + halanta + হ

ন + halanta + থ versus ন + halanta + হ

ণ + halanta + থ versus ণ + halanta + হ

The above combinations, if written in traditional orthography, could be little confusing, where the tha (থ) in conjunct appears like a ha (হ). It could be typed wrong as well.

1. স্থ স্হ

2. ন্থ ন্হ

3. ন্হ ণ্হ

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

# Whole Label Evaluation Rules (WLE)

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

## 7.1 The Vowel Sequence

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

The possibility of a Visarga or Anusvara following a Chandrabindu exists in Bangla. Vowel can optionally be followed by a combination of Halant/ Hasant/ Virama [H], Consonant [C] to form a Ya-phalaa. “Ya-phalaa 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-phalaa has a special form: য়. Again, when combined with U+09BE া , i.e. Bengali vowel sign for ‘aa’, it is used for transcribing [æ] as in the “a” in the English word“bat” written in Bengali as ব্যাট.

The vowel sequence in Bangla is, therefore, the following:

V[D|B|X|BD|BX|HCM[D|B|X|BD|BX]]

A Vowel-sequence admits the following combinations:

### 7.1.1. A Single Vowel

Examples: V অ अ

7.1.2. A Vowel which can be followed by set of characters

A vowel can optionally be followed by Anusvara [D] or Chandrabindu [B] or Visarga [X] or Chandrabindu+Anusvara [BD] or Chandrabindu+Visarga [BX] or combination of Halant (also known as Hasant/Virama) [H] followed by Consonant [C] followed by Matra [M].

Examples:

 VD অং अं

 VB অঁ अँ

 VX অঃ अः

 VBD অঁং अँंं

 VHCM অঁঃ अँंः

 7.1.3. A VHCM sequence

A VHCM sequence can optionally be followed by Anusvara [D] or Chandrabindu [B] or Visarga [X] or Chandrabindu+Anusvara[BD] or Chandrabindu+Visarga[BX].

Examples:

 VHCMD অ্য়াং

 VHCMB অ্য়াঁ

 VHCMX অ্য়াঃ

 VHCMBD অ্য়াঁং

 VHCMBX অ্য়াঁঃ

## 7.2. The Consonant Sequence

 7.2.1 A Single Consonant (C)

Example: C ক क

7.2.2 A Consonant optionally followed a specific set of characters

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

C[M|D|B|X|H|BD|BX]

Example:

 CM কি कि

 CD কং कं

 CB কঁ कँ

 CX কঃ कः

 CH ক্ क् (Pure consonant)

 CBD কঁং कँं ं

 CBX কঁঃ कँः

7.2.3 A CM sequence can be optionally followed by D, B, X, BD or BX.

(CM)[D|B|X|BD|BX]

Example:

 CMD কীং कीं

 CMB কাঁ काँ

 CMX বীঃ वीः

 CMBD কাঁং काँ ं

 CMBX কাঁঃ काँः

7.2.4 A sequence of consonants (up to 4) joined by Halant

(also known as Hasant/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

####  7.2.4.1 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, D, B, X, BD or BX.

Example:

 CHCM ক্কী →ক ্ ক ী क्की → क ं्क ী

 CHCD ক্কং →ক ্ ক ং क्कं → क ं्क ंं

CHCB ক্কঁ →ক ্ ক ঁ क्कँ→ क ं्क ंँ

CHCX ক্কঃ →ক ্ ক ঃ क्कः → क ं्क ঃ

CHCBD ক্কঁ ং →ক ্ ক ঁ ং क्कँंं→क ं्क ंँंं

CHCBX ক্কঁঃ →ক ্ ক ঁ ঃ क्कँंः→ क ं्क ंँंः

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

(CHCM)[D|B|X|BD|BX]

Example:

CHCMD ক্কীং → ক ্ ক ী ং क्की → क ं्क ंी

CHCMB ক্কাঁ → ক ্ ক া ঁ क्का ँ→ क ं्क ंा ंँ

CHCMX ক্কীঃ → ক ্ ক ী ঃ क्कीः → क ं्क ंी ंः

CHCMBD ক্কাঁং→ ক ্ ক া ঁ ং क्कांंँ→ क ं्क ंा ंँंं

CHCMBX ক্কাঁঃ → ক ্ ক া ঁ ঃ क्कांःँ → क ं्क ंा ंँंः

7.2.5 A single ‘Khanda’-Ta (Z)

 Example: Z ৎ = ত্

7.2.6 A Khanda Ta and Hasant

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

[CH]Z

Example: র + ্ + ৎ = র্ৎ

The final canonical structure of Bengali consonant sequence can thus be defined in ABNF as:

 \*3(C[N]H)C[N][H|D|B|X|BD|BX|M[D|B|X|BD|BX]] | [CH]Z

## 7.3. ABNF Applied to the Bangla IDN

 The formalism can be applied to create/validate IDN labels in Bangla. So a valid IDN label in Bangla can be defined as follows.

