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

LGR Version: 3.0
Date: 2018-08-08
Document version: 2.2
Authors: Neo-Brahmi Generation Panel [NBGP]

1. General Information/ Overview/ Abstract

The purpose of this document is to give an overview of the proposed Kannada LGR in the XML format and the rationale behind the design decisions taken. It includes a discussion of relevant features of the script, the communities or languages using it, the process and methodology used and information on the contributors. The formal specification of the LGR can be found in the accompanying XML document:

Proposal-LGR-knda_20180808.xml

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

Kannada-test-Labels-20180808.txt

2. Script for which the LGR is Proposed

ISO 15924 Code: Knda
ISO 15924 N°: 345
ISO 15924 English Name: Kannada
Latin transliteration of the native script name:
Native name of the script: ಕನ್ನಡ

Maximal Starting Repertoire (MSR) version: MSR-3

Some languages using the script and their ISO 639-3 codes: Kannada (kan), Tulu (tcy), Beary, Konkani (kok), Havyaka, Kodava (kfa)

3. Background on Script and Principal Languages Using It

3.1 Kannada language

Kannada is one of the scheduled languages of India. It is spoken predominantly by the people of Karnataka State of India. It is one of the major languages among the Dravidian languages. Kannada is also spoken by significant linguistic minorities in the states of Andhra Pradesh, Telangana, Tamil Nadu, Maharashtra, Kerala, Goa and abroad. As per scholars, Kannada was a spoken language during the 3rd century B.C. Ptolemy, a scholar from Alexandria, in his The Geography written during the first half of the second century A.D. mentions some Kannada words. Ptolemy speaks of many places in Karnataka such as Kalgeris (identified as Kalkeri), Modogoulla (Mudugal), Badamios (Badami) and so on. All these are not only places in Karnataka, but are also names of Kannada origin.

The famous Halmidi Record of the Kadambas which is an inscription of the 5th century A.D., is the oldest available evidence of Kannada language written in the pre-Old Kannada script. Kappe Arabhatta's Record at Badami (700 A.D.) has the first Kannada poem in ತ್ರಿಪದಿ tripadi metre. The oldest available literary work in Kannada is ಕವಿರಾಜಮಾರ್ಗ – Kavirajamarga, a book on poetics belonging to 9th century. This work speaks of some earlier poets in Kannada. Hence, Kannada must have been a fully developed language by the 5th or the 6th century A.D. and must have been a spoken language for at least a few centuries earlier. Kannada is attested epigraphically for about one and a half millennia, and literary Old Kannada flourished in the 6th-century Ganga dynasty and during the 9th-century Rashtrakuta Dynasty. Kannada has an unbroken literary history of over a thousand years.

3.2 Evolution of Kannada script

The Kannada language is written using the Kannada script, which evolved from the 5th-century Kadamba script. The oldest form of Kannada script begins in 3rd century B.C. The first popular and well-known Kannada script was called Kadamba script used by the Kadamba dynasty during 5th century A.D. Buhler, the famous epigraphist says that the Kadamba script is the earliest form of the present day Kannada script. During Ganga dynasty, in the 6th century A.D., the script used is known as Adi Ganga script, which resembles Kadamba script. During 6-7th century A.D., the Chalukyas of Badami used a script which is now called by historian as Badami Chalukya script. Rashtrakuta was the next famous dynasty which ruled during 8-10th century A.D. and the script used during those time is referred to as Rashtrakuta script. The script used by the Kalyana Chalukya rulers is called Kalyana Chalukya script. It can be seen in the records of 10-12 century AD. Cursive writing was started during the 13th century by Hoysala kings. They built the decorative cursive way of writing based on the script of Kalyana Chalukyas. Inscriptions at Beluru and Halebeedu have text written using this kind of script. The Vijaynagar kings ruled during the 14-16th century A.D. did not make any major modifications to the script. The last dynasty of Karnataka, the kings of Mysore developed what is known as Modi script. It is called Modi script or ಮೋಡಿ ಬರಹ (Modi baraha). Most of the public records that were written during the period of the Mysore kings are in the Modi script. No inscriptions were written in the Modi script as this style is difficult to inscribe on a stone. This may be considered the latest developed form of the script, and is taught even now in schools as cursive writing for Kannada.

