

# Proposal for a Devanagari Script Root Zone Label Generation Rule-Set [LGR]

---

*LGR Version: 2.0*

*Date: 4<sup>th</sup> July 2017*

*Document version: 1.1*

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

## 1 General Information/ Overview/ Abstract

This document lays down the Label Generation Rule Set for Devanagari script. Three main components of the Devanagari Script LGR i.e. Code point repertoire, Variants and Whole Label Evaluation Rules have been described in detail here. All these components have been incorporated in a machine-readable format in the accompanying XML file named "**Proposed-LGR-Deva-20170323.xml**".

## 2 Script for which the LGR is proposed

ISO 15924 Code: Deva

ISO 15924 Key N°: 315

ISO 15924 English Name: Devanagari (Nagari)

Latin transliteration of native script name: dévanâgarî

Native name of the script: देवनागरी

Maximal Starting Repertoire [MSR] version: 2

## 3 Background on Script and Principal Languages Using It

The script called Nagari or Devanagari is written from left to right. Historically it derives from the Brahmi alphabet of the Ashokan inscriptions. Devanagari is currently used for 11

out of 22 scheduled languages of India (Boro/Bodo, Dogri, Hindi, Kashmiri, Konkani, Maithili, Marathi, Nepali, Sanskrit, Santhali and Sindhi) and around 45 other languages especially the related Indo-Aryan languages: Bagheli, Bhili, Bhojpuri, Himachali dialects, Magahi, Newari and Rajasthani and its dialects: Marwari, Mewati, Shekhawati, Bagri, Dhundhari, Harauti and Wagri. Closely associated with Sanskrit and Prakrit, it is an alternative script for Kashmiri (by Hindu speakers), Sindhi and Santhali. It is growing popular in use by speakers of tribal languages of Arunachal Pradesh, Bihar and Andaman & Nicobar Islands. The script is also used in Fiji to represent Fiji Hindi. Hindi is also a language of communication in Mauritius, Malaysia, England, Canada, South Africa, Indonesia as well as emigrant communities around the world.

Devanagari is used by over 120 languages both in India and in South-east-Asia.

### 3.1 The Evolution of the Script

It is well-known that Devanagari has evolved from the parent script Brahmi, with its earliest historical form known as Aśokan Brahmi, traced to the 4th century B.C. Brahmi was deciphered by Sir James Prinsep in 1837. The study of Brahmi and its development has shown that it has given rise to most of the scripts in India as well as in other countries viz. Sri Lanka, Myanmar, Kampuchea, Thailand, Laos, and Tibet to name a few.

The evolution of Brahmi into present-day Devanagari involved intermediate forms, common to other scripts such as Gupta and Śāradā in the north and Grantha and Kadamba in the South. Devanagari can be said to have developed from the Kutila script, a descendant of the Gupta script, in turn a descendent of Brahmi. The word kutila, meaning 'crooked', was used as a descriptive term to characterize the curving shapes of the script, compared to the straight lines of Brahmi. This inheritance is the reason for some of the characters across the scripts that will be considered under the Neo-Brahmi GP to look similar to each other despite belonging to totally different code blocks.

A look at the development of Devanagari from Brahmi gives an insight into how the Indic scripts have come to be diversified: the handiwork of engravers and writers who used different types of strokes leading to different regional styles. The development of the script

is outlined below. Figure 1: Pictorial depiction of Evolution of Devanagari illustrates the stages in the evolution of the script<sup>1</sup>.

Period	Description
300 BCE	Mauryan : Early Brahmi form the Asokan edicts. Some scholars believe that Brahmi itself evolved from "karoshti" a script written right to left.
200 CE	Kushan/Satavahana Dynasties.
400 CE	Gupta Dynasty
600 CE	Yasodharman
800 CE	Origins of the present day Nagari Script. Vardhana dynasty in the North and Pallava period in the South.
900 CE	The period of the Chalukyas and Rashtrakutas
1100 CE	Continuation of the Chalukya Rule
1300 CE	Yadavas in the north and Kakatiyas in the south.
1500 CE	The Vijayanagar empire.

