

Proposal for a Devanagari Script Root Zone Label Generation Rule-Set (LGR)

LGR Version: 3.0

Date: 2019-04-22

Document version: 6.5

Authors: Neo-Brahmi Generation Panel [NBGP]

1 General Information/ Overview/ Abstract

This document lays down the Label Generation Rule Set for the 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 "proposal-devanagari-lgr-22apr19-en.xml".

In addition, a document named "devanagari-test-labels-22apr19-en.txt" has been provided. It contains a list of valid and invalid labels as per the Whole Label Evaluation laid down in Section 7 of this document. The labels have been tagged as valid and invalid under the specific rules¹. In addition, the file also lists the set of labels which can produce variants as laid down in Section 6 of this document.

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: 4

¹ The categorization of invalid labels under specific rules is given as per the general understanding of the LGR Tool by the NBGP. During testing with any LGR tool, whether a particular label gets flagged under the same rule or the different one is totally dependent on the internal implementation of the LGR Tool. In case of discrepancy among the same, the fact that it is an invalid label should only be considered.

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, Santali 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 Wagdi. Closely associated with Sanskrit and Prakrit, it is an alternative script for Kashmiri (by Hindu speakers), Sindhi and Santali. It is growing popular in use by speakers of tribal languages of Arunachal Pradesh, Bihar, Chattisgarh, Jharkhand, Madhya Pradesh 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. The script is also used in Nepal for writing the Nepali language. Nepali is the official language of Nepal as well as one language of the state of Sikkim in India. It is spoken by over 30 million people.

Devanagari is used by over 120 languages in India, Bangladesh, Nepal and in Southeast 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 BC. 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, Cambodia, Thailand, Laos, and the region of Tibet to name a few.

The evolution of Brahmi into present-day Devanagari involved intermediate forms, common to other scripts such as Gupta, and its two generates – Siddaṃ 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 why some of the characters across the scripts that will be considered under the Neo-Brahmi GP look similar to each other despite belonging to totally different code blocks of the Unicode Standard.

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 led 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 script².

Period	Description
300 BCE	Mauryan: Early Brahmi form in the Asokan edicts. Some scholars believe that Brahmi itself evolved from "Kharoshthi" 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

²http://www.acharya.gen.in:8080/sanskrit/script_dev.php

300 BCE	†	ε	Ϝ	।	ॠ	ॡ
200 CE	‡	Ǝ	⋈	।	ॠ	ॡ
400 CE	†	Ǝ	ॠ	।	ॠ	ॡ
600 CE	‡	Ǝ	ॠ	।	ॠ	ॡ
800 CE	‡	Ǝ	ॠ	।	ॠ	ॡ
900 CE	‡	ॠ	ॠ	।	ॠ	ॡ
1100 CE	‡	ॠ	ॠ	।	ॠ	ॡ
1300 CE	‡	ॠ	ॠ	।	ॠ	ॡ
Modern	क	ज	म	र	स	अ

Figure 1: Pictorial depiction of evolution of Devanagari

3.2 Languages considered

Devanagari is used by over 120 languages which makes it one of the most used scripts in the world. Languages using Devanagari as their primary script belong to varying geo-political scenarios as given below:

- designated as official (scheduled) languages of some countries
- used by communities living in urban areas
- used by communities living in rural yet accessible areas
- used by communities living in far-flung areas which are not easily connected either by roads or by communication mechanisms.

Information about official (scheduled) languages of countries is easily available. Information about languages used by communities living in urban areas is also easily obtainable. There was some effort needed to cover the languages which are spoken by communities living in rural yet accessible areas. However, it was quite difficult to cover the rest of the languages being spoken by the communities living in remote tribal areas, which are generally not connected by road or by communication means. Defining the scope of language coverage was hence essential to limit the scope of the work to be undertaken for the analysis of the Devanagari LGR.

NBGP decided to employ “Expanded Graded Intergenerational Disruption Scale” [EGIDS], which is designed to measure the status of the languages of the world in terms of endangerment or development. The EGIDS consists of 13 levels with each higher number on the scale representing a greater level of disruption to the intergenerational transmission of the language. NBGP decided to accommodate all the languages belonging to EGIDS Scale 1 to 4 for its analysis which represents languages in one form or the other are still in usage. Following are the descriptions³ of those scales.

Scale	Label	Description
1	National	The language is widely used between nations in trade, knowledge exchange, and international policy.
2	Provincial	The language is used in education, work, mass media, and government at the national level.
3	Wider Communication	The language is used in education, work, mass media, and government within major administrative subdivisions of a nation.
4	Educational	The language is in vigorous use, with standardization and literature being sustained through a widespread system of institutionally supported education.

Languages belonging to Level 5 and higher are not in widespread usage.

Below is the tabular representation of the languages that have been considered for the Devanagari LGR.

³<https://www.ethnologue.com/about/language-status>

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

Table 2: Languages considered under Devanagari LGR

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

Apart from the above-mentioned languages, Braj, Dhundari, Mundari, and Kharia have also been considered for the analysis as the community using them was accessible and they provided their inputs.

3.2.1 Case of Sanskrit

Sanskrit is generally perceived as an archaic language used only in ancient religious texts. However, it is worth noting that there is a quite vibrant and active user community of Sanskrit in India which practices Sanskrit on day to day basis. Sanskrit is still taught in schools under various State and Central educational boards. There is increasing use of Sanskrit on social media as well. The same is reflected in EGIDS scale where Sanskrit is categorized in Scale 4 indicating status of the language as “Educational”.

⁴ Though listed in EGIDS scale 4, Saraiki is not covered by the NBGP. As per Ethnologue, the Devanagari script is "no longer in use" by the Saraiki community.

Ref: <https://www.ethnologue.com/language/skr>

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

3.3.1 The Consonants

Devanagari consonants have an implicit schwa⁵ /ə/ vowel included in them. As per traditional classification, they are categorized according to their phonetic properties (especially in terms of place plus manner of articulation). 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

⁵Although representing the implicit vowel as /a/ is more correct orthographically, the schwa /ə/, although not part of the orthographic system has been used since the /a/ would be misunderstood and read as अ/आ/ा.

3.3.2 The Implicit Vowel Killer: Halant⁶

All consonants contain an implicit vowel (schwa). 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 empirical. It is just an observation drawn from the words that have been observed to date. Given the confluence of languages happening 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. Hence, in the LGR work, this limit will not be enforced⁷.

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 sign (Matra) is attached to the consonant. Since the consonant has a built-in schwa, there are equivalent Matras for all vowels excepting the ॐ.

The correlation is shown as follows:

Vowel	Corresponding vowel sign (Matra)
अ U+0905	
आ U+0906	ा U+093E
इ U+0907	ि U+093F

⁶ 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.

⁷This can be the case when a foreign language word, which admits a large number of consonants, is transliterated into Devanāgarī

ई U+0908	ी U+0940
उ U+0909	ु U+0941
ऊ U+090A	ू U+0942
ऋ U+090B	ृ U+0943
ए U+090F	े U+0947
ऐ U+0910	ै U+0948
ओ U+0913	ो U+094B
औ U+0914	ौ U+094C
अं U+0973	ं U+093A
अं U+0974	ं U+093B
ऐ/अं U+090E/ U+0904	ं U+0946
ओ U+0912	ो U+094A
ऐ/अं U+090D/ U+0972	ं U+0945
ऋ U+0960	ृ U+0944
ऑ U+0911	ॉ U+0949
औ U+0975	ौ U+094F

ॐ U+0976	ॐ U+0956
ॐ U+0977	ॐ U+0957

Table 5: Vowels with corresponding Matras

Marathi uses ॐ (U+0972) instead of ॐ (U+090D).

3.3.4 The Anusvara (ँ - U+0902)

The Anusvara represents a homorganic nasal. It replaces a conjunct group of a Nasal Consonant + Halant + Consonant belonging to that particular varga. Before a non-varga consonant the Anusvara represents a nasal sound. Modern Hindi, Marathi and Konkani languages prefer the Anusvara to the corresponding Half-nasal⁸:

सन्त vs. संत /sənt/saint

चम्पा vs. चंपा /tʃəmpa/ A flower: belonging to the genus Plumeria family

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: Candrabindu (ँ - U+0901)

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

3.3.6 Nukta (ँ - U+093C)⁹

The Nukta sign is placed below a certain number of consonants to represent sounds found only in words borrowed from Perso-Arabic. It is pre-dominantly used in this manner in Bodo, Hindi, Kashmiri, Maithili, Santali, Sindhi and Tamang. It can be adjoined to

⁸ A half-nasal is used in epigraphy to indicate a nasal consonant conjoined to its corresponding “Varga” through a Halant.

⁹The possible sets of consonants/vowels have been derived from various sources viz. Prior research carried out by Centre for Development of Advanced Computing's [C-DAC] Graphics Intelligence based Script Technologies [GIST] Research Labs (https://cdac.in/index.aspx?id=mlc_gist_about), Omniglot and inputs provided by various experts on-board the NBGP for specific languages. Only Omniglot references have been provided as they are available online.