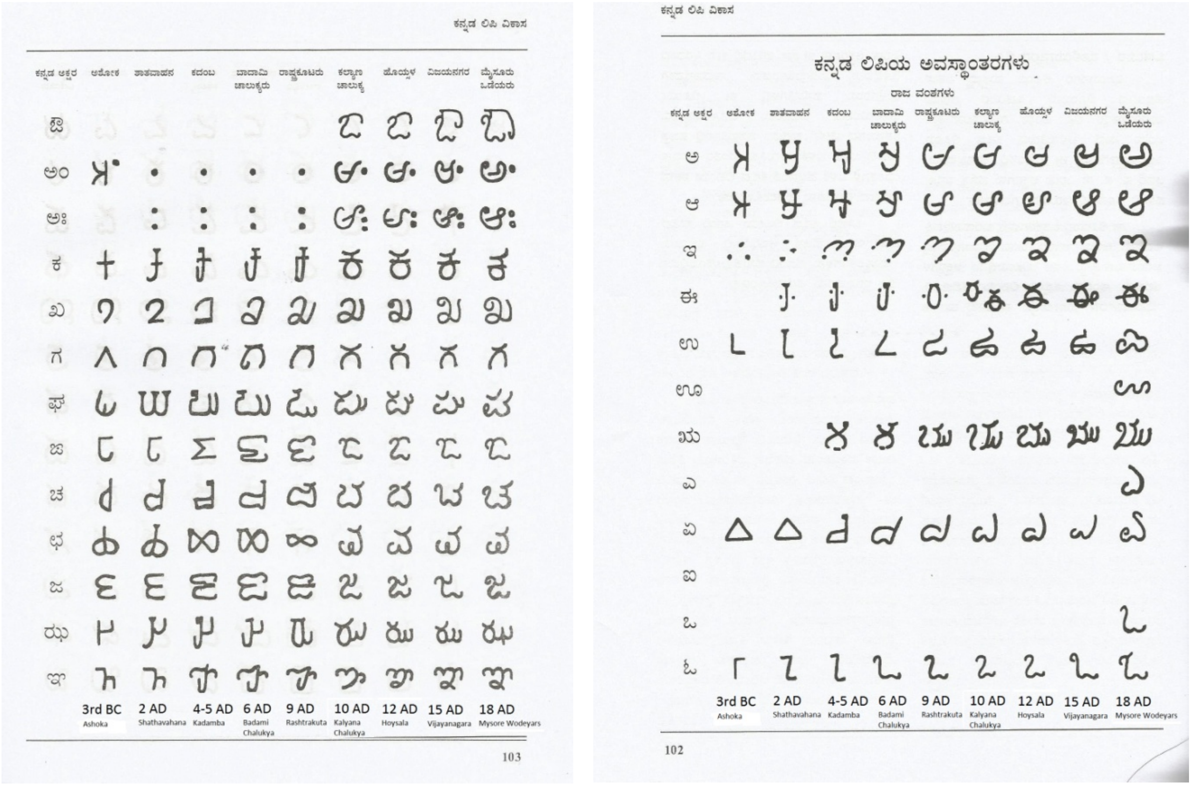

ಕನ್ನಡ ಲಿಪಿ ವಿಕಾಸ
3rd BC Ashoka
2 AD Shathavahana
4-5 AD Kadamba
6 AD Badami Chalukya
9 AD Rashtrakuta
10 AD Kalyana Chalukya
12 AD Hoysala
15 AD Vijayanagara
18 AD Mysore Wodeyars
103
ಕನ್ನಡ ಲಿಪಿಯ ಅವಸ್ಥಾಂತರಗಳು
102

Figure 1: Evolution of Kannada script from 3rd century B.C. to 18th century A.D.
(from https://karnatakaitihasaacademy.org/karnataka-history/evolution-of-kannada-script/)

3.3 Languages considered

Apart from the Kannada language, other languages that use the Kannada script are -Tulu, Kodava (Coorgi), Konkani, Havyaka, Sanketi, Beary (byaari), Arebaase, Koraga, etc. Tulu had its own script which is not in much use nowadays even though lot of efforts are being done of late to revive the Tulu script. The Konkani language is written in Devanagari, Roman, and Malayalam scripts also.