Table 1: Evolution of Devanagari

300 BCE	†	ε	୪	।	୮	୩
200 CE	†	Ε	⋈	।	୮	୩
400 CE	†	Ε	୮	।	୮	୩
600 CE	†	Ε	୮	।	୮	୩
800 CE	†	ε	୮	।	୮	୩
900 CE	†	୮	୮	।	୮	୩
1100 CE	†	୮	୮	।	୮	୩
1300 CE	†	୮	୮	।	୮	୩
Modern	क	ज	म	र	स	अ

Figure 1: Pictorial depiction of Evolution of Devanagari

<sup>1</sup> [http://www.acharya.gen.in:8080/sanskrit/script\\_dev.php](http://www.acharya.gen.in:8080/sanskrit/script_dev.php)

### 3.2 Languages considered

Below is the tabular representation of the languages that have been considered for the Devanagari LGR. As per the requirement of the LGR procedure, languages belonging to the EGIDS scale 1 to 4 have been considered.

EGIDS Scale 1	EGIDS Scale 2	EGIDS Scale 3	EGIDS Scale 4
Hindi Nepali	Konkani Maithili Marathi Sindhi	Bhatrī Halbi Kinnauri Kukna Panchpargania Sadri Wagdi	Bhojpuri Chhattisgarhi Dogri Kashmiri Limbu Magahi Sanskrit Santhali Tamang, Eastern Avadhi Newar Saraiki

Table 2: Main languages considered under Devanagari LGR

Despite of being classified under EGIDS Scale 5, Boro language is also considered under the Devanagari LGR as it is one of the scheduled languages of India and is widely spoken.

### 3.3 The structure of written Devanagari

Devanagari is an alphasyllabary and the heart of the writing system is the Akshar. It is this unit, which is instinctively recognized by users of the script. To understand the notion of akshar, a brief overview of the writing system is provided in this Section and the akshar itself will be treated in depth in Section 3.4.

The writing system of Devanagari could be summed up as composed of the following:

### 3.3.1 The Consonants

Devanagari consonants have an implicit schwa /ə/ included in them. As per traditional classification they are categorized according to their phonetic properties. There are 5 Varga groups (classes) and one non-Varga group. Each Varga, which corresponds to Stops, contains five consonants classified as per their properties. The first four consonants are classified on the basis of Voicing and Aspiration and the last is the corresponding nasal.

Varga	Unvoiced		Voiced		Nasal
	-Asp	+Asp	-Asp	+Asp	
<b>Velar</b>	क	ख	ग	घ	ङ
<b>Palatal</b>	च	छ	ज	झ	ञ
<b>Retroflex</b>	ट	ठ	ड	ढ	ण
<b>Dental</b>	त	थ	द	ध	न
<b>Bi-labial</b>	प	फ	ब	भ	म

Table 3: Varga classification of consonants

<b>Non-Varga</b>	य	र	ल	ळ	व	श	ष	स	ह
------------------	---	---	---	---	---	---	---	---	---

Table 4: Non-Varga consonants

### 3.3.2 The Implicit Vowel Killer: Halanta<sup>2</sup>

All consonants have an implicit vowel sign (schwa) within them. A special sign is needed to denote that this implicit vowel is stripped off. This is known as the Halanta (्). The Halanta thus joins two consonants and creates conjuncts, which can be generally from 2 to 4 consonant combinations. In rare cases it can join up to 5 consonants. However the notion of maximum number of consonants joining to form one akshar is not empirical. It is just an observation drawn from the words that have been observed till date. Given the confluence of languages happening in the Internet age, the possibility that one may want a generic Top

<sup>2</sup> Unicode (cf. Unicode 3.0 and above) prefers the term Virama. In this report both the terms have been used to denote the character that suppresses the inherent vowel.

Level Domain [gTLD] which may have more than the observed maximum cannot be ruled out. Hence, in the LGR work, this limit will not be enforced<sup>3</sup>.

### 3.3.3 Vowels

Separate symbols exist for all Vowels, which are pronounced independently either at the beginning or after a vowel sound. To indicate a Vowel sound other than the implicit one, a Vowel modifier (Matra) is attached to the consonant. Since the consonant has a built in schwa, there are equivalent Matras for all vowels excepting the अ.