"क"(U+0915), "ख"(U+0916), "ग"(U+0917), "ज"(U+091C) and "फ"(U+092B) 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) to indicate flapped sounds, e.g.:

बढ़ /bādh/(U+092C U+0922 U+093C)

Web Publication "DEVANĀGARĪ ALPHABET AND ITS ROMANIZATION" [109] by the Central Hindi Directorate, Ministry of HRD, Government of India, clearly states such a use of Nukta in Hindi.

In Bodo the Nukta is adjoined to "ड"(U+0921) [110]. In Maithili it is adjoined to "क" (U+0915), "ज" (U+091C), "ड" (U+0921) and "ढ" (U+0922) [111]. In Sindhi, it is adjoined to "ख" (U+0916), "ग" (U+0917), "ज" (U+091C), "फ" (U+092B), "ड" (U+0921) and "ढ" (U+0922) [104].

In Kashmiri, it can also be adjoined to "च" (U+091A), "छ" (U+091B) and "ज" (U+091C) [108] to indicate the laterally released affricates.

चाय /čāy/tea (U+091A U+093C U+093E U+092F)

छल /čhal/wash-Imperative (U+091B U+093C U+0932)

पोज़ /póz/fact (U+092A U+094A U+091C U+093C)

Normally a Nukta is appended to a Consonant. However, the Santali language uses Nukta in a unique way. The Nukta is adjoined to following vowels and vowel signs:

- a. आ (U+0906)
- b. ओ (U+0913)
- c. ा (U+093E)
- d. ो (U+094B)

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

The Visarga is frequently used in Sanskrit and represents a sound very close to /h/, for example: दुःख /dukh/ sorrow, unhappiness (U+0926 U+0941 U+0903 U+0916).

The Avagraha "s" (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.

3.3.8 Zero Width Non-joiner (U+200C) and Zero Width Joiner (U+200D)

The Zero Width Non-joiner (ZWNJ) is an invisible character used in certain cases (after Halant) where default conjunct formation is to be explicitly restricted and the Halant joining the two consonants participating in the conjunct formation needs to be explicitly shown. For example, the conjunct क्ष /ksha/ which gets formed by क /ka/ + ् (halant) + ष /sha/ gets rendered as क्ष – when formed by क /ka/ + ् (halant) + Zero Width Non-joiner + ष /sha/. In certain cases, for certain communities, this visual rendition creates a difference in the manner in which those combinations are pronounced.

The Zero Width Joiner (ZWJ) is another invisible character which is used in certain cases (mostly after Halant) in which a particular conjunct combination gets rendered such that constituting consonant shapes may not be directly visible in the conjunct shape. For example, the conjunct क्ष /ksha/ which gets formed by क /ka/ + ् (halant) + ष /sha/ does not show half form of ka joining with sha. However, using ZWJ, the constituting consonant's shapes are preserved in the visual depiction: कष – formed by क /ka/ + ् (halant) + Zero Width Joiner + ष /sha/.

Earlier the ZWJ was recommended by the Unicode Consortium to be used to generate certain special conjuncts like Eyelash Ra (more details in Section 5.2).. However, with the new recommendations in place, this usage of ZWJ is now not encouraged.

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 has been assigned a separate LGR; however, the Neo-Brahmi GP ensured 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, mostly 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. 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 community about its usage in the language.

4.1.2 Exclusion principles

The main exclusion principle is that of External Limits on Scope. These comprise protocols or standards that 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 *External Limits on Scope*

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 Standard

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 the Unicode consortium.

ii. IDNA Protocol

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

For 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.

IDNA also imposes restrictions on invisible characters Zero Width Non-Joiner (U+200C) and Zero Width Joiner (U+200D) in the form of CONTEXTJ rules. These are required in certain cases where a typical visual shape of an akshar is desired.

In domain names, due to absence of space " " or tab "-", there will be cases where inability to use ZWNJ can pose some issues where two words need to be joined together where previous word needs to end in an Explicit Halant and the next word begins with a consonant. In that case, a conjunct will be formed between last consonant of the first word

and the first consonant of the second word. This visual display may not be desired. For example, if two words देश् (/deṣh/ nation) and विदेश (/videṣh/ foreign land) are juxtaposed to each other, the resultant word i.e. “देश्विदेश”¹⁰ is not the appropriate way of rendering it. Appropriate rendering of the same would be “देशविदेश” which can be achieved by adding a ZWNJ in between the two words.

As the ZWNJ is not part of the MSR, it is not permissible to make such combinations. If and when the ZWNJ is permitted by the MSR, the then NBGP may consider adding it to the Devanagari repertoire if necessary.

However, there may not be much of an impact of exclusion of the ZWJ from MSR as there are better alternatives already available for depicting the cases for which ZWJ was earlier used. Some specific shapes¹¹ may not be able to be made, however there will not be any impact on the phonetic level.

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 Zone LGR Procedure introduces additional exclusions on IDNA allowed set of characters. For example, the Devanagari Sign Avagraha "s" (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.

¹⁰ In this particular case though it is possible to get the required display by dropping the explicit Halant at the end of the word, however in that case, one can argue that the pronunciation of the two words i.e. देश् and देश is different and hence it changes the fundamental word.

¹¹ Case of क्ष and क्ष: the first is composed with क्+्ष while the latter is with क्+्ष+ZWJ+ष. The pronunciation of both the conjuncts is same.

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 "८" (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 "ठ" (U+0960) and DEVANAGARI LETTER VOCALIC LL "ड" (U+0961) as well as their Matra forms "ॠ" (U+0944) and "ॡ" (U+0963). No such characters will be included. This is in compliance with the Conservatism 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 compliance with the Letter principle as laid down in the Root Zone LGR procedure.

4.2 Methodology to incorporate the feedback received through Public Comment process:

The Devanagari script LGR proposal was published for public comment to allow those who had not participated in the NBGP to make their views known. The Devanagari LGR received various comments during the public comments process. Most of the comments received were of the editorial nature. Some comments demanded the attention to the normative section of the document. The NBGP at-large and the Devanagari team in specific, went through the comments in detail and decided for each of the individual comments received if it needed a change in the overall LGR recommendation.

Wherever the Devanagari team decided that a change was necessary, the change was made. Rests of the comments were addressed by a detailed explanation about why the said change is not necessary. An elaborate document with all such explanations was shared with the NBGP at-large. On overall agreement of the entire NBGP, the Devanagari LGR was finalized. The analysis of public comments can be accessed online given at [114].

5 Repertoire

Section 5.1 provides the section of the [MSR] applicable to the Devanagari script on which the Devanagari code point repertoire is based. Section 5.2 details the code point repertoire that the Neo-Brahmi Generation Panel [NBGP] proposes to be included in the Devanagari LGR.

5.1 Devanagari section of Maximal Starting Repertoire [MSR] Version 4

Devanagari

	090	091	092	093	094	095	096	097
0	ॐ	ऐ	ठ	र	ी	ॐ	ऋ	०
1	ँ	ऑ	ड	र	ु	ं	ऌ	ं
2	ं	ओ	ढ	ल	्र	ॠ	ॡ	ँ
3	ः	ओ	ण	ळ	ॠ	ॡ	ॢ	अं
4	ऐ	औ	त	ळ	ॠ	ॡ	।	आं
5	अ	क	थ	व	ॠ	ॡ	॥	औ
6	आ	ख	द	श	ॠ	ॡ	०	अु
7	इ	ग	ध	ष	ॠ	ॡ	१	अु
8	ई	घ	न	स	ॠ	ॡ	२	रु
9	उ	ङ	न्	ह	ँ	ख	३	ज़
A	ऊ	च	प	ं	ो	ग	४	ष
B	ऋ	ॡ	फ	ा	ो	ज	५	ग
C	ऌ	ॢ	ब	्	ौ	ड़	६	ज़
D	ँ	झ	भ	ऽ	्	ढ	७	१
E	ऐ	ज	म	ा	ि	फ़	८	ड
F	ए	ट	य	ि	ौ	य़	९	ब

Color convention¹²:

All characters that are included in the [MSR] - Yellow background

PVALID in IDNA2008 but excluded from the MSR for various reasons - Pinkish background

Not PVALID in IDNA2008 - White background

Figure 2: Devanagari Code Page from [MSR]

¹²This document needs to be printed in color for this to be read correctly.

5.2 Code Point Repertoire

For each of the code points, language references have been given in the last column titled "Reference". For the entire coverage of Devanagari code points, references of Hindi, Marathi, Sanskrit, Sindhi and Kashmiri have been given. Though only five representative languages have been chosen for referencing, they together cover all the code points required for all the languages that NBGP has considered as given in Section 3.2.