3.4 Structure of written Kannada

The structure of Kannada is similar to other Indian languages, especially to Telugu. The heart of the writing system is the Akshar. The Kannada alphabet is known as aksharamale or varnamale. The modern alphabet contains 49 characters. This has been arrived at by removing two characters that are mainly used to write classical Kannada texts. These two characters were in use just about 50 years ago. Characters combine to form compound characters called as samyuktakshara (conjuncts). These compound characters have distinct display forms. The total number of such combinations will be about 650,000. The basic characters in varnamale are classified into three main categories. They are - swara (vowels), vyanjana (consonants) and yogavahas.

3.4.1 Swaras (vowels)

There are thirteen vowels

Letter	Diacritic	ISO notation
ಅ	N/A	a
ಆ	◌ಾ	ā
ಇ	◌ಿ	i
ಈ	◌ೀ	ī
ಉ	◌ು	u
ಊ	◌ೂ	ū
ಋ	◌ೃ	rū
ಎ	◌ೆ	e
ಏ	◌ೇ	ē
ಐ	◌ೈ	ai
ಒ	◌ೊ	o
ಓ	◌ೋ	ō
ಔ	◌ೌ	au

Table 1: Kannada Swaras (vowels)

(from https://en.wikipedia.org/wiki/Kannada_alphabet)

When a vowel follows a consonant, it is written with a diacritic rather than as a separate letter. Sometimes these are referred to as vowel signs or *matras*. Vowel signs or matras are attached only to consonants.

3.4.2 Yogavahas

The Yōgavāha (part-vowel, part consonant) include two letters:

1. The anusvara: ಅಂ (*aṁ*)
2. The visarga: ಅಃ (*aḥ*)

3.4.3. Vyanjanas (consonants)

Two categories of consonant characters (Vyañjanas) are defined in Kannada: the structured consonants (Vargīya Vyañjana) and the unstructured consonants (Avargīya Vyañjana.

The structured consonants are classified according to where the tongue touches the palate of the mouth and are classified accordingly into five structured groups. These consonants are shown here

	voiceless	voiceless aspirate	voiced	voiced aspirate	nasal
Velars	ಕ (ka)	ಖ (kha)	ಗ (ga)	ಘ (gha)	ಙ (ṅa)
Palatals	ಚ (ca)	ಛ (cha)	ಜ (ja)	ಝ (jha)	ಞ (ña)
Retroflex	ಟ (ṭa)	ಠ (ṭha)	ಡ (ḍa)	ಢ (ḍha)	ಣ (ṇa)
Dentals	ತ (ta)	ಥ (tha)	ದ (da)	ಧ (dha)	ನ (na)
Labials	ಪ (pa)	ಫ (pha)	ಬ (ba)	ಭ (bha)	ಮ (ma)

Table 2: Kannada Consonants

(from https://en.wikipedia.org/wiki/Kannada_alphabet)

The unstructured consonants are consonants that do not fall into any of the above structures: ಯ (ya), ರ (ra), ಱ (ṟa) (obsolete), ಲ (la), ವ (va), ಶ (śa), ಷ (ṣa), ಸ (sa), ಹ (ha), ಳ (ḷa), ೞ (ḻ) (obsolete). From this list the two obsolete characters (ಱ and ೞ) have been removed in modern varnamale bringing the total number of characters to 49.

3.4.4 Implicit vowel ಅ (a) in consonants

All consonants (vyanjanas) in Kannada when written as ಕ (ka), ಖ (kha), ಗ (ga), etc. contain an implicit vowel ಅ (a). The consonants ಕ್, ಖ್, ಗ್, etc., are shown after removing the implicit vowel ಅ (a). In fact many grammar books on Kannada list the consonants by removing the implicit ಅ (a). The Unicode character U+0CCD, which is the Kannada equivalent of Devanagari's Halant U+094D (or VIRAMA as Unicode calls it), follows consonants to remove the implicit ಅ (a). Halant can only follow a consonant and no other characters. A vowel sign (matra) following the consonant replaces the implicit vowel by a different vowel.