The correlation is shown as under:

अ	आ	इ	ई	उ	ऊ	ऋ	ए	ऐ	ओ	औ
	ा	ि	ी	ु	ू	ृ	े	ै	ो	ौ

Table 5: Vowels with corresponding Matras

In addition to show sounds borrowed from English, some languages using Devanagari such as Hindi, Marathi, and Konkani also admit 2 vowels and their corresponding Matras as in

एँ            ओँ  
 ऐण्ड /and/    और /or/

Marathi replaces the एँ by ऐँ.

### 3.3.4 The Anusvara (ँ)

The Anusvara represents a homo-organic nasal. It replaces a conjunct group of a Nasal Consonant+Halanta+Consonant belonging to that particular varga. Before a non-varga consonant the anusvara represents a nasal sound. Modern Hindi, Marathi and Konkani prefer the anusvara to the corresponding Half-nasal:

सन्त vs. संत /sənt/    saint    चम्पा vs. चंपा /tʃəmpa/

<sup>3</sup> This can be the case when a foreign language word, which admits a large number of consonants, is transliterated into Devanāgarī

### 3.3.5 Nasalization: Chandrabindu (ँ)

Chandrabindu/Anunasika denotes nasalization of the preceding vowel as in आँख (eye) /ākh/ eye. Present-day Hindi users tend to replace the chandrabindu by the anusvara

### 3.3.6 Nukta (ँ)

Mainly used in Hindi, the nukta sign is placed below a certain number of consonants to represent words borrowed from Perso-Arabic. It can be adjoined to क ख ग ज फ to show that words having these consonants with a nukta are to be pronounced in the Perso-Arabic style.

e.g. फ़िरोज़ /firoz/

It is also placed under ड and ढ in Hindi to indicate flapped sounds

बढ़ /bədh/

With the exception of flaps, users of modern-day Hindi hardly use the nukta characters today.

### 3.3.7 Visarga (ः) and Avagraha (s)

The Visarga is frequently used in Sanskrit and represents a sound very close to /h/. दुःख /du:kh/ sorrow, unhappiness.

The Avagraha (s) creates an extra stress on the preceding vowel and is used in Sanskrit texts. It is rarely used in other languages using Devanagari. In case of LGR, the Avagraha is not part of the repertoire as it is barred in the Maximal Starting Repertoire.

## 4 Overall Development Process and Methodology

Under the Neo-Brahmi Generation Panel, there are many different scripts belonging to separate Unicode blocks. Each of these scripts will be assigned a separate LGR; however Neo-Brahmi GP will ensure that the fundamental philosophy behind building those LGRs

are all in sync with all other Brahmi derived scripts. This is the Devanagari LGR, which caters to multiple languages written using Devanagari belonging to EGIDS scale 1 to 4.

#### 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 scripts within its ambit.

The main principle is that of Acknowledgement to Environmental Limitations. These comprise protocols or standards. All further principles are in fact subsumed under these limitations but have been spelt out separately for the sake of clarity.

##### 4.1.1 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. Following three main protocols/standards act as successive filters:

###### *i. The Unicode Chart:*

Out of all the characters that are needed by the given script, if the character in question is not encoded in Unicode, it cannot be incorporated in the code point repertoire. Such cases are quite rare, 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.



Example: Devanagari Letter Qa (क़) is not allowed to be a part of domain name. Its decomposed form, i.e. Devanagari Letter Ka followed by Devanagari Sign Nukta (क़+़) can be used instead.

### *iii. Maximal Starting Repertoire:*

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: Devanagari Sign Avagraha (्) 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 No Punctuation Marks:**

The TLDs being identifiers, punctuation markers present in brahmi based languages such as Danda (।) and double Danda (॥) will not be included.

#### **4.1.3 No Symbols and Abbreviations:**

Abbreviations, weights and measures and other such iconic characters like Isshar (॰), Abbreviation sign (◌) etc. will not be included.

#### **4.1.4 No Rare and Obsolete Characters:**

There are characters which have been added to Unicode to accommodate rare forms especially like DEVANAGARI LETTER VOCALIC RR (ठ) and DEVANAGARI LETTER

VOCALIC LL (ॠ) as well as their matra forms (ॡ) and (ॢ). All such characters will not be included. This is in consonance with the Letter principle as laid down in the Root Zone LGR procedure.