Sr. No.	Unicode Code Point	Glyph	Character Name	Category	Example languages using the code point (Not exhaustive list)	Language with lowest EGIDS scale using the code point	Reference
1.	0901	◌ं	DEVANAGARI SIGN CANDRABINDU	Candrabindu	Bodo, Hindi, Kashmiri, Konkani, Maithili, Marathi, Nepali, Santali and Sanskrit	1 Hindi, Nepali	[0], [101], [102], [103], [105], [108], [110], [111], [112], [113]
2.	0902	◌ँ	DEVANAGARI SIGN ANUSVARA	Anusvara (Bindu)	Most of the languages given in section 3.2	1 Hindi, Nepali	[0], [101], [102], [103], [113]
3.	0903	◌ः	DEVANAGARI SIGN VISARGA	Visarga	Most of the languages given in section 3.2	1 Hindi, Nepali	[0], [101], [102], [103], [113]
4.	0905	अ	DEVANAGARI LETTER A	Vowel	Most of the languages given in section 3.2	1 Hindi, Nepali	[0], [101], [102], [103], [104], [113]
5.	0906	आ	DEVANAGARI LETTER AA	Vowel	Most of the languages given in section 3.2	1 Hindi, Nepali	[0], [101], [102], [103], [104], [113]
6.	0907	इ	DEVANAGARI LETTER I	Vowel	Most of the languages given in section 3.2	1 Hindi, Nepali	[0], [101], [102], [103], [104], [113]
7.	0908	ई	DEVANAGARI LETTER II	Vowel	Most of the languages given in section 3.2	1 Hindi, Nepali	[0], [101], [102], [103], [104], [113]

8.	0909	उ	DEVANAGARI LETTER U	Vowel	Most of the languages given in section 3.2	1 Hindi, Nepali	[0], [101], [102], [103], [104], [113]
9.	090A	ऊ	DEVANAGARI LETTER UU	Vowel	Most of the languages given in section 3.2	1 Hindi, Nepali	[0], [101], [102], [103], [104], [113]
10.	090B	ऋ	DEVANAGARI LETTER VOCALIC R	Vowel	Hindi, Marathi, Sanskrit	1 Hindi	[0], [101], [102], [103]
11.	090D	ऌ	DEVANAGARI LETTER CANDRA E	Vowel	Hindi	1 Hindi	[0], [101]
12.	090E	ऐ	DEVANAGARI LETTER SHORT E	Vowel	Kashmiri	4 Kashmiri	[0], [105], [108]
13.	090F	ए	DEVANAGARI LETTER E	Vowel	Most of the languages given in section 3.2	1 Hindi, Nepali	[0], [101], [102], [103], [104], [105], [108], [113]
14.	0910	ऐ	DEVANAGARI LETTER AI	Vowel	Most of the languages given in section 3.2	1 Hindi, Nepali	[0], [101], [102], [103], [104], [105], [108], [113]
15.	0911	ऑ	DEVANAGARI LETTER CANDRA O	Vowel	Hindi, Konkani, Marathi, Kashmiri	1 Hindi	[0], [100], [101], [102], [108], [112]
16.	0912	औ	DEVANAGARI LETTER SHORT O	Vowel	Kashmiri	4 Kashmiri	[0], [105], [108]
17.	0913	ओ	DEVANAGARI LETTER O	Vowel	Most of the languages given in section 3.2	1 Hindi, Nepali	[0], [101], [102], [103], [104], [105], [108], [113]
18.	0914	औ	DEVANAGARI LETTER AU	Vowel	Most of the languages given in section 3.2	1 Hindi, Nepali	[0], [101], [102], [103], [104], [105], [108], [113]
19.	0915	क	DEVANAGARI LETTER KA	Consonant	Most of the languages given in section 3.2	1 Hindi, Nepali	[0], [101], [102], [103], [104], [105], [108], [113]

20.	0916	ख	DEVANAGARI LETTER KHA	Consonant	Most of the languages given in section 3.2	1 Hindi, Nepali	[0], [101], [102], [103], [104], [105], [108], [113]
21.	0917	ग	DEVANAGARI LETTER GA	Consonant	Most of the languages given in section 3.2	1 Hindi, Nepali	[0], [101], [102], [103], [104], [105], [108], [113]
22.	0918	घ	DEVANAGARI LETTER GHA	Consonant	Most of the languages given in section 3.2	1 Hindi, Nepali	[0], [101], [102], [103], [104], [113]
23.	0919	ङ	DEVANAGARI LETTER NGA	Consonant	Most of the languages given in section 3.2	1 Hindi, Nepali	[0], [101], [102], [103], [113]
24.	091A	च	DEVANAGARI LETTER CA	Consonant	Most of the languages given in section 3.2	1 Hindi, Nepali	[0], [101], [102], [103], [104], [105], [108], [113]
25.	091B	छ	DEVANAGARI LETTER CHA	Consonant	Most of the languages given in section 3.2	1 Hindi, Nepali	[0], [101], [102], [103], [104], [105], [108], [113]
26.	091C	ज	DEVANAGARI LETTER JA	Consonant	Most of the languages given in section 3.2	1 Hindi, Nepali	[0], [101], [102], [103], [104], [105], [108], [113]
27.	091D	झ	DEVANAGARI LETTER JHA	Consonant	Most of the languages given in section 3.2	1 Hindi, Nepali	[0], [101], [102], [103], [104], [113]
28.	091E	ञ	DEVANAGARI LETTER NYA	Consonant	Most of the languages given in section 3.2	1 Hindi, Nepali	[0], [101], [102], [103], [113]
29.	091F	ट	DEVANAGARI LETTER TTA	Consonant	Most of the languages given in section 3.2	1 Hindi, Nepali	[0], [101], [102], [103], [104], [105], [108], [113]

30.	0920	ठ	DEVANAGARI LETTER TTHA	Consonant	Most of the languages given in section 3.2	1 Hindi, Nepali	[0], [101], [102], [103], [104], [105], [108], [113]
31.	0921	ड	DEVANAGARI LETTER DDA	Consonant	Most of the languages given in section 3.2	1 Hindi, Nepali	[0], [101], [102], [103], [104], [105], [108], [113]
32.	0922	ढ	DEVANAGARI LETTER DDHA	Consonant	Most of the languages given in section 3.2	1 Hindi, Nepali	[0], [101], [102], [103], [104], [113]
33.	0923	ण	DEVANAGARI LETTER NNA	Consonant	Most of the languages given in section 3.2	1 Hindi, Nepali	[0], [101], [102], [103], [104], [113]
34.	0924	त	DEVANAGARI LETTER TA	Consonant	Most of the languages given in section 3.2	1 Hindi, Nepali	[0], [101], [102], [103], [104], [105], [108], [113]
35.	0925	थ	DEVANAGARI LETTER THA	Consonant	Most of the languages given in section 3.2	1 Hindi, Nepali	[0], [101], [102], [103], [104], [105], [108], [113]
36.	0926	द	DEVANAGARI LETTER DA	Consonant	Most of the languages given in section 3.2	1 Hindi, Nepali	[0], [101], [102], [103], [104], [105], [108], [113]
37.	0927	ध	DEVANAGARI LETTER DHA	Consonant	Most of the languages given in section 3.2	1 Hindi, Nepali	[0], [101], [102], [103], [104], [105], [108], [113]
38.	0928	न	DEVANAGARI LETTER NA	Consonant	Most of the languages given in section 3.2	1 Hindi, Nepali	[0], [101], [102], [103], [104], [105], [108], [113]
39.	092A	प	DEVANAGARI LETTER PA	Consonant	Most of the languages given in section 3.2	1 Hindi, Nepali	[0], [101], [102], [103], [104], [105], [108], [113]

40.	092B	फ	DEVANAGARI LETTER PHA	Consonant	Most of the languages given in section 3.2	1 Hindi, Nepali	[0], [101], [102], [103], [104], [105], [108], [113]
41.	092C	ब	DEVANAGARI LETTER BA	Consonant	Most of the languages given in section 3.2	1 Hindi, Nepali	[0], [101], [102], [103], [104], [105], [108], [113]
42.	092D	भ	DEVANAGARI LETTER BHA	Consonant	Most of the languages given in section 3.2	1 Hindi, Nepali	[0], [101], [102], [103], [104], [105], [108], [113]
43.	092E	म	DEVANAGARI LETTER MA	Consonant	Most of the languages given in section 3.2	1 Hindi, Nepali	[0], [101], [102], [103], [104], [105], [108], [113]
44.	092F	य	DEVANAGARI LETTER YA	Consonant	Most of the languages given in section 3.2	1 Hindi, Nepali	[0], [101], [102], [103], [104], [105], [108], [113]
45.	0930	र	DEVANAGARI LETTER RA	Consonant	Most of the languages given in section 3.2	1 Hindi, Nepali	[0], [101], [102], [103], [104], [105], [108], [113]
46.	0932	ल	DEVANAGARI LETTER LA	Consonant	Most of the languages given in section 3.2	1 Hindi, Nepali	[0], [101], [102], [103], [104], [105], [108], [113]
47.	0933	ळ	DEVANAGARI LETTER LLA	Consonant	Bodo, Konkani, Marathi, Nepali, Sanskrit	1 Nepali	[0], [102], [103], [110], [112], [113]
48.	0935	व	DEVANAGARI LETTER VA	Consonant	Most of the languages given in section 3.2	1 Hindi, Nepali	[0], [101], [102], [103], [104], [105], [108], [113]
49.	0936	श	DEVANAGARI LETTER SHA	Consonant	Most of the languages given in section 3.2	1 Hindi, Nepali	[0], [101], [102], [103], [104], [105], [108], [113]