3.4.5 Conjuncts

Kannada is known to have a large number of conjuncts which are nothing but combination of consonants and vowel signs (matras). These are also known as syllables. Different types of consonant and vowel sign combinations are possible. They are the following:

- Consonant + Vowel sign,
  e.g., ಕ (ka, U+0C95) + ◌ೊ (U+0CCA, matra of vowel ಒ) = ಕೊ
- Consonant + Halant + Consonant,
  e.g., ಕ (ka, U+0C95) + Halant (U+0CCD) + ಕ (ka, U+0C95) = ಕ್ಕ
- Consonant + Halant + Consonant + Vowel sign,
  e.g. ಕ (ka, U+0C95) + Halant (U+0CCD) + ಕ (ka, U+0C95) + ◌ೊ (U+0CCA, matra of vowel ಒ) = ಕ್ಕೊ
- Consonant + Halant + Consonant + Halant + Consonant
  e.g., ಷ (ṣa, U+0CB7) + Halant (U+0CCD) +ಟ (ṭa, U+0C9F) + Halant (U+0CCD) + ರ (ra, U+0CB0) = ಷ್ಟ್ರ
- Consonant + Halant + Consonant + Halant + Consonant + Vowel sign
  e.g., ಷ (ṣa, U+0CB7) + Halant (U+0CCD) + ಟ (ṭa, U+0C9F) + Halant (U+0CCD) + ರ (ra, U+0CB0) + ◌ೊ (U+0CCA, matra of vowel ಒ) = ಷ್ಟ್ರೊ

Conjuncts cluster having more than 3 consonants in one syllable are normally not seen in Kannada.

3.4.6 Pure vowels in the middle of a word

In Kannada, it is common to have words starting with a vowel. Sometimes, of late, people are writing words having pure vowels in the middle of a word. This kind of writing was originally normally not seen in Kannada. This kind of writing has been made a requirement to write the words imported from other languages, especially from English. Linguistically this is not invalid and hence can be allowed.

3.4.7 Illegal combinations

There are some combinations which are invalid as per Kannada grammar. They are listed below:

3.4.7.1 Having two or more consecutive vowel signs (matras).
3.4.7.2 Having a vowel sign (matra) after a vowel.
3.4.7.3 Having a vowel sign (matra) after a Yōgavāha (anusvara or visarga).
3.4.7.4 Having a Halant after a vowel or vowel sign (matra).
3.4.7.5 Having a Yōgavāha after a Halant.

For 3.4.7.4 there could be cases involving multi-word domains where V may need to be allowed to follow an H. This is the case where two different words are joined together first of which ends with a Halant and the second word begins with a Vowel. Some sections of the linguistic community require the explicit presence of H for full representation of the sound intended. However, by and large, the form of the first word without a H is considered enough for full representation of the sound intended for the first word.

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

Depending on the prevailing requirements by the community, afuture NBGP may consider revisiting this rule.

4. Overall Development Process and Methodology

Neo-Brahmi Generation Panel (NBGP) has been formed by members having experience in linguistics and computational linguistics. Under the Neo-Brahmi Generation Panel, there are nine scripts belonging to separate Unicode blocks. Each of these scripts has been a separate LGR; however, the Neo-Brahmi GP ensures that the fundamental philosophy behind building those LGRs are all in sync with all other Brahmi-derived scripts.

NBGP considered all the languages with EGIDS scale 1 to 4 and found that Kannada script is being used for Kannada, Tulu, Beary, Konkani, Havyaka, Kodava, among other languages.

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 of 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 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. 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: Kannada sign CANDRABINDU ◌ಁ (U+0C81) is not allowed to be part of the domain name.

