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

LGR Version: 3.0

Date: 8th September 2017

Document version: 1.3

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-20170908.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, Newar 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. Nepali is the official language of Nepal spoken by over 30 million people.

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[1].

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

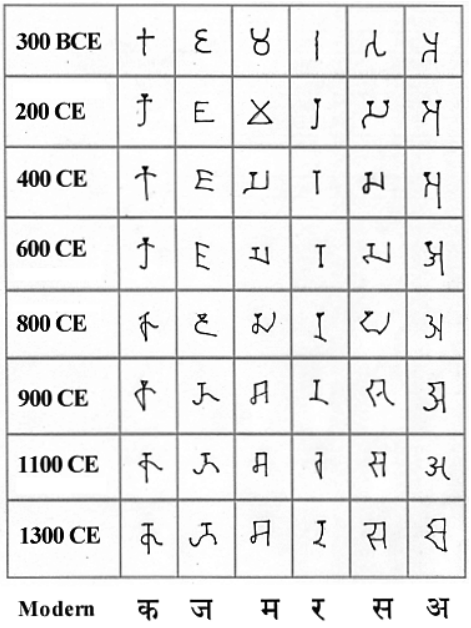

Figure 1: Pictorial depiction of Evolution of Devanagari

[1] 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	Bhatri 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	
Velar	क U+0915	ख U+0916	ग U+0917	घ U+0918	ङ U+0919
Palatal	च U+091A	छ U+091B	ज U+091C	झ U+091D	ञ U+091E
Retroflex	ट U+091F	ठ U+0920	ड U+0921	ढ U+0922	ण U+0923
Dental	त U+0924	थ U+0925	द U+0926	ध U+0927	न U+0928
Bi-labial	प U+092A	फ U+092B	ब U+092C	भ U+092D	म U+092E

Table 3: Varga classification of consonants

Non-Varga	य U+092F	र U+0930	ल U+0932	ळ U+0933	व U+0935	श U+0936	ष U+0937	स U+0938	ह U+0939

Table 4: Non-Varga consonants

3.3.2 The Implicit Vowel Killer: Halant[2]

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 Halant "◌्" (U+094D).

The Halant 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

[2] 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.

generic Top 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[3].

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
U+0905	U+0906	U+0907	U+0908	U+0909	U+090A	U+090B	U+090F	U+0910	U+0913	U+0914
	ा	ि	ी	ु	ू	ृ	े	ै	ो	ौ
	U+093E	U+093F	U+0940	U+0941	U+0942	U+0943	U+0947	U+0948	U+094B	U+094C

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

ऍ	ऑ
U+090D	U+0911
ऍण्ड /and/	ऑर /or/
U+090D U+0923 U+094D U+0921	U+0911 U+0930

Marathi replaces the ऍ (U+090D) by ॲ (U+0972).

3.3.4 The Anusvara (ं - U+0902)

The Anusvara represents a homo-organic nasal. It replaces a conjunct group of a Nasal Consonant+Halant+Consonant belonging to that particular varga. Before a non-varga

[3] This can be the case when a foreign language word, which admits a large number of consonants, is transliterated into Devanāgarī

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/
U+0938 U+0928 U+094D U+0924 vs. U+0938 U+0902 U+0924	U+091A U+092E U+094D U+092A U+093E vs. U+091A U+0902 U+092A U+093E

3.3.5 Nasalization: Chandrabindu (ँ - U+0901)

Chandrabindu/Anunasika denotes nasalization of the preceding vowel as in आँख /ãkh/ eye (U+0906 U+0901 U+0916). Present-day Hindi users tend to replace the chandrabindu by the Anusvara.

3.3.6 Nukta (़ - U+093C)

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/ (U+092B U+093C U+093F U+0930 U+094B U+091C U+093C)

It is also placed under "ड" (U+0921) and "ढ" (U+0922) in Hindi to indicate flapped sounds

बढ़ /bədh/ (U+092C U+0922 U+093C)

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

3.3.7 Visarga (ः - U+0903) and Avagraha (ऽ - U+093D)

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

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 comprise of protocols or standards which are pre-requisites to the Label Generation

Rulesets. 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. 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" (U+0958) is not allowed to be a part of domain name. Its decomposed form, i.e. Devanagari Letter Ka followed by Devanagari Sign Nukta "क" (U+0915) + "़" (U+093C) 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 "ऽ" (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+0964) and double Danda "॥" (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+0970) etc. will 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 DEVANAGARI LETTER VOCALIC RR "ॠ" (U+0960) and DEVANAGARI LETTER VOCALIC LL "ऌ" (U+0961) as well as their matra forms "ॄ" (U+0944) and "ॣ" (U+0963). All such characters will not be included. This is in consonance with the Letter principle as laid down in the Root Zone LGR procedure.