50.	0937	ष	DEVANAGARI LETTER SSA	Consonant	Most of the languages given in section 3.2	1 Hindi, Nepali	[0], [101], [102], [103], [104], [113]
51.	0938	स	DEVANAGARI LETTER SA	Consonant	Most of the languages given in section 3.2	1 Hindi, Nepali	[0], [101], [102], [103], [104], [105], [108], [113]
52.	0939	ह	DEVANAGARI LETTER HA	Consonant	Most of the languages given in section 3.2	1 Hindi, Nepali	[0], [101], [102], [103], [104], [105], [108], [113]
53.	093A	ं	DEVANAGARI VOWEL SIGN OE	Matra	Kashmiri	4 Kashmiri	[11], [105], [108]
54.	093B	ऋ	DEVANAGARI VOWEL SIGN OOE	Matra	Kashmiri	4 Kashmiri	[11], [105], [108]
55.	093C	ँ	DEVANAGARI SIGN NUKTA	Nukta	Bodo, Hindi, Kashmiri, Maithili, Santali, Sindhi	1 Hindi	[0], [101], [105], [108], [110], [109], [111]
56.	093E	ा	DEVANAGARI VOWEL SIGN AA	Matra	Most of the languages given in section 3.2	1 Hindi, Nepali	[0], [101], [102], [103], [113]
57.	093F	ि	DEVANAGARI VOWEL SIGN I	Matra	Most of the languages given in section 3.2	1 Hindi, Nepali	[0], [101], [102], [103], [113]
58.	0940	ी	DEVANAGARI VOWEL SIGN II	Matra	Most of the languages given in section 3.2	1 Hindi, Nepali	[0], [101], [102], [103], [113]
59.	0941	ु	DEVANAGARI VOWEL SIGN U	Matra	Most of the languages given in section 3.2	1 Hindi, Nepali	[0], [101], [102], [103], [113]
60.	0942	ू	DEVANAGARI VOWEL SIGN UU	Matra	Most of the languages given in section 3.2	1 Hindi, Nepali	[0], [101], [102], [103], [113]
61.	0943	ृ	DEVANAGARI VOWEL SIGN VOCALIC R	Matra	Hindi, Marathi, Sanskrit	1 Hindi	[0], [101], [102], [103]

62.	0945	ँ	DEVANAGARI VOWEL SIGN CANDRA E = candra	Matra	Hindi, Konkani, Marathi, Sanskrit, Kashmiri	1 Hindi	[0], [100], [101], [108]
63.	0946	े	DEVANAGARI VOWEL SIGN SHORT E	Matra	Kashmiri	4 Kashmiri	[0], [105], [108]
64.	0947	ै	DEVANAGARI VOWEL SIGN E	Matra	Most of the languages given in section 3.2	1 Hindi, Nepali	[0], [101], [102], [103], [105], [108], [113]
65.	0948	ी	DEVANAGARI VOWEL SIGN AI	Matra	Most of the languages given in section 3.2	1 Hindi, Nepali	[0], [101], [102], [103], [113]
66.	0949	ॉ	DEVANAGARI VOWEL SIGN CANDRA O	Matra	Hindi, Konkani, Marathi, Kashmiri	1 Hindi	[0], [100], [108]
67.	094A	ो	DEVANAGARI VOWEL SIGN SHORT O	Matra	Kashmiri	4 Kashmiri	[0], [105], [108]
68.	094B	ी	DEVANAGARI VOWEL SIGN O	Matra	Most of the languages given in section 3.2	1 Hindi, Nepali	[0], [101], [102], [103], [105], [108], [113]
69.	094C	ौ	DEVANAGARI VOWEL SIGN AU	Matra	Most of the languages given in section 3.2	1 Hindi, Nepali	[0], [101], [102], [103], [105], [108], [113]
70.	094D	्	DEVANAGARI SIGN VIRAMA	Halant / Virama	Most of the languages given in section 3.2	1 Hindi, Nepali	[0], [101], [102], [103], [105], [108], [113]
71.	094F	ी	DEVANAGARI VOWEL SIGN AW	Matra	Kashmiri	4 Kashmiri	[0], [105], [108]
72.	0956	े	DEVANAGARI VOWEL SIGN UE	Matra	Kashmiri	4 Kashmiri	[11], [105], [108]
73.	0957	ै	DEVANAGARI VOWEL SIGN UUE	Matra	Kashmiri	4 Kashmiri	[11], [105], [108]
74.	0972	अँ	DEVANAGARI LETTER CANDRA A	Vowel	Konkani, Marathi, Kashmiri	2 Konkani, Marathi	[9], [100], [102], [108], [112]

75.	0973	अ	DEVANAGARI LETTER OE	Vowel	Kashmiri	4 Kashmiri	[11], [105], [108]
76.	0974	आ	DEVANAGARI LETTER OOE	Vowel	Kashmiri	4 Kashmiri	[11], [105], [108]
77.	0975	औ	DEVANAGARI LETTER AW	Vowel	Kashmiri	4 Kashmiri	[11], [105], [108]
78.	0976	उ	DEVANAGARI LETTER UE	Vowel	Kashmiri	4 Kashmiri	[11], [105], [108]
79.	0977	ऊ	DEVANAGARI LETTER UUE	Vowel	Kashmiri	4 Kashmiri	[11], [105], [108]
80.	097B	ग	DEVANAGARI LETTER GGA	Consonant	Sindhi	2 Sindhi	[8], [104]
81.	097C	ज	DEVANAGARI LETTER JJA	Consonant	Sindhi	2 Sindhi	[8], [104]
82.	097E	ड	DEVANAGARI LETTER DDDA	Consonant	Sindhi	2 Sindhi	[8], [104]
83.	097F	ब	DEVANAGARI LETTER BBA	Consonant	Sindhi	2 Sindhi	[8], [104]

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 for enabling inclusion of "Eyelash Reph"¹³ construct.

Sr. No.	Unicode Code Points	Sequence	Character Names	Example languages using the code-point (Not exhaustive list)	Reference
1.	0931 094D 092F	य	DEVANAGARI LETTER RRA DEVANAGARI SIGN VIRAMA DEVANAGARI LETTER YA	Konkani, Marathi, Nepali	[106], [107]

¹³ Unicode uses the term "Eyelash Ra" instead. Since the construct that is formed by this sequence is a special form of Reph (which is otherwise formed by Normal Ra U+0930), the term "Reph" is used here.

2.	0931	ह	DEVANAGARI LETTER RRA	Konkani, Marathi, Nepali	[106], [107]
	094D		DEVANAGARI SIGN VIRAMA		
	0939		DEVANAGARI LETTER HA		

Table 7: Sequences

5.3 Code points not included

The following code points have not been included in the repertoire.

Sr. No.	Unicode Code Point	Glyph	Character Name	Reason for exclusion
1.	U+0904	ॐ	DEVANAGARI LETTER SHORT A	Usage unknown. Not required explicitly by any language.
2.	U+090C	ॢ	DEVANAGARI LETTER VOCALIC L	Not in modern usage. Excluded as per conservatism principle.
3.	U+0929	ॣ	DEVANAGARI LETTER NNNA	Not required in any spoken language. Required only for transcribing Dravidian alveolar n.
4.	U+0934	।	DEVANAGARI LETTER LLLA	Not required in any spoken language. Required only for transcribing Dravidian l.
5.	U+0944	॥	DEVANAGARI VOWEL SIGN VOCALIC RR	Not in modern usage. Excluded as per conservatism principle.
6.	U+0979	३	DEVANAGARI LETTER ZHA	Not required in any spoken language. Required only in transliteration of Avestan.
7.	U+097A	ॣ	DEVANAGARI LETTER HEAVY YA	Usage unknown. Not required explicitly by any language.

5.4 Structural Formation of Devanagari:

All the languages written in Brahmi derived scripts follow a particular way of formation of their words, known as *akshar*. In the next section there are detailed akshar formation rules as applicable to representation of the Hindi language when written in the Devanagari Script. These rules need slight additions 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 Reph construct is required in Marathi, Konkani and Nepali but not in Hindi).

It is worth noting that the rules required for accommodation of additional languages in the Devanagari ruleset apart from those required for Hindi are never in conflict with one another.

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

5.5 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. These rules are based on an Indian Standard (IS 13194:1991) popularly known as "Indian Script Code for Information Interchange" [ISCII].

5.5.1 Variables involved

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

5.5.2 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.

5.5.3 The Vowel Sequence

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

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

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

Examples:

Sequence Description	Sequence	Example	Constituting characters
Vowel	V	अ /a/ U+0905	
Vowel + Anusvara	V[B]	अं /aṁ/ U+0905 U+0902	अ ँ U+0905 U+0902
Vowel + Candrabindu	V[D]	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

5.5.4 Consonant Sequence

A consonant sequence begins with a consonant. It may be optionally followed by a Nukta (N), Matra (M), Anusvara (B), Candrabindu (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	Constituting characters
Consonant	C	क /ka/ U+0915	<single character>
Consonant + Nukta	C[N]	क ः /kaḥ/ U+0915 U+093C	क ः U+0915 U+093C

Table 10

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

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

Examples:

Sequence Description	Sequence	Example	Constituting characters
Consonant + Matra	C[M]	कि /ki/	क ि U+0915 U+093F
Consonant + Anusvara	C[B]	कं /kaṁ/	क ँ U+0915 U+0902
Consonant + Candrabindu	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	Constituting characters
Consonant + Matra + Anusvara	CM[B]	कीं /kīṁ/	क ि ँ U+0915 U+0940 U+0902
Consonant + Matra + Candrabindu	CM[D]	काँ /kāṁ/	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 Halant¹⁴ *3(CH)C

Example:

Sequence Description	Sequence	Example	Constituting characters
Consonant + Halant + Consonant + Halant + Consonant + Halant + Consonant	CHCHCHC	न्क्रय /nkrya/	न्क्रय U+0928 U+094D U+0915 U+094D U+0930 U+094D U+092F

Table 13

However, in the WLE rules proposed in Section 7 do not impose any restriction on the number of consonants that can be joined by a Halant.