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: Kannada Sign AVAGRAHA "ಽ" (U+0CBD) 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 by admitting only such characters as are part of the code-block of the given script/language. This is further narrowed down by the IDNA 2008 Protocol and finally an additional filter in the form of Maximal Starting Repertoire restricts the character set associated with the given language even more.

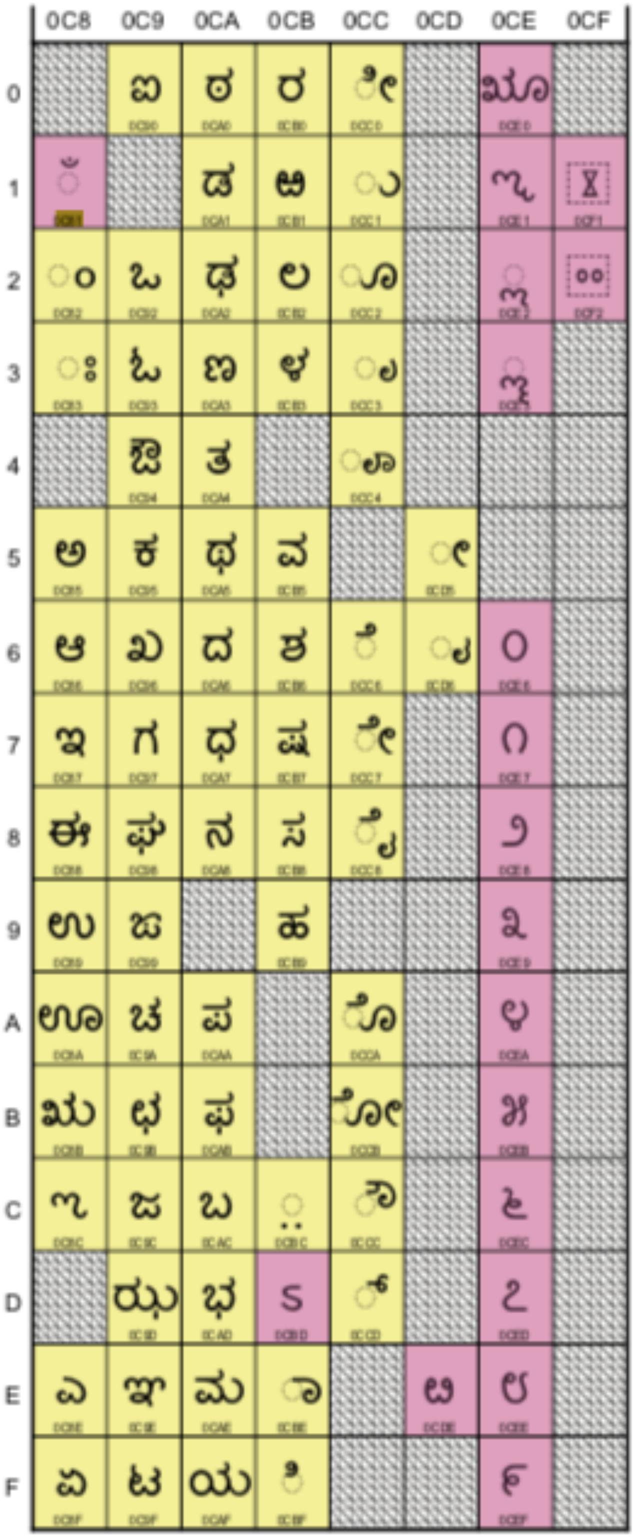
4.1.2.2 No Rare and Obsolete Characters:

There are characters which have been added to Unicode to accommodate rare forms especially like KANNADA LETTER VOCALIC L "ಌ" (U+0C8C) as well as its matra forms "◌ೢ" (U+0CE2). All such characters will not be included. This is in consonance with the Conservatism principle as laid down in the Root Zone LGR procedure.

5. Repertoire

Section 5.1 provides the section of the [MSR] applicable to the Kannada script on which the Kannada 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 Kannada LGR.