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

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

## 5 Repertoire

This section details the code-point repertoire that the Neo-Brahmi Generation Panel [NBGP] proposes to be included in the Devanagari LGR.

One of the major sources of reference to the justification for inclusion of the code-point is the Indian National Standard 'Enhanced Inscript Keyboard layouts' [INSCRIPT]" laying down the language specific keyboard layouts for all the scheduled languages of India. It is officially published and notified in the Gazette of India. The standard specifies key-layouts for each of the scheduled languages of India. The standard among other things provides a comprehensive language-wise list of various characters as used by the scheduled (of which the set of languages under the ambit of the NBGP is a sub-set) languages of the India. The [INSCRIPT] standard carves out a sub-set of characters applicable to each of the languages out of the respective code-page of the script used by that language.

## 5.1 Code Point Repertoire:

Sr. No.	Unicode Code Point	Character	Character Name	Unicode General Category (gc)	Indic Syllabic Category	Reference
1.	0901	ँ	DEVANAGARI SIGN CANDRABINDU	Mn	Chandrabindu	[INSCRIPT]
2.	0902	ं	DEVANAGARI SIGN ANUSVARA	Mn	Anusvara (Bindu)	[INSCRIPT]
3.	0903	ः	DEVANAGARI SIGN VISARGA	Mc	Visarga	[INSCRIPT]
4.	0905	अ	DEVANAGARI LETTER A	Lo	Vowel	[INSCRIPT]
5.	0906	आ	DEVANAGARI LETTER AA	Lo	Vowel	[INSCRIPT]
6.	0907	इ	DEVANAGARI LETTER I	Lo	Vowel	[INSCRIPT]
7.	0908	ई	DEVANAGARI LETTER II	Lo	Vowel	[INSCRIPT]
8.	0909	उ	DEVANAGARI LETTER U	Lo	Vowel	[INSCRIPT]
9.	090A	ऊ	DEVANAGARI LETTER UU	Lo	Vowel	[INSCRIPT]
10.	090B	ऋ	DEVANAGARI LETTER VOCALIC R	Lo	Vowel	[INSCRIPT]
11.	090D	ॠ	DEVANAGARI LETTER CANDRA E	Lo	Vowel	[INSCRIPT]
12.	090F	ए	DEVANAGARI LETTER E	Lo	Vowel	[INSCRIPT]
13.	0910	ऐ	DEVANAGARI LETTER AI	Lo	Vowel	[INSCRIPT]
14.	0911	ऑ	DEVANAGARI LETTER CANDRA O	Lo	Vowel	[INSCRIPT]
15.	0913	ओ	DEVANAGARI LETTER O	Lo	Vowel	[INSCRIPT]
16.	0914	औ	DEVANAGARI LETTER AU	Lo	Vowel	[INSCRIPT]
17.	0915	क	DEVANAGARI LETTER KA	Lo	Consonant	[INSCRIPT]
18.	0916	ख	DEVANAGARI LETTER KHA	Lo	Consonant	[INSCRIPT]