Subsets:

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

Example:

¹⁴ In case of Sanskrit, it can join upto 5 consonants.

Sequence Description	Sequence	Example	Constituting characters
Consonant + Halant + Consonant + Matra	CHC[M]	क्की /kkī/	क्की U+0915 U+094D U+0915 U+0940
Consonant + Halant + Consonant + Anusvara	CHC[B]	क्कं /kkām/	क्कं U+0915 U+094D U+0915 U+0902
Consonant + Halant + Consonant + Candrabindu	CHC[D]	क्कँ /kkām̐/	क्कँ U+0915 U+094D U+0915 U+0901
Consonant + Halant + Consonant + Visarga	CHC[X]	क्कः /kkāḥ/	क्कः U+0915 U+094D U+0915 U+0903

Table 14

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

Example:

Sequence Description	Sequence	Example	Constituting characters
Consonant + Halant + Consonant + Matra + Anusvara	CHCM[B]	क्कीं /kkīm/	क्कीं U+0915 U+094D U+0915 U+0940 U+0902
Consonant + Halant + Consonant + Matra + Candrabindu	CHCM[D]	क्कीँ /kkīm̐/	क्कीँ U+0915 U+094D U+0915 U+0940 U+0901
Consonant + Halant + Consonant + Matra + Visarga	CHCM[X]	क्कीः /kkīḥ/	क्कीः U+0915 U+094D U+0915 U+0940 U+0903

Table 15

These are the basic akshar formation rules on which the overall Devanagari LGR is based. As languages other than Hindi are considered, some additional 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 that look exactly alike. However, Devanagari has ample cases of confusingly similar variants. The NBGP categorizes these confusingly similar variants in two groups.

Group1: Confusing due to pure visual similarity

Group2: Confusing due to 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. "Table 21: Visually confusables" in "Appendix A: Visually confusable characters/sequences" lists them.

Cases which belong to Group 2, however, are proposed to be considered as variants. These cases are not of mere visual similarity as they involve some deviations from the widely accepted norms of Devanagari akshar formations. 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

The Santali language has a unique requirement for Nukta character "◌̣"(U+093C) positioning, which is not common in other Devanagari based languages. Santali requires the Nukta character to follow certain Vowels and Matras. Complete representation of these Santali combinations necessitated the Whole Label Evaluation rules (given in the Section 6.1) to be opened up for these specific cases. A regular non-Santali user mostly cannot even anticipate the possibility of such a combination and can confuse it for something else.

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

Variant 1	Variant 2
अ U+0906	अ U+0906 U+093C
ओ U+0913	ओ U+0913 U+093C
ः U+093E	ः U+093E U+093C
ो U+094B	ो U+094B U+093C

Table 16: Proposed Variants - Set 1

6.1.1 Variant context rule for Santali Nukta variants:

All of the Nukta variants given in "Table 16: Proposed Variants - Set 1" have a typical characteristic which is, within a variant pair, **Variant 1** is a subset of the **Variant 2**, e.g. in the first pair, अ (U+0906) is a subset of अ (U+0906 U+093C). This implies a regenerative tendency, in theory, i.e. if an अ (U+0906) is substituted with अ (U+0906 U+093C), it introduces a new instance of अ (U+0906) as seen here in bold: अ (**U+0906** U+093C). By definition, this new case of अ (U+0906) may also need to be substituted with अ (U+0906 U+093C) thereby creating an invalid akshar combination अ (U+0906 **U+093C U+093C**) where a Nukta will need to follow another Nukta. To prevent this, a variant context rule has been added to all the above nukta variants as given below.

Rule: As per the "Table 16: Proposed Variants - Set 1" the **Variant 1** to **Variant 2** relationship exists if and only if any of the **Variant 1** set character is not followed by a Nukta (U+093C) character. Thus, following variant relations are bound by the above condition:

अ (U+0906) → अ (U+0906 U+093C)

ओ (U+0913) → ओ (U+0913 U+093C)

ः (U+093E) → ः (U+093E U+093C)

ो (U+094B) → ो (U+094B U+093C)

The variant relationship from **Variant 2** to **Variant 1** should be equally constrained for two reasons. First, a variant is uniquely defined by both the variant mapping and the context condition imposed on it. (See [RFC 7940]). In order to maintain a symmetric

definition of variants, it is necessary to define both forward and symmetric variants using the same condition. (See also [RFC 8828]). Second, this type of variant pair is an “effective null variant” where the U+093C in one sequence maps to “nothing” in the other. In order to maintain a fully transitive system of variant definitions, it is necessary to prevent a label like U+0906 U+093C U+093C from having a variant U+0906 U+093C. The same condition would ensure this second constraint. However, as sequences of U+093C followed by U+093C are already invalid due to context rules on the Nukta, the condition is only required for the first reason in this case.

6.1.2 Overlapped variant analysis involving Nukta:

Considering the following variant sets A, B, C, D. Each of them contains 0906 or 093E.

Set	Mapping
Variant Set A	093E 0901 <--> 0949 0902
Variant Set B	0906 0901 <--> 0911 0902
Variant Set C	0906 0902 <--> 0974
Variant Set D	093B <--> 093E 0902

Overlapping variant sets involving 0906 and 093E plus Nukta

Source	Glyph	Target	Glyph		Type	Variant Context
0906	ॐ	0906 093C	ॐ	↔	blocked	not: followed-by-N
093E	ॐ	093E 093C	ॐ	↔	blocked	not: followed-by-N

When substituting the variant from the overlapping sets for 0906 or 093E respectively, into the left hand side sequences of Variant Sets A, B, C and E, the result is a valid sequence that displays with a Nukta. As the presence/absence of Nukta is the basis for several variant sets, therefore these variant are extended to ensure the symmetry and transitivity:

- 093E 093C 0901 is added to Set A,
- 0906 093C 0901 is added to Set B,
- 0906 093C 0902 is added to set C, and
- 093E 093C 0902 is added to Set D.

For set C, the implicit code point context of 0902 and 0974 are not equal: 0902 can be followed by Vowels and Consonants only, but 0974 can also be followed by 0901, 0902, 0903. (Both can be at

the end of the label.) Therefore, the variant mapping should receive a context rule: when(followed-by-V-C-or-end). This matches the intersection between these contexts.

Likewise, for set D, 0902 can be followed by Vowels and Consonants only, but 093B can also be followed by 0901, 0902, 0903 (Both can be at the end of the label). Therefore, the variant mapping should receive a context rule: when(followed-by-V-C-or-end).

The conclusion of these variant sets and variant contextual rules are:

Set	Mapping	Variant Contextual Rule
Variant Set A	093E 0901 <--> 093E 093C 0901 <--> 0949 0902	-
Variant Set B	0906 0901 <--> 0906 093C 0901 <--> 0911 0902	-
Variant Set C	0906 0902 <--> 0906 093C 0902 <--> 0974	when(followed-by-V-C-or-end)
Variant Set D	093B <--> 093E 0902 <--> 093E 093C 0902	when(followed-by-V-C-or-end)

Additional Notes

1. Overlapped variants involving Candrabindu, 0901 <--> 0945 0902, is technically overlapped with the four sets above. However, because the variant context rule “when(follows-only-C-or-N)” (See 6.4.1) , none of the sequences lead to a variant where 0901 is expanded to 0945 0902.
2. Another overlapped variant set with these four sets is 0902 (B, Anusvara) <- -> 093A (M, Matra). However, they are all invalid as a Matra can only follow C or N, while all sequences including 0902 as second element have V or M as first element.

6.2 Unique Vowels and Vowel Signs required for Kashmiri

Kashmiri, when written in Devanagari script, requires a unique set of Vowels and Vowel signs which only a Kashmiri speaker can understand. The majority of Devanagari users who are not conversant with Kashmiri can easily confuse them with some of the Vowels / Vowel signs which look similar to the Kashmiri ones. There are also cases where a Kashmiri Vowel / Vowel signs can be confused with certain akshar formations. Hence, they are being proposed as variants.

Variant 1	Variant 2
ॐ U+0973	ॐ U+0905 U+0902
ॐ U+093A	ॐ U+0902

अं U+0974	अं U+0906 U+0902
ीं U+093B	ीं U+093E U+0902
ँ U+090E	ँ U+0910
ं U+0946	ं U+0947
अं U+0975	अं U+0914
ीं U+094F	ीं U+094C

Table 17: Proposed Variants - Set 2

No variant contexts are required for these mappings, but the code point sequences introduced as mapping will need to be listed in the repertoire with matching code point context so that both source and target of the variant mapping have matching contexts (not preceded by H).