5.1 Kannada section of Maximal Starting Repertoire [MSR] Version 3



	0C8	0C9	0CA	0CB	0CC	0CD	0CE	0CF
0		ಐ	ಠ	ರ	◌ೀ		ೠ	
1	◌ಁ		ಡ	ಱ	◌ು		ೡ	ೱ
2	◌ಂ	ಒ	ಢ	ಲ	◌ೂ		◌ೢ	ೲ
3	◌ಃ	ಓ	ಣ	ಳ	◌ೃ		◌ೣ	
4		ಔ	ತ		◌ೄ			
5	ಅ	ಕ	ಥ	ವ		◌ೕ		
6	ಆ	ಖ	ದ	ಶ	◌ೆ	◌ೖ	೦	
7	ಇ	ಗ	ಧ	ಷ	◌ೇ		೧	
8	ಈ	ಘ	ನ	ಸ	◌ೈ		೨	
9	ಉ	ಙ		ಹ			೩	
A	ಊ	ಚ	ಪ		◌ೊ		೪	
B	ಋ	ಛ	ಫ		◌ೋ		೫	
C	ಌ	ಜ	ಬ	◌಼	◌ೌ		೬	
D		ಝ	ಭ	ಽ	◌್		೭	
E	ಎ	ಞ	ಮ	◌ಾ		ೞ	೮	
F	ಏ	ಟ	ಯ	◌ಿ			೯	

Figure 2: Kannada Code Page from [MSR]

Color convention[1]**:**

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

PVALID in IDNA2008 but excluded from the [MSR] - Pinkish background

Not PVALID in IDNA2008 - White background

[1] This document needs to be printed in color for this to be read correctly.

5.2 Code point repertoire

Given below is the repertoire for Kannada based on Unicode character set.

Sr. No.	Unicode Code Point	Glyph	Character Name	Indic Syllabic Category	Reference
1	0C82	ಂ	KANNADA SIGN ANUSVARA	Anusvara	110
2	0C83	ಃ	KANNADA SIGN VISARGA	Visarga	110
3	0C85	ಅ	KANNADA LETTER A	Vowel	110
4	0C86	ಆ	KANNADA LETTER AA	Vowel	110
5	0C87	ಇ	KANNADA LETTER I	Vowel	110
6	0C88	ಈ	KANNADA LETTER II	Vowel	110
7	0C89	ಉ	KANNADA LETTER U	Vowel	110
8	0C8A	ಊ	KANNADA LETTER UU	Vowel	110
9	0C8B	ಋ	KANNADA LETTER VOCALIC R	Vowel	110
10	0C8E	ಎ	KANNADA LETTER E	Vowel	110
11	0C8F	ಏ	KANNADA LETTER EE	Vowel	110
12	0C90	ಐ	KANNADA LETTER AI	Vowel	110
13	0C92	ಒ	KANNADA LETTER O	Vowel	110
14	0C93	ಓ	KANNADA LETTER OO	Vowel	110
15	0C94	ಔ	KANNADA LETTER AU	Vowel	110
16	0C95	ಕ	KANNADA LETTER KA	Consonant	110
17	0C96	ಖ	KANNADA LETTER KHA	Consonant	110
18	0C97	ಗ	KANNADA LETTER GA	Consonant	110

Sr. No.	Unicode Code Point	Glyph	Character Name	Indic Syllabic Category	Reference
19	0C98	ಘ	KANNADA LETTER GHA	Consonant	110
20	0C99	ಙ	KANNADA LETTER NGA	Consonant	110
21	0C9A	ಚ	KANNADA LETTER CA	Consonant	110
22	0C9B	ಛ	KANNADA LETTER CHA	Consonant	110
23	0C9C	ಜ	KANNADA LETTER JA	Consonant	110
24	0C9D	ಝ	KANNADA LETTER JHA	Consonant	110
25	0C9E	ಞ	KANNADA LETTER NYA	Consonant	110
26	0C9F	ಟ	KANNADA LETTER TTA	Consonant	110
27	0CA0	ಠ	KANNADA LETTER TTHA	Consonant	110
28	0CA1	ಡ	KANNADA LETTER DDA	Consonant	110
29	0CA2	ಢ	KANNADA LETTER DDHA	Consonant	110
30	0CA3	ಣ	KANNADA LETTER NNA	Consonant	110
31	0CA4	ತ	KANNADA LETTER TA	Consonant	110
32	0CA5	ಥ	KANNADA LETTER THA	Consonant	110
33	0CA6	ದ	KANNADA LETTER DA	Consonant	110
34	0CA7	ಧ	KANNADA LETTER DHA	Consonant	110
35	0CA8	ನ	KANNADA LETTER NA	Consonant	110
36	0CAA	ಪ	KANNADA LETTER PA	Consonant	110
37	0CAB	ಫ	KANNADA LETTER PHA	Consonant	110