4.1.2.5 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.

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	ऍ	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	ॄ	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	ॏ	DEVANAGARI VOWEL SIGN AW	Mc	Matra	[INSCRIPT]
69.	0956	ॖ	DEVANAGARI VOWEL SIGN UE	Mn	Matra	[INSCRIPT]
70.	0957	ॗ	DEVANAGARI VOWEL SIGN UUE	Mn	Matra	[INSCRIPT]

71.	0972	ऍ	DEVANAGARI LETTER CANDRA A	Lo	Consonant	[INSCRIPT]
72.	0973	ऎ	DEVANAGARI LETTER OE	Lo	Consonant	[INSCRIPT]
73.	0974	ऑ	DEVANAGARI LETTER OOE	Lo	Consonant	[INSCRIPT]
74.	0975	ऒ	DEVANAGARI LETTER AW	Lo	Consonant	[INSCRIPT]
75.	0976	ऄ	DEVANAGARI LETTER UE	Lo	Consonant	[INSCRIPT]
76.	0977	अ	DEVANAGARI LETTER UUE	Lo	Consonant	[INSCRIPT]
77.	0979	ज़	DEVANAGARI LETTER ZHA	Lo	Consonant	[INSCRIPT]
78.	097A	ष	DEVANAGARI LETTER HEAVY YA	Lo	Consonant	[INSCRIPT]
79.	097B	ग़	DEVANAGARI LETTER GGA	Lo	Consonant	[INSCRIPT]
80.	097C	ज़	DEVANAGARI LETTER JJA	Lo	Consonant	[INSCRIPT]
81.	097E	ड़	DEVANAGARI LETTER DDDA	Lo	Consonant	[INSCRIPT]
82.	097F	ब़	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 094D 092F	र्य	DEVANAGARI LETTER RRA DEVANAGARI SIGN VIRAMA DEVANAGARI LETTER YA	Lo Mn Lo	[INSCRIPT]
2.	0931 094D 0939	र्ह	DEVANAGARI LETTER RRA DEVANAGARI SIGN VIRAMA DEVANAGARI LETTER HA	Lo Mn Lo	[INSCRIPT]

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 0, 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/ U+0905	
Vowel + Anusvara	V[D]	अं /aṁ/ U+0905 U+0902	अ ं U+0905 U+0902
Vowel + Chandrabindu	V[B]	अँ /aṃ/ U+0905 U+0901	अ ँ U+0905 U+0901
Vowel + Visarga	V[X]	अः /aḥ/ U+0905 U+0903	अ ः U+0905 U+0903

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 Halant (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/ U+0915	<No Decomposition>
Consonant + Nukta	C[N]	क़ /ḳa/	क ़ U+0915 U+093C

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 ि U+0915 U+093F
Consonant + Anusvara	C[B]	कं /kaṁ/	क ं U+0915 U+0902
Consonant + Chandrabindu	C[D]	कँ /kaṃ/	क ँ U+0915 U+0901

Consonant + Visarga	C[X]	कः /kaḥ/	क ◌ः U+0915 U+0903
Consonant + Halant	C[H]	क् /k/ (Pure Consonant)	क ◌् U+0915 U+094D

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]	कीं /kīṁ/	क ◌ी ◌ं U+0915 U+0940 U+0902
Consonant + Matra + Chandrabindu	CM[D]	काँ /kāṃ/	क ◌ा ◌ँ U+0915 U+093E U+0901
Consonant + Matra + Visarga	CM[X]	कीः /kīḥ/	क ◌ी ◌ः U+0915 U+0940 U+0903

Table 12

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

Example:

Sequence Description	Sequence	Example	Example Decomposition
Consonant + Halant + Consonant + Halant + Consonant + Halant + Consonant	CHCHCHC	न्क्र्य /nkrya/	न ◌् क ◌् र ◌् य U+0928 U+094D U+0915 U+094D U+0930 U+094D U+092F