6.3 Halant in Final Position (Only a discussion, not proposed as variants)

Another case of deceptive similarity to a majority of the Devanagari user base is of a word ending in Halant "ं" (U+094D) vis-à-vis the same word without the final Halant. As the function of Halant is of a vowel killer, coming at the end, many users tend to ignore the phonetic effect of its presence/absence. The majority of users would pronounce both words in the same way, thereby creating a perception of (false) equivalence. However, there also exist some users who clearly require the final Halant to achieve the peculiar phonetic effect of a truncated implicit vowel sound in the end. These users make a clear distinction between the two words (with and without the final Halant). It is for this reason that the final Halant is being accommodated in the Whole Label Evaluation rules for Devanagari.

In these cases, the presence or absence of final Halant is clearly visible, and there is no apparent case to make them variant pairs. Eventually, in the light of practical experience, a future NBGP revision may assess if these cases need to be considered as variant pairs.

6.4 Variants based on Candrabindu and Candra Vowel Signs followed by Anusvara

This is a case of pairs of similar looking variants involving a Candrabindu ँ (U+0901). In Devanagari, there are two Candra vowel signs viz. ँ (U+0945) and ऌ (U+0949) which when succeeded by an Anusvara ं (U+0902) create a shape which resembles a Candrabindu ँ (U+0901). This gives rise to pairs which get rendered exactly alike in many fonts. Though some fonts can render them differently, the behavior is not consistent and keeps room for ambiguity. The actual pairs of the variants are as follows:

Variant 1	Variant 2
ँ U+0901	ँँ U+0945 U+0902
ँँ U+093E U+0901	ँँँ U+0949 U+0902
ँँ U+0905 U+0901	ँँँ U+0972 U+0902
ँँ U+090F U+0901	ँँ U+090D U+0902
ँँ U+0906 U+0901	ँँ U+0911 U+0902

Table 18: Proposed Variants - Set 3


As the first two suggested pairs are composed only of the dependent signs, they do not get properly joined in the absence of an independent character like Consonant or a Vowel before the same. For the sake of clarity, both the pairs preceded by a Devanagari Letter Ka (क - U+0915) are shown here:

Variant Pair 1: कँ (U+0901) and कँ (U+0945 U+0902)









Variant Pair 2: कँ (U+0949 U+0902) and कँ (U+0949 U+0902)




Ideally, the case of U+0945 U+0902 should be rendered as ँँ and the case of (U+0949



ँँ

U+0902) should be rendered as  where the Anusvara is clearly separated from the Candra shape. However, not all fonts follow the same convention and many of them render the two shapes exactly like a Candrabindu. This gives rise to an ambiguity and hence the variant possibility.

6.4.1 Variant context rule for Candrabindu and Candra-Anusvara variant pair

The variant pair  (U+0901) -  (U+0945 U+0902) necessitates that for creating a variant label, every Candrabindu  (U+0901) in a label be replaced by the sequence  (U+0945 U+0902). It should be noted that the beginning of the said sequence is with a Matra sign i.e.  (U+0945). While the Candrabindu  (U+0901) can be preceded by either of Consonants, Vowels, Nukta or a Matra, the same is not true for the  (U+0945) which is a Matra. A Matra can only be preceded by a consonant or a consonant followed by a Nukta. To handle this, a variant context rule has been added to  (U+0945 U+0902) as given below.

Rule: As per the "Table 18: Proposed Variants - Set 3" the variant relationship between the pair  (U+0901) -  (U+0945 U+0902) exists if and only if Candrabindu  (U+0901) is preceded by a Consonant or a Consonant followed by a Nukta.

There is no additional constraint on the preceding character of  (U+0945 U+0902), as the Candrabindu can be preceded by any character which can precede the sequence  (U+0945 U+0902). However, to express the mapping, the sequence U+0945 U+0902 is formally part of the repertoire, and as such must have a context rule that matches the context rule for U+0945, which happens to be the same context rule as for the variant mapping. For the same symmetry reason as discussed in Section 6.1.1 above, the formal definition of the variant mapping (U+0945 U+0902) -> (U+0901) has the same variant context condition to that for the mapping (U+0901) -> (U+0945 U+0902).

6.5 Variant Disposition

As variants mentioned in Table 16, Table 17 and Table 18 are confusingly similar, albeit of a peculiar nature, it is proposed that they be considered of "blocked" nature.

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

6.6 Cross-script Variants

A cross-script variant, also sometimes referred to as "Whole Label variant", is the variant case where one label in one script can be composed in such a way that it resembles another entire label in a different script.

Every individual LGR under NBGP is supposed to provide a set of cross script variants it identifies with all other scripts under NBGP.

NBGP has ensured that not only the individual characters but also most of the akshar variations are taken into consideration during the Cross-script variant analysis of Devanagari with all the other scripts under NBGP. This was achieved by sharing a list of most of the Devanagari akshar combinations with all the other script teams. (The word 'most' is used here as it is not practical to cover all the possible "Consonant + Halant + Consonant +" cases. However, for Devanagari, all cases of "Consonant + Halant + Consonant" combinations were included in the analysis.)

The Devanagari script has a major set of possible cross-script variants only with the Gurmukhi script. Cases listed in Table 19 are of the variants that are proposed to be cross-script variants between Devanagari and Gurmukhi. Similarly, Table 20 has the cases proposed to be cross-script variants between Devanagari and Bengali.

It is to be noted that none of the combinations listed in Table 19 and Table 20 are termed to be equivalents of each other semantically or otherwise. They are only grouped based on possible visual confusability.

NBGP has ensured that Devanagari, Bengali and Gurmukhi LGR teams propose a same set of cross-script variants by meeting face-to-face on many occasions as well as through mail communications. The same set of cross-script variants (with Devanagari) is supposed to be found in the Bengali and Gurmukhi LGR documents.

Devanagari	Gurmukhi
ं U+0902	ੰ U+0A02
इ U+0907	ਙ U+0A19
उ U+0909	ਤ U+0A24
ग U+0917	ਗ U+0A17
घ U+0918	ਬ U+0A2C
ट U+091F	ਟ U+0A1F
ठ U+0920	ਠ U+0A20
ढ U+0922	ਢ U+0A2B
प U+092A	ਧ U+0A27
भ U+092D	ਮ U+0A2E
म U+092E	ਸ U+0A38

व U+0935	𑖠 U+0A15
ह U+0939	𑖡 U+0A35
ं U+093A	𑖢 U+0A02
ः U+093C	𑖣 U+0A3C
ि U+093F	𑖤 U+0A3F
ी U+0940	𑖥 U+0A40
ँ U+0945	𑖦 U+0A71
े U+0946	𑖧 U+0A47
ै U+0946	𑖨 U+0A4B
े U+0947	𑖩 U+0A47
े U+0947	𑖪 U+0A4B

ै U+0948	ै U+0A48
ँ U+0956	ँ U+0A41
ं U+0957	ं U+0A42
ष्टि U+092A U+094D U+091F U+093F	ष्टि U+0A07
ष्टी U+092A U+094D U+091F U+0940	ष्टी U+0A08
ष्टे U+092A U+094D U+091F U+0947	ष्टे U+0A0F
ष्टै <u>092A 094D 091F 0946</u>	ष्टै U+0A0F
त्त U+0924 U+094D U+0924	त्त U+0A1C

Table 19: Proposed Cross-script Devanagari-Gurmukhi Variants

Devanagari	Bengali
म U+092E	ম U+09AE
ि U+093F	ি U+09BF

Table 20: Proposed Cross-script Devanagari-Bengali Variants

In addition to above cases, the Devanagari and Gurmukhi scripts have a possible set of cross-script confusables, which look similar but not similar enough to be recommended as cross-script variants. The "Table 22: Devanagari Cross-script confusables" in "Appendix B: Cross-script Confusables" lists them.

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 "Category" as mentioned in the Table 6: Code point repertoire.

C	→	Consonant
M	→	Matra
V	→	Vowel
B	→	Anusvara (Bindu)
D	→	Candrabindu
X	→	Visarga
H	→	Halant / Virama
N	→	Nukta
S	→	Eyelash Reph (C2HC3) where C2 is 0931 (ꣳ - DEVANAGARI LETTER RRA) H is 094D (् - DEVANAGARI SIGN VIRAMA) C3 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 a member of C1, V1 or M1.

The set C1 consists of these consonants:

- a. क (U+0915)
- b. ख (U+0916)
- c. ग (U+0917)
- d. च (U+091A)
- e. छ (U+091B)
- f. ज (U+091C)
- g. ङ (U+0921)
- h. ढ (U+0922)
- i. फ (U+092B)

The set V1 consists of these vowels:

- a. आ (U+0906)(Required in Santali language)
- b. औ (U+0913)(Required in Santali language)

The set M1 consists of these matras:

- a. ा (U+093E) (Required in Santali language)
- b. ो (U+094B) (Required in Santali language)

2. H: must be preceded by C or CN¹⁵
3. M: must be preceded by C or CN¹⁶
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")

¹⁵ where CN is a C followed by an N

¹⁶ where CN is a C followed by an N

Additional rules are used only for variants where a Nukta maps to a "null", or that are overlapped:

- Variant is not defined if followed by a Nukta (see 6.4.1)
- Variant undefined if it is not followed by V or C (including RRA) or end of label (See 6.1.2)

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

As any valid *akshar* in Devanagari begins either with a Consonant or a Vowel, in case of multi-words domains, it was necessary to check the compatibility of both of these to succeed any of the valid *akshar* ending character. It is to be noted that only the case "V preceded by H" needs a special discussion as given below.