Sr. No.	Unicode Code Point	Glyph	Character Name	Indic Syllabic Category	Reference
38	0CAC	ಬ	KANNADA LETTER BA	Consonant	110
39	0CAD	ಭ	KANNADA LETTER BHA	Consonant	110
40	0CAE	ಮ	KANNADA LETTER MA	Consonant	110
41	0CAF	ಯ	KANNADA LETTER YA	Consonant	110
42	0CB0	ರ	KANNADA LETTER RA	Consonant	110
43	0CB2	ಲ	KANNADA LETTER LA	Consonant	110
44	0CB3	ಳ	KANNADA LETTER LLA	Consonant	110
45	0CB5	ವ	KANNADA LETTER VA	Consonant	110
46	0CB6	ಶ	KANNADA LETTER SHA	Consonant	110
47	0CB7	ಷ	KANNADA LETTER SSA	Consonant	110
48	0CB8	ಸ	KANNADA LETTER SA	Consonant	110
49	0CB9	ಹ	KANNADA LETTER HA	Consonant	110
50	0CBE	◌ಾ	KANNADA VOWEL SIGN AA	Matra	110
51	0CBF	◌ಿ	KANNADA VOWEL SIGN I	Matra	110
52	0CC0	◌ೀ	KANNADA VOWEL SIGN II	Matra	110
53	0CC1	◌ು	KANNADA VOWEL SIGN U	Matra	110
54	0CC2	◌ೂ	KANNADA VOWEL SIGN UU	Matra	110
55	0CC3	◌ೃ	KANNADA VOWEL SIGN VOCALIC R	Matra	110

Sr. No.	Unicode Code Point	Glyph	Character Name	Indic Syllabic Category	Reference
56	0CC6	◌ೆ	KANNADA VOWEL SIGN E	Matra	110
57	0CC7	◌ೇ	KANNADA VOWEL SIGN EE	Matra	110
58	0CC8	◌ೈ	KANNADA VOWEL SIGN AI	Matra	110
59	0CCA	◌ೊ	KANNADA VOWEL SIGN O	Matra	110
60	0CCB	◌ೋ	KANNADA VOWEL SIGN OO	Matra	110
61	0CCC	◌ೌ	KANNADA VOWEL SIGN AU	Matra	110
62	0CCD	◌್	KANNADA SIGN VIRAMA	Halant / Virama	110

Table 3: Code point repertoire

5.3 Codepoints not included

Following code points have not been included in the repertoire.

Sr. No.	Unicode Code Point	Glyph	Character Name	Reason for Exclusion
1.	0C8C	ಌ	KANNADA LETTER VOCALIC L	Not used in Kannada
2.	0CB1	ಱ	KANNADA LETTER RRA	Obsolete character, not used in modern Kannada
3.	0CBC	◌಼	KANNADA SIGN NUKTA	Does not belong to Kannada, not needed in LGR
4.	0CC4	◌ೄ	KANNADA VOWEL SIGN VOCALIC RR	Not used in Kannada
5.	0CD5	◌ೕ	KANNADA LENGTH MARK	Not in use
6.	0CD6	◌ೖ	KANNADA AI LENGTH MARK	Not in use

Table 4: Code point not included

6. Variants

6.1 In-script variants

There are no variants within the Kannada script.

6.2 Cross-script variants analysis

Some characters of Kannada look the same as some characters in Devanagari, Gujarati, Telugu, Malayalam and Sinhala. The tables below list them.