19.	0917	ग	DEVANAGARI LETTER GA	Lo	Consonant	[INSCRIPT]
20.	0918	घ	DEVANAGARI LETTER GHA	Lo	Consonant	[INSCRIPT]
21.	0919	ङ	DEVANAGARI LETTER NGA	Lo	Consonant	[INSCRIPT]
22.	091A	च	DEVANAGARI LETTER CA	Lo	Consonant	[INSCRIPT]
23.	091B	छ	DEVANAGARI LETTER CHA	Lo	Consonant	[INSCRIPT]
24.	091C	ज	DEVANAGARI LETTER JA	Lo	Consonant	[INSCRIPT]
25.	091D	झ	DEVANAGARI LETTER JHA	Lo	Consonant	[INSCRIPT]
26.	091E	ञ	DEVANAGARI LETTER NYA	Lo	Consonant	[INSCRIPT]
27.	091F	ट	DEVANAGARI LETTER TTA	Lo	Consonant	[INSCRIPT]
28.	0920	ठ	DEVANAGARI LETTER TTHA	Lo	Consonant	[INSCRIPT]
29.	0921	ड	DEVANAGARI LETTER DDA	Lo	Consonant	[INSCRIPT]
30.	0922	ढ	DEVANAGARI LETTER DDHA	Lo	Consonant	[INSCRIPT]
31.	0923	ण	DEVANAGARI LETTER NNA	Lo	Consonant	[INSCRIPT]
32.	0924	त	DEVANAGARI LETTER TA	Lo	Consonant	[INSCRIPT]
33.	0925	थ	DEVANAGARI LETTER THA	Lo	Consonant	[INSCRIPT]
34.	0926	द	DEVANAGARI LETTER DA	Lo	Consonant	[INSCRIPT]
35.	0927	ध	DEVANAGARI LETTER DHA	Lo	Consonant	[INSCRIPT]
36.	0928	न	DEVANAGARI LETTER NA	Lo	Consonant	[INSCRIPT]
37.	092A	प	DEVANAGARI LETTER PA	Lo	Consonant	[INSCRIPT]
38.	092B	फ	DEVANAGARI LETTER PHA	Lo	Consonant	[INSCRIPT]
39.	092C	ब	DEVANAGARI LETTER BA	Lo	Consonant	[INSCRIPT]
40.	092D	भ	DEVANAGARI LETTER BHA	Lo	Consonant	[INSCRIPT]

41.	092E	म	DEVANAGARI LETTER MA	Lo	Consonant	[INSCRIPT]
42.	092F	य	DEVANAGARI LETTER YA	Lo	Consonant	[INSCRIPT]
43.	0930	र	DEVANAGARI LETTER RA	Lo	Consonant	[INSCRIPT]
44.	0932	ल	DEVANAGARI LETTER LA	Lo	Consonant	[INSCRIPT]
45.	0933	ळ	DEVANAGARI LETTER LLA	Lo	Consonant	[INSCRIPT]
46.	0935	व	DEVANAGARI LETTER VA	Lo	Consonant	[INSCRIPT]
47.	0936	श	DEVANAGARI LETTER SHA	Lo	Consonant	[INSCRIPT]
48.	0937	ष	DEVANAGARI LETTER SSA	Lo	Consonant	[INSCRIPT]
49.	0938	स	DEVANAGARI LETTER SA	Lo	Consonant	[INSCRIPT]
50.	0939	ह	DEVANAGARI LETTER HA	Lo	Consonant	[INSCRIPT]
51.	093A	◌	DEVANAGARI VOWEL SIGN OE	Mn	Matra	[INSCRIPT]
52.	093B	◌†	DEVANAGARI VOWEL SIGN OOE	Mc	Matra	[INSCRIPT]
53.	093C	◌◌	DEVANAGARI SIGN NUKTA	Mn	Nukta	[INSCRIPT]
54.	093E	◌ा	DEVANAGARI VOWEL SIGN AA	Mc	Matra	[INSCRIPT]
55.	093F	◌ि	DEVANAGARI VOWEL SIGN I	Mc	Matra	[INSCRIPT]
56.	0940	◌ी	DEVANAGARI VOWEL SIGN II	Mc	Matra	[INSCRIPT]
57.	0941	◌ु	DEVANAGARI VOWEL SIGN U	Mn	Matra	[INSCRIPT]
58.	0942	◌ू	DEVANAGARI VOWEL SIGN UU	Mn	Matra	[INSCRIPT]
59.	0943	◌ृ	DEVANAGARI VOWEL SIGN VOCALIC R	Mn	Matra	[INSCRIPT]
60.	0944	◌्र	DEVANAGARI VOWEL SIGN VOCALIC RR	Mn	Matra	[INSCRIPT]