There could be cases involving multi-word domains where V may need to be allowed to follow an H, e.g. आम्रअचार /a:məcha:r/ Mango pickle (U+0906 U+092E U+094D U+0905 U+091A U+093E U+0930).

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.

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

8 Contributors

NBGP Co-chairs: Dr. Udaya Narayan Singh, Mr. Mahesh D Kulkarni and Dr. Ajay Data

Following is the full list of NBGP members with their Language expertise.

Position	Name	Organization	Country	Language Expertise
Co-Chair	Ajay Data	Data Xgen Technologies	India	Hindi, English
Co-Chair	Mahesh D. Kulkarni	C-DAC	India	Marathi, Hindi, English
Co-Chair	Udaya Narayana Singh	Visva-Bharati, Santiniketan, West Bengal	India	Bengali, Maithili, Hindi, English
Member	Abhijit Dutta	Wikimedia	India	Bengali, Hindi
Member	Akshat S. Joshi (Editor)	C-DAC	India	Hindi, Marathi, English
Member	Anivar A. Aravind	Indic Project	India	Malayalam
Member	Anupam Agrawal	Tata Consultancy Service	India	Hindi, Bengali
Member	Arvind Bhandari	Gujarat University	India	Gujarati
Member	Ashish Modi	Data Xgen Technologies	India	Hindi
Member	Atiur Rahman Khan	C-DAC	India	Bangla
Member	Bal Krishna Bal	Kathmandu University	Nepal	Nepali
Member	Balaram Prasain	Tribhuvan University	Nepal	Nepali
Member	BASANTA KUMAR PANDA	Regional Institute of Education (NCERT)	India	Odia
Member	Bhim Dhoj Shrestha	Consultant	Nepal	Nepali, Newar
Member	Chitrita Chatterjee	Internet and Mobile Association of India (IAMAI)	India	Multiple languages represented by members of IAMAI
Member	DEBAJIT SHARMA	Anundoram Borooh Institute of Language Art and Culture	India	Assamese
Member	Dev Dass Manandhar	Consultant	Nepal	Nepali, Newar
Member	Dhanalakshmi KT	Northern Trust	India	Kannada
Member	Ganesh Murmu	Ranchi University	India	Santali
Member	Gangadhar Panday	Babul Films Society	India	Telugu
Member	Ghanashyam Nepal	Benares Hindu University & University of North Bengal	India	Nepali
Member	Girish Chandra Mishra	Language Technology Centre, Ravenshaw University	India	Odia
Member	Gurpreet Singh Lehal	Punjabi University Patiala	India	Panjabi
Member	Harish Chowdhary	NIXI	India	Hindi
Member	Hempal Shrestha	Nepal Entrepreneurs' Hub (NEHUB)	Nepal	Nepali, Newar
Member	Jay Paudyal	Consultant	India	Hindi

Member	Jijo Pappachan	DN. Domains	India	Malayalam
Member	K. C. Tikayat ray	Odia Bhasa Pratisthan	India	Odia
Member	Kalyan Vasudeo Kale	Formerly affiliated with University of Pune	India	Marathi
Member	Kuldeep Patnaik	Visualize thy soul	India	Odia
Member	Mukesh Saini	Essel Group	India	Hindi
Member	N. DeivaSundaram	NDS Lingsoft Solutions Pvt Ltd	India	Tamil
Member	Neha Gupta	C-DAC	India	Hindi, English
Member	Nirajan Parajuli	NREN	Nepal	Nepali
Member	Nishit Jain	C-DAC	India	Hindi, English
Member	Pawan Chitrakar	Gapsco	Nepal	Nepali
Member	Prabhakar Pandey	C-DAC	India	Hindi
Member	Prasad PK	A-one Publishers	India	Malayalam
Member	Prateek Pathak	ISOC Mumbai	India	Devanagari
Member	Raiomond Doctor	NLP Consultant	India	English, Hindi, Marathi, Gujarati
Member	Rajib Chakraborty	Society for Natural Language Technology Research	India	Bangla (Bengali)
Member	Rajiv Kumar	NIXI	India	
Member	S. Maniam	International Forum IT for Tamil	Singapore	Tamil
Member	Santhosh Thottingal	Wikimedia foundation	India	Malayalam, Sourashtra, Tamil
Member	Saroja Bhate	University of Pune	India	Sanskrit
Member	Shambhu Kumar Singh	National Translation Mission, Mysore	India	Maithili
Member	Shanmugam R	C-DAC	India	Tamil
Member	Shantaram S. Warde Walawalikar	Independent Researcher	India	Konkani
Member	Shashi Pathania	P.G.D. of Dogri, University of Jammu	India	Dogri
Member	Shubham Saran	NIXI	India	
Member	Sinnathambi Shanmugarajah	University of Colombo School of Computing	Sri Lanka	Tamil
Member	Sujith Kartha	Digitalkz.com	India	Malayalam
Member	Suraj Adhikari	Mercantile Communications (and .np ccTLD)	Nepal	Nepali
Member	Swarna Prabha Chainary	Guwahati University	India	Bodo
Member	U.B. Pavanaja	http://vishvakannada.com/	India	Kannada
Member	Uma Maheshwar G	CALTS, Univ. of Hyderabad	India	Telugu
Member	Uttam Shrestha Rana	NPNOG	Nepal	Nepali
Member	Veena Solomon	(freelancer)	India	Malayalam
Member	Vinay Murarka	Consultant; https://मेरा.भारत	India	Hindi

In addition, following members externally gave inputs to NBGP for the respective languages/scripts.

Name	Language/Script Expertise
Ajit Kumar	Awadhi, Braj Language
Amar Tummyahang	Limbu Language
Amrit Yonjan	Tamang Language
Aprana Kulkarni	Hindi, Marathi
Basil Baa	Sadri Language
Basil Kiro	Kharia Language
Biswa Limbu	Limbu Language
Devdass Manandhar	Newar
Devendra Kumar Devesh	Bhojpuri Language
Dinbandhu Mahto	Panchpargania Language
Dipika Sangma Narzary	Bodo Language
Dr K.P. Lekhwani	Sindhi
Dr. Birendra Kumar Soy	Mundari Language
Dr. Dinesh Kumar Shrivastav	Magahi Language
Dr. Harvinder Kaur	Gurmukhi Script
Dr. Laxmi Prasad Khatiwada	Nepali Language
Harihar Vaishnav	Halbi
Indra Kumar Tamang	Tamang Language
Jagannath Singh	Panchpargania Language
Narendra Kumar Negi	Kinnauri Language
Prateek Harshwal	Wagdi and Dhundhari Language
Rayem Olem Dungdung	Sadri Language
Tej Man Angdembe	Limbu Language
Urmila Harshwal	Wagdi Language

9 References

[MSR] Integration Panel, "Maximal Starting Repertoire — MSR-4 Overview and Rationale", 7 February 2019 <https://www.icann.org/en/system/files/files/msr-4-overview-25jan19-en.pdf> (Accessed on 18th Feb. 2019)

[EGIDS] Expanded Graded Intergenerational Disruption Scale, <https://www.ethnologue.com/about/language-status> (Accessed on 13th Nov. 2017)

[NBGP] Neo-Brahmi Generation Panel

[RFC7940] Davies, K. and A. Freytag, "Representing Label Generation Rulesets using XML", RFC 7940, August 2016, <https://tools.ietf.org/html/rfc7940>. (Accessed on 1st Dec. 2017)

- [RCF8228] A. Freytag, "Guidance on Designing Label Generation Rulesets (LGRs) Supporting Variant Labels, August 2017, <https://tools.ietf.org/html/rfc8228>. (Accessed on 1st Dec. 2017)
- [gTLD] generic Top Level Domain
- [ISCII] Indian Script Code for Information Interchange, https://cdac.in/index.aspx?id=mlc_gist_iscii (Accessed on 2ndFeb. 2018)
- [GIST] Graphics Intelligence based Script Technologies, <https://cdac.in/index.aspx?id=gist> (Accessed on 2ndFeb. 2018)
- [C-DAC] Centre for Development of Advanced Computing, <https://cdac.in> (Accessed on 2nd Feb. 2018)
- [0] The Unicode Standard 1.1, <http://www.unicode.org/versions/Unicode1.1.0/> (Accessed on 12th Dec. 2017)
- [8] The Unicode Standard 5.0, <http://www.unicode.org/versions/Unicode5.0.0/> (Accessed on 12th Dec. 2017)
- [9] The Unicode Standard 5.1, <http://www.unicode.org/versions/Unicode5.1.0/> (Accessed on 12th Dec. 2017)
- [11] The Unicode Standard 6.0, <http://www.unicode.org/versions/Unicode6.0.0/> (Accessed on 12th Dec. 2017)
- [100] Devanāgarī VIP Team. "Variant Issues Report", ICANN, 3rd Oct. 2011, <https://archive.icann.org/en/topics/new-gtlds/devanagari-vip-issues-report-03oct11-en.pdf> (Accessed on 10th Oct. 2017)
- [101] Omniglot, "Hindi", <https://www.omniglot.com/writing/hindi.htm> (Accessed on 10th Oct. 2017)
- [102] Omniglot, "Marathi", <https://www.omniglot.com/writing/marathi.htm> (Accessed on 10th Oct. 2017)
- [103] Omniglot, "Sanskrit", <https://www.omniglot.com/writing/sanskrit.htm> (Accessed on 10th Oct. 2017)