Table 13

Subsets:

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

Example:

Sequence Description	Sequence	Example	Example Decomposition
Consonant + Halant + Consonant + Matra	CHC[M]	क्की /kkī/	क ् क ी U+0915 U+094D U+0915 U+0940
Consonant + Halant + Consonant + Anusvara	CHC[B]	क्कं /kkaṁ/	क ् क ं U+0915 U+094D U+0915 U+0902
Consonant + Halant + Consonant + Chandrabindu	CHC[D]	क्कँ /kkaṃ/	क ् क ँ U+0915 U+094D U+0915 U+0901
Consonant + Halant + Consonant + Visarga	CHC[X]	क्कः /kkaḥ/	क ् क ः U+0915 U+094D U+0915 U+0903

Table 14

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

Example:

Sequence Description	Sequence	Example	Example Decomposition
Consonant + Halant + Consonant + Matra + Anusvara	CHCM[B]	क्कीं /kkīṁ/	क ् क ी ं U+0915 U+094D U+0915 U+0940 U+0902
Consonant + Halant + Consonant + Matra + Chandrabindu	CHCM[D]	क्कीँ /kkīṃ/	क ् क ी ँ U+0915 U+094D U+0915 U+0940 U+0901
Consonant + Halant + Consonant + Matra + Visarga	CHCM[X]	क्कीः /kkīḥ/	क ् क ी ः U+0915 U+094D U+0915 U+0940 U+0903

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, Devanagari has ample cases of confusingly similar variants. The NBGP categorizes these confusingly similar variants in two groups.

Group 1: Confusing because of pure visual similarity

Group 2: Confusing because of deviation from normally perceived character formations by larger linguistic community

As advised by ICANN, no cases belonging to Group 1 are 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 in which Devanagari words are formed. These can cause confusion even to a careful observer and hence being proposed as variants. Following is the brief description of these variants followed by variants in Table 16 and Table 17.

6.1 Vowel/Vowel sign followed by Nukta:

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 0) 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. Being a unique case of homographic similarity, following variants are being proposed.

Variant 1	Variant 2
आ U+0906	आ़ U+0906 U+093C
ओ	ओ़

U+0913	U+0913 U+093C
◌ा U+093E	◌ा◌़ U+093E U+093C
◌ो U+094B	◌ो◌़ U+094B U+093C

Table 16: Proposed Variants - Set 1

6.2 Halant ending:

Another case of deceptive similarity to a majority of the Devanagari user-base is of a word ending in Halant "्" (U+094D). Even in this case, as majority of Devanagari users do not anticipate an ending Halant, it gives rise to a confusion. Following label-wide variant is being proposed:

Variant 1	Variant 2
A label ending in Halant "्" (U+094D)	Same label ending without Halant "्" (U+094D)

Table 17: Proposed Variants - Set 2

e.g.

Variant 1: मानव् /mɑːnəv/ (U+092E U+093E U+0928 U+0935 **U+094D**)

Variant 2: मानव /mɑːnəvə/ (U+092E U+093E U+0928 U+0935)

6.3 Variant Disposition:

As both the categories are of confusingly similar, albeit of a peculiar nature, it is proposed that they be considered of "blocking" nature.

There is no preference among these variants. Whichever label containing either of these variants is chosen earlier, the other one equivalent variant label should be blocked.

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 → Chandrabindu

X → Visarga

H → Halant / Virama

N → Nukta

S → Eyelash Reph (C1HC2)
where
C1 is 0931 (र - 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. ड (U+0921)

f. ढ (U+0922)

g. फ (U+092B)

The specific Vs are:

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

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

The specific Ms are:

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

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

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

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. आम्अचार */a:m əcha:r/* (U+0906 U+092E U+094D U+0905 U+091A U+093E U+0930) (meaning: *Mango pickle*)

This is the case where two different words are joined together first of which ends in an H and the second word begins with a V. 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 an 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.