Vowel-sequence →

 V [D | B | X | BD | BX | HCM[D|B|X|BD|BX]]

Consonant-sequence →

\*3(C[N]H)C[N][H|D|B|X|BD|BX|M[D|B|X|BD|BX]] | [CH]Z

Sequence → Consonant-sequence[Y] | Vowel-sequence[Y]

IDN-label → (sequence | digit) \* ([dash] (sequence |digit))

## 7.4. Additional Examples from Bangla 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, D, B or X cannot occur in the beginning of a Bangla IDN. Example:

ं्क ং্ক

कंक কংক

ंं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, D, B, X, M, Digit or Dash. Example:

अ् অ্

कं ्् অং্্

कँ ्् কঁ্্

कः्् কঃ্্

कि् কি্

1्् 1্্

-्् - ্্

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

Example:

कं ंं কংং

कँ ंँ কঁ্ঁ

कःंः কংঃঃ

कंंं কংং

कांँँ কাঁ্ঁ

कीःंः কীংঃঃ

अंंं অংং

अँंँ অঁ্ঁ

अःंः অংঁ্ঁ

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

की्ी কিি

5. M is not permitted after V Example:

ईा ইা

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

Example:

कं ्ः কংঃ

कः्ं কঃং

## 7.5. Restriction Rules

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 Bangla the following rules apply:

1. Khanda ta (ৎ) is NOT allowed in the beginning of an IDN label. Same applies to ঞ and the velar nasal ঙ in the Bengali *barga*. Moreover, Bengali does not allow *ya* (য়) in the beginning of a word also.

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

র্ৎ as in ভর্ৎসনা

3. Nukta shall be allowed only after following characters:

 জ (জ়), ড (ড়), ঢ (ঢ়)

4. Only following combinations with VHCM will be allowed.

→ ( অ্যা) as in অ্যাসিড (acid)

→ এ্যা as in এ্যাসিড, এ্যাসোসিয়েশান (acid, association)

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

 আল্লাহ্-

 শাহ্ ১

 Or even combination of hyphen and a digit as in বাহাদুর শাহ্-১

6. Consecutive Hyphens will not be permitted in a domain name.

7. 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

8. 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)

**Additional Note:**

Wherever a variant is present in a given label, the variants shall be strictly

symmetric and non-transitive. Thus given a variant and, the variants

of a label such as shall be generated by adding an extra to shall not be permitted regardless of its ABNF Validity status. This ensures that over generativity does not take place.

# Contributors

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

Dr Atiur Rahman Khan, C-DAC, Pune

# References

Bandyopadhyay, Chittaranjan. 1981. Dui Shataker Bangla Mudran o Prakashan. Kolkata: Ananda Publishers.

Banerji, R.D. 1919. The Origin of the bengali Script. Kolkata. New Delhi; Asian Educational Services; 2003 reprint.

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

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

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

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

Majumdar, R.C. 1971. shows, in his History of Ancient Bengal,

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

Pal, Palash Baran. 2001. Dhwanimala Barnamala. Kolkata: Papyrus.

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

Ross, Fiona. 1999. The Printed Bengali Character and its Evolution. London: Curzon.

Shastri, Mahamahopadhyay Hara Prasad. 1916. Hājār Bacharēr Purāṇa Bāṅgālā Bhāṣāy Bauddha Gān ō Dōhā. Calcutta: Bangiya Sahitya Parishad.

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

-----. 1987. A Bibliography of Bengali Linguistics. Mysore : CIIL. xii+316 pp.

-----. 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.

-----. 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.

Sripantha. 1996. Jakhan Chapakhana Elo. Kolkata: Paschim-Banga Bangla Academy.

Sur, Atul. 1986. Bangla Mudraner Dusho Bachar. Kolkata: Jijnasa.

# Appendix