- [104] Omniglot, "Sindhi", <https://www.omniglot.com/writing/sindhi.htm>
(Accessed on 10th Oct. 2017)
- [105] Omniglot, "Kashmiri", <https://www.omniglot.com/writing/kashmiri.htm>
(Accessed on 10th Oct. 2017)
- [106] Unicode 10.0.0, "South and Central Asia-I - Official Scripts of India", Page 456 (R5 and R5a) ", <http://www.unicode.org/versions/Unicode10.0.0/ch12.pdf>
(Accessed on 13th Nov. 2017)
- [107] Unicode Indic Group, "Devanagari Eyelash Ra",
<http://unicode.org/~emuller/iwg/p8/utcdoc.html> (Accessed on 13th Nov. 2017)
- [108] M.K. Raina, "How to read and write Kashmiri in Devanagari?",
<http://www.koshur.org/pdf/Let%20Us%20Learn%20Kashmiri.pdf> (Accessed on 12th Dec. 2017)
- [109] Central Hindi Directorate-Ministry of HRD-Govt. of India, "Devanāgarī Alphabet and its Romanization",
http://hindinideshalaya.nic.in/english/hindi_organ/devnagarithesymbols.html (Accessed on 12th Dec. 2017)
- [110] Omniglot, "Bodo", <https://www.omniglot.com/writing/bodo.htm> (Accessed on 12th Dec. 2017)
- [111] Omniglot, "Maithili", <https://www.omniglot.com/writing/maithili.htm>
(Accessed on 12th Dec. 2017)
- [112] Omniglot, "Konkani", <https://www.omniglot.com/writing/konkani.htm>
(Accessed on 20th May. 2018)
- [113] Omniglot, "Nepali", <https://www.omniglot.com/writing/nepali.htm>
(Accessed on 20th May. 2018)
- [114] NBGP, "Public comment feedback for Devanagari, Gujarati, Gurmukhi Script LGR Proposals",
https://docs.google.com/document/d/1CLKdJBTNDcC_sFFs5s0a_Bk0zQUER2BIruYuyCNgkAw (Accessed on 18th Feb. 2019)

10 Books, articles and webographies consulted

Following is a thematically sorted set of documents, books, articles and webographies consulted in the drafting of this report

10.1 WRITING SYSTEMS

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

10.2 DEVANĀGARĪ

1. Agrawala, V. S. (1966). *The Devanāgarī script*. In: *Indian Systems of Writing*. (Pp. 12-16) Delhi: Publications Division.
2. Agyeya, Sacchidanand Hiranand Vatsyayan. 1972. *Bhavanti*. Delhi: Rajpal and Sons.
3. 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.]
4. Bhatia, Tej K. 1987. *A History of the Hindi Grammatical Tradition: Hindi-Hindustani Grammar, Grammarians, History and Problems*. Leiden/New York: E. J. Brill.
5. 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.
6. Cardona, George. 1987. *Sanskrit*. In *The World's Major Languages*. Bernard Comrie (ed.). London: Croom Helm. 448-469.
7. Dwivedi, Ram Awadh. 1966. *A Critical Survey of Hindi Literature*. Delhi: Motilal Banarsidass.
8. Faruqi, Shamsur Rahman. 2001. *Early Urdu Literary Culture and History*. Delhi: Oxford University Press.
9. Guru, Kamta Prasad. 1919. *Hindi Vyakaran*. Varanasi: Nagari Pracharini Sabha. (1962 edition).
10. Kachru, Yamuna. 1965. *A Transformational Treatment of Hindi Verbal Syntax*. London: University of London Ph.D. dissertation (Mimeographed).

11. Kachru, Yamuna. 1966. An Introduction to Hindi Syntax. Urbana: University of Illinois, Department of Linguistics.
12. Kalyan Kale and Anjali Soman, 1986. Learning Marathi. Shri Vishakha Prakashan, Pune :
13. McGregor, R. S. (1977). Outline of Hindi Grammar. 2nd ed. Delhi: Oxford University Press.
14. McGregor, R. S. 1972. Outline of Hindi Grammar with Exercises. Delhi: Oxford University Press.
15. McGregor, R. S. 1974. Hindi Literature of the Nineteenth and Early Twentieth Centuries. Wiesbaden: Harrassowitz.
16. McGregor, R. S. 1984. Hindi Literature from Its Beginnings to the Nineteenth Century. Wiesbaden: Harrassowitz.
17. Pandey, P. K. (2007). Phonology-orthography interface in Devanāgarī for Hindi. *Written Language and Literacy*, 10 (2): 139-156. 2007.
18. Rai, Amrit. 1984. A House Divided. The Origin and Development of Hindi/Hindavi. Delhi: Oxford University Press.
19. Sharad, Onkar. 1969. Lohiyake Vicar. Allahabad: Lokbharati Prakashan.
20. 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.
21. Sproat, R. (2000). A Computational Theory of Writing Systems. Cambridge University Press.
22. Tiwari, Pandit Udaynarayan. 1961. Hindi Bhasha ka Udgamaur Vikas [The Origin and Development of the Hindi Language]. Prayag: Leader Press.
23. Verma, M. K. 1971. The Structure of the Noun Phrase in English and Hindi. Delhi: Motilal Banarsidass.

10.3 INDIC COMPUTING SPECIFIC

1. IS 10401: 8-bit code for information interchange. 1982
2. IS 10315: 7-bit coded character set for information interchange. 1985

3. IS 12326: 7-bit and 8-bit coded character sets-Code extension techniques. 1987
4. ISO 15919, Information and documentation - Transliteration of Devanāgarī and related Indic scripts into Latin characters. 2001
5. ISO 2375: Procedure for registration of escape sequences. 2003
6. ISO 8859: 8-bit single-byte coded graphic character sets - Parts 1-13. 1998-2001
7. IDN POLICY <http://meity.gov.in/writereaddata/files/India-IDN-Policy.pdf>

11 Appendix A: Visually confusable characters/sequences

The Table 21 below shows characters / character sequences which may appear visually confusing to some of the users of the Devanagari script. However, they are not considered confusing enough to be categorized as variants.

Confusable 1	Confusable 2
क U+0915	क़ U+0915U+093C
ख U+0916	ख़ U+0916U+093C
ग U+0917	ग़ U+0917 U+093C
च U+091A	च़ U+091A U+093C
छ U+091B	छ़ U+091B U+093C
ज U+091C	ज़ U+091C U+093C
ड U+0921	ड़ U+0921 U+093C
ढ U+0922	ढ़ U+0922 U+093C
फ U+092B	फ़ U+092B U+093C

Table 21: Visually confusables

12 Appendix B: Cross-script Confusables

The Devanagari script has a major set of possible cross-script confusables with the Gurmukhi script. The Table 22 lists them.

In addition to Gurmukhi, some instances of cross-script confusable are found with Bengali, Gujarati, Telugu, Kannada, Malayalam and Sinhala.

None of the combinations listed in Table 22 are considered equivalents of each other, whether semantically or otherwise. They are only grouped based on possible visual confusability.

At first, they may not look exactly the same, however, in the given context e.g. in a browser bar as a part of a domain name, or as a single word where there is no surrounding text from the same script for distinguishing, they can create visual confusion.

A label can be considered to have a cross-script variant label only if "all" the constituent characters/aksharas have an equivalent confusable in the other script. If there is even one single character/akshara which does not have an equivalent visual confusable in other script, it essentially provides a visual distinction and hence a non-confusable string.

Devanagari confusable	Other script confusable	From script
ॐ U+0903	ॐ U+0A83	Gujarati
ॐ U+0903	ॐ U+0C03	Telugu
ॐ U+0903	ॐ U+0C83	Kannada
ॐ U+0903	ॐ U+0D03	Malayalam
ॐ U+0903	ॐ U+0A28	Sinhala

ः U+0903	ः ¹⁷ U+0983	Bengali
उ U+0909	ঔ U+0993	Bengali
घ U+0918	ষ U+0998	Bengali
ठ U+0920	ਨ U+0A28	Gurmukhi
ठ U+0920	ਰ U+0A30	Gurmukhi
ड U+0921	ਡ U+0A21	Gurmukhi
ड U+0921	ਤ U+0A24	Gurmukhi
ढ U+0922	ਢ U+0A22	Gurmukhi
त U+0924	ਜ U+0A1C	Gurmukhi
य U+092F	ਯ U+0A27	Gurmukhi
ँ U+0945	ঁ U+0981	Bengali

Table 22: Devanagari Cross-script confusables

¹⁷ The Bengali and Devanagari Visarga pair was discussed at length for inclusion in normative part of the Devanagari and Bengali LGRs. It was decided that both look different enough not to be included in the normative part and hence have been added in the Appendix as confusables only.