61.	0945	ँ	DEVANAGARI VOWEL SIGN CANDRA E = candra	Mn	Matra	[INSCRIPT]
62.	0947	े	DEVANAGARI VOWEL SIGN E	Mn	Matra	[INSCRIPT]
63.	0948	ै	DEVANAGARI VOWEL SIGN AI	Mn	Matra	[INSCRIPT]
64.	0949	ॉ	DEVANAGARI VOWEL SIGN CANDRA O	Mc	Matra	[INSCRIPT]
65.	094B	ो	DEVANAGARI VOWEL SIGN O	Mc	Matra	[INSCRIPT]
66.	094C	ौ	DEVANAGARI VOWEL SIGN AU	Mc	Matra	[INSCRIPT]
67.	094D	्	DEVANAGARI SIGN VIRAMA	Mn	Halant / Virama	[INSCRIPT]
68.	094F	इ	DEVANAGARI VOWEL SIGN AW	Mc	Matra	[INSCRIPT]
69.	0956	उ	DEVANAGARI VOWEL SIGN UE	Mn	Matra	[INSCRIPT]
70.	0957	ऊ	DEVANAGARI VOWEL SIGN UUE	Mn	Matra	[INSCRIPT]
71.	0972	अ	DEVANAGARI LETTER CANDRA A	Lo	Consonant	[INSCRIPT]
72.	0973	अ	DEVANAGARI LETTER OE	Lo	Consonant	[INSCRIPT]
73.	0974	आ	DEVANAGARI LETTER OOE	Lo	Consonant	[INSCRIPT]
74.	0975	औ	DEVANAGARI LETTER AW	Lo	Consonant	[INSCRIPT]
75.	0976	अ	DEVANAGARI LETTER UE	Lo	Consonant	[INSCRIPT]
76.	0977	अ	DEVANAGARI LETTER UUE	Lo	Consonant	[INSCRIPT]
77.	0979	झ	DEVANAGARI LETTER ZHA	Lo	Consonant	[INSCRIPT]
78.	097A	य	DEVANAGARI LETTER HEAVY YA	Lo	Consonant	[INSCRIPT]
79.	097B	ग	DEVANAGARI LETTER GGA	Lo	Consonant	[INSCRIPT]
80.	097C	ज	DEVANAGARI LETTER JJA	Lo	Consonant	[INSCRIPT]

81.	097E	ड	DEVANAGARI LETTER DDDA	Lo	Consonant	[INSCRIPT]
82.	097F	ब	DEVANAGARI LETTER BBA	Lo	Consonant	[INSCRIPT]

Table 6: 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 "DEVANAGARI LETTER RRA" in the repertoire.

Sr. No.	Unicode Code Points	Sequence	Character Names	Unicode General Category (gc)	Reference
1.	0931	य	DEVANAGARI LETTER RRA	Lo	[INSCRIPT]
	094D		DEVANAGARI SIGN VIRAMA	Mn	
	092F		DEVANAGARI LETTER YA	Lo	
2.	0931	ह	DEVANAGARI LETTER RRA	Lo	[INSCRIPT]
	094D		DEVANAGARI SIGN VIRAMA	Mn	
	0939		DEVANAGARI LETTER HA	Lo	

Table 7: Sequences

## 5.2 Structural Formation of Devanagari:

All the languages written in Brahmi derived scripts follow a particular way of formation of its words, known as "akshar". In the next section there are detailed akshar formation rules as applicable to representation of "Hindi" language when written in Devanagari Script. These rules need slight changes for different languages written in Devanagari in terms of

- Character addition/deletion (e.g. Nukta [U+093C] character is applicable for Hindi but not Marathi)

- Presence or absence of a particular rule (e.g. Eyelash Ra construct is required in Marathi, Konkani and Nepali but not in Hindi).

In section 7, the Whole Label Evaluation (WLE) rules are given which cover all the languages under the purview of the NBGP for Devanagari script.

### 5.3 Akshar formation rules for Hindi:

This section details the Akshar formation rules as applicable to Hindi. The first section lists the categories of the characters in the form of variables. In the rules, instead of their descriptive names, the variable names are used. The second section lists four operators along with their functions which are assumed while specifying the rules. The following two sections describe the two major categories of the Akshar formations first of which begins with the vowels and the second one with the consonants.