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

10 Bibliography

The bibliography given below and sorted thematically is a set of documents, books, articles and webographies consulted in the drafting of this report

WRITING SYSTEMS

1. Dillinger. D., The Alphabet. A Key to the History of Mankind. 3rd Edition in 2 Volumes. Hutchison. London. 1968.

DEVANĀGARĪ

2. Agrawala, V. S. (1966). The Devanāgarī script. In: Indian Systems of Writing. (Pp. 12-16) Delhi: Publications Division.
3. Agyeya, Sacchindanand Hiranand Vatsyayan. 1972. Bhavanti. Delhi: Rajpal and Sons.
4. Beames, John. 1872-79. A Comparative Grammar of the Modern Aryan Languages of India. 3 vols. London, Trubner and Co. [Reprinted by Munshiram Manoharlal, New Delhi, 1966.]
5. Bhatia, Tej K. 1987. A History of the Hindi Grammatical Tradition: Hindi-Hindustani Grammar, Grammarians, History and Problems. Leiden/New York: E. J. Brill.
6. Bright, W. (1996). The Devanāgarī script. In P. Daniels and W. Bright (eds), The World's Writing Systems. (Pp. 384-390). New York: Oxford University Press.
7. Cardona, George. 1987. Sanskrit. In The World's Major Languages. Bernard Comrie (ed.). London: Croom Helm. 448-469.
8. Dwivedi, Ram Awadh. 1966. A Critical Survey of Hindi Literature. Delhi:Motilal Banarsidass.
9. Faruqi, Shamsur Rahman. 2001. Early Urdu Literary Culture and History.Delhi: Oxford University Press.
10. Guru, Kamta Prasad. 1919. Hindi Vyakaran. Varanasi: Nagari Pracharini Sabha. (1962 edition).
11. Kachru, Yamuna. 1965. A Transformational Treatment of Hindi Verbal Syntax. London: University of London Ph.D. dissertation (Mimeographed).
12. Kachru, Yamuna. 1966. An Introduction to Hindi Syntax. Urbana: University of Illinois, Department of Linguistics.

13. Kalyan Kale and Anjali Soman, 1986.Learning Marathi. Shri Vishakha Prakashan, Pune :
14. McGregor, R. S. (1977). Outline of Hindi Grammar. 2nd ed. Delhi: Oxford University Press.
15. McGregor, R. S. 1972. Outline of Hindi Grammar with Exercises. Delhi: Oxford University Press.
16. McGregor, R. S. 1974. Hindi Literature of the Nineteenth and Early Twentieth Centuries. Wiesbaden: Harrassowitz.
17. McGregor, R. S. 1984. Hindi Literature from Its Beginnings to the Nineteenth Century. Wiesbaden: Harrassowitz.
18. Pandey, P. K. (2007). Phonology-orthography interface in Devanāgarī for Hindi. Written Language and Literacy, 10 (2): 139-156. 2007.
19. Rai, Amrit. 1984. A House Divided. The Origin and Development of Hindi/Hindavi. Delhi: Oxford University Press.
20. Sharad, Onkar. 1969. Lohiya ke Vicar. Allahabad: Lokbharati Prakashan.
21. Singh, A. K. (2007). Progress of modification of Brāhmī alphabet as revealed by the inscriptions of sixth-eighth centuries. In P.G. Patel, P. Pandey and D. Rajgor (eds), The Indic Scripts: Paleographic and Linguistic Perspectives. (Pp. 85-107). New Delhi: DK Printworld.
22. Sproat, R. (2000). A Computational Theory of Writing Systems. Cambridge University Press.
23. Tiwari, Pandit Udaynarayan. 1961. Hindi Bhasha ka Udgam aur Vikas [The Origin and Development of the Hindi Language]. Prayag: Leader Press.
24. Verma, M. K. 1971. The Structure of the Noun Phrase in English and Hindi. Delhi: Motilal Banarsidass.

LANGUAGE SPECIFIC

25. IS 10401: 8-bit code for information interchange. 1982
26. IS 10315: 7-bit coded character set for information interchange. 1985
27. IS 12326: 7-bit and 8-bit coded character sets-Code extension techniques. 1987

28. ISO 15919, Information and documentation - Transliteration of Devanāgarī and related Indic scripts into Latin characters. 2001
29. ISO 2375: Procedure for registration of escape sequences. 2003
30. ISO 8859: 8-bit single-byte coded graphic character sets - Parts 1-13. 1998-2001
31. IDN POLICY http://mit.gov.in/sites/upload_files/dit/files/India-IDN-Policy.pdf

Appendix