#### 5.3.1 Variables involved

Dash	→ Hyphen -
Digit	→ Indo-Arabic digits [0-9]
C	→ Consonant
M	→ Matra
V	→ Vowel
B	→ Anusvara (Bindu)
D	→ Chandrabindu (Anunasika)
X	→ Visarga
H	→ Halant / Virama
N	→ Nukta

#### 5.3.2 Operators used:

Symbol	Function
	Alternative
[ ]	Optional
*	Variable Repetition
( )	Sequence Group

Table 8: Symbol functions

In what follows, the Vowel Sequence and the Consonant Sequence pertinent to Devanagari, when used to write Hindi, are given.



### 5.3.3 The Vowel Sequence

A vowel sequence begins with a vowel. It may be optionally followed by an Anusvara (D), Chandrabindu (B) or a Visarga (X). The number of D, B or X which can follow a V in Devanagari are restricted to one.

The possibility of a Visarga following a Chandrabindu or Anusvara is ruled out, since it is used only in Vedic and in Bengali script.

The vowel sequence in Hindi is therefore V [D | B | X]

Examples:

Sequence Description	Sequence	Example	Example Decomposition
Vowel	V	अ /a/	
Vowel + Anusvara	V[D]	अँ /aṁ/	अ ँ
Vowel + Chandrabindu	V[B]	अँ /aṁ/	अ ँ
Vowel + Visarga	V[X]	अः /aḥ/	अ ः

Table 9

### 5.3.4 Consonant Sequence

A consonant sequence begins with a consonant. It may be optionally followed by a Nukta (N), Matra (M), Anusvara (B), Chandrabindu (D), Visarga (X) or a Halanta (H). The number of instances of these characters occurring after a consonant is restricted to one. There is a possibility of further extension of the Consonant sequence after the N, M and H. Each of these has been discussed in the following sections:

#### 1. A single consonant (C)

(The consonant shall be treated as coterminous with the Consonant along with the Nukta sign wherever such a case is pertinent.)

Examples:

Sequence Description	Sequence	Example	Example Decomposition
Consonant	C	क /ka/	
Consonant + Nukta	C[N]	कँ /kaṁ/	क ँ

Table 10

2. A consonant optionally followed by dependent vowel sign/Matra [M] or Anusvara [D] Chandrabindu [B] or Visarga[X] or Halant [H]

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

Examples:

Sequence Description	Sequence	Example	Example Decomposition
Consonant + Matra	C[M]	कि /ki/	क ि
Consonant + Anusvara	C[B]	कं /kaṁ/	क ँ
Consonant + Chandrabindu	C[D]	कँ /kaṁ̐/	क ँ̐
Consonant + Visarga	C[X]	कः /kaḥ/	क ः
Consonant + Halanta	C[H]	क् /k/ (Pure Consonant)	क ्

Table 11

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

(CM)[D|B|X]

Example:

Sequence Description	Sequence	Example	Example Decomposition
Consonant + Matra + Anusvara	CM[B]	की /kīm/	क ि ँ
Consonant + Matra + Chandrabindu	CM[D]	काँ /kāṁ̐/	क ा ँ̐
Consonant + Matra + Visarga	CM[X]	कीः /kīḥ/	क ि ः

Table 12

3. A sequence of consonants (up to 4) joined by Halant \*3(CH)C

Example:

Sequence Description	Sequence	Example	Example Decomposition
Consonant + Halanta + Consonant + Halanta + Consonant + Halanta + Consonant	CHCHCHC	न्क्रय /nkrya/	न ् क र् य

Table 13

**Subsets:**

3. A. The combination may be followed by M, D, B or X

Example:

Sequence Description	Sequence	Example	Example Decomposition
Consonant + Halanta + Consonant + Matra	CHC[M]	क्की /kkī/	क् क ी
Consonant + Halanta + Consonant + Anusvara	CHC[B]	क्कं /kkaṁ/	क् क ं
Consonant + Halanta + Consonant + Chandrabindu	CHC[D]	क्कँ /kkaṃ/	क् क ँ
Consonant + Halanta + Consonant + Visarga	CHC[X]	क्कः /kkaḥ/	क् क ः

Table 14

3. B. \*3(CH)CM may be followed by a D, B or X

Example:

Sequence Description	Sequence	Example	Example Decomposition
Consonant + Halanta + Consonant + Matra + Anusvara	CHCM[B]	क्कीं /kkīm/	क् क ी ं
Consonant + Halanta + Consonant + Matra + Chandrabindu	CHCM[D]	क्कीँ /kkīm/	क् क ी ँ
Consonant + Halanta + Consonant + Matra + Visarga	CHCM[X]	क्कीः /kkīḥ/	क् क ी ः

Table 15

These are the basic akshar rules on which the overall Devanagari LGR is based. As languages other than Hindi are considered, some language specific characters and rules are introduced. There are some additional finer aspects to these rules as one takes into account the digits, punctuations and special standalone characters like Avagraha. Those aspects are not discussed here as the MSR on which the LGRs are supposed to be based, excludes those characters.

## 6 Variants

There are no characters/character sequences in Devanagari which can be created by using the characters permitted as per the [MSR] and look exactly alike.

However, Santhali language has a unique requirement for Nukta character (◌᳚ U+093C ) positioning which is not common in other Devanagari based languages. Santhali requires the Nukta character to be followed after certain Vowels and Matras. Complete representation of these Santhali combinations necessitated the Whole Label Evaluation rules (given in the 7) to be opened up for these specific cases. A regular non-Santhali user mostly cannot even anticipate possibility of such a combination and can mistake it for something else.

This gives rise to a possibility of creation of certain labels which can be deceptively similar to a majority of the Devanagari user-base. This being a unique case of homographic similarity, following variants are being proposed.

Variant 1	Variant 2
आ U+0906	आ᳚ U+0906 U+093C
ओ U+0913	ओ᳚ U+0913 U+093C
ा U+093E	ा᳚ U+093E U+093C
ो U+094B	ो᳚ U+094B U+093C

## 7 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 Devanagari Script. The rules have been drafted in such a way that they can be easily translated into the LGR specification.

Below are the symbols used in the WLE rules, for each of the "Indic Syllabic Category" as mentioned in the Table 6: Code point repertoire.

C	→	Consonant
M	→	Matra
V	→	Vowel
B	→	Anusvara (Bindu)
D	→	Chandrabinu
X	→	Visarga
H	→	Halant / Virama
N	→	Nukta
S	→	Eyelash Reph (C1HC2) where C1 is 0931 (ऋ - DEVANAGARI LETTER RRA) H is 094D (् - DEVANAGARI SIGN VIRAMA) C2 is either - 092F (य - DEVANAGARI LETTER YA) or 0939 (ह - DEVANAGARI LETTER HA)

Below are the specific WLE rules:

1. N: must be preceded only by either of specific set of Cs, Vs and Ms

The specific Cs are:

- a. क (U+0915)
- b. ख (U+0916)
- c. ग (U+0917)

d. ज (U+091C)

e. ङ (U+0921)

f. ढ (U+0922)

g. ढ (U+092B)

The specific Vs are:

a. ञ (U+0906) (Required in Santhali language)

b. ण (U+0913) (Required in Santhali language)

The specific Ms are:

a. ण (U+093E) (Required in Santhali language)

b. ण (U+094B) (Required in Santhali language)

2. H: must be preceded by C or CN
3. X: must be preceded by either of V, C, N or M
4. B: must be preceded by either of V, C, N or M
5. D: must be preceded by either of V, C, N or M
6. M: must be preceded either by C or CN
7. V: Can **NOT** be preceded by H

### Case of Eyelash Reph:

In the WLE rules, there is no specific mention of the Eyelash Reph for two reasons:

1. As the U+0931 is added as a part of permissible sequences in Table 7: Sequences, it gets permitted only with the specific sequences.
2. The last characters of both the sequences of which the U+0931 is part, are consonants. As the Eyelash-Reph can take all the combinations as that of a consonant, no specific handling in terms of context rule is required.

## 8 Contributors

Neo-Brahmi Generation Panel members.

## 9 References

[MSR] Maximal Starting Repertoire

[INSCRIPT] Bureau of Indian Standards (BIS), "Enhanced Inscript Keyboard layouts" (IS 16350: 2016)

*<This is a paid resource managed by Bureau of Indian Standards. NBGP will try to get a copy of the same and then share the same with IP>*

[NBGP] Neo-Brahmi Generation Panel

## Appendix