6.2.1 Cross-script variants for Kannada and Telugu analysis

The Telugu and Kannada code point sets in Table 5 are cross-script variant code points.

Variant Set	Telugu Code Point	Kannada Code Point
1	◌ం (0C02)	◌ಂ (0C82)
2	◌ః (0C03)	◌ಃ (0C83)
3	అ (0C05)	ಅ (0C85)
4	ఆ (0C06)	ಆ (0C86)
5	ఇ (0C07)	ಇ (0C87)
6	ఈ (0C08)	ಈ (0C88)
7	ఐ (0C10)	ಐ (0C90)
8	ఒ (0C12)	ಒ (0C92)
9	ఓ (0C13)	ಓ (0C93)
10	ఔ (0C14)	ಔ (0C94)
11	ఖ (0C16)	ಖ (0C96)
12	గ (0C17)	ಗ (0C97)
13	జ (0C1C)	ಜ (0C9C)
14	ఝ (0C1D)	ಝ (0C9D)
15	ఞ (0C1E)	ಞ (0C9E)

16	ట (0C1F)	ಟ (0C9F)
17	ఠ (0C20)	ಠ (0CA0)
18	డ (0C21)	ಡ (0CA1)
19	ఢ (0C22)	ಢ (0CA2)
20	ణ (0C23)	ಣ (0CA3)
21	థ (0C25)	ಥ (0CA5)
22	ద (0C26)	ದ (0CA6)
23	ధ (0C27)	ಧ (0CA7)
24	న (0C28)	ನ (0CA8)
25	బ (0C2C)	ಬ (0CAC)
26	భ (0C2D)	ಭ (0CAD)
27	మ (0C2E)	ಮ (0CAE)
28	య (0C2F)	ಯ (0CAF)
29	ర (0C30)	ರ (0CB0)
30	ల (0C32)	ಲ (0CB2)
31	ళ (0C33)	ಳ (0CB3)
32	◌ి (0C3F)	◌ಿ (0CBF)
33	◌ు (0C41)	◌ು (0CC1)
34	◌ృ (0C43)	◌ೃ (0CC3)

Table 5: Telugu and Kannada code point analysis

6.2.2 Cross-script variants for Kannada and Devanagari analysis

Visarga is the only potential variant code point that exhibits shape similarity between the Kannada and Devanagari scripts. However, as this is a combining mark and there are no other variant code points between the two languages, it is not defined as a variant code point.

Devanagari Code Point	Kannada Code Point
◌ः (0903)	◌ಃ (0C83)

Table 6: Devanagari and Kannada code point analysis

6.2.3 Cross-script variants for Kannada and Gujarati analysis

Visarga is the only potential variant code point that exhibits shape similarity between the Kannada and Gujarati scripts. However, as this is a combining mark and there are no other variant code points between the two languages, it is not defined as a variant code point.

Gujarati Code Point	Kannada Code Point
◌ઃ (0A83)	◌ಃ (0C83)

Table 7: Gujarati and Kannada code point analysis

6.2.4 Cross-script variants for Kannada and Malayalam analysis

Anusvara and Visarga are the only potential variant code points that exhibit shape similarity between the Kannada and Malayalam scripts. However, as they are combining marks and there are no other variant code points between the two languages, it is not defined as a variant code point.

Kannada Code Point	Malayalam Code Point
◌ಂ (0C82)	◌ം (0D02)
◌ಃ (0C83)	◌ഃ (0D03)

Table 8: Kannada and Malayalam code point analysis

6.2.5 Cross-script variants for Kannada and Sinhala analysis

These pairs are defined as variant code points with the Sinhala script as they are identical and there are enough variant code points for them to form multiple labels. This analysis follows the NBGP Cross-script Variant inclusion policy available in Appendix III.

Variant Set	Kannada Code Point	Sinhala Code Point
1	◌ಂ (0C82)	ං (0D82)
2	◌ಃ (0C83)	ඃ (0D83)
3	ರ (0CB0)	ර (0DBB)

Table 9: Kannada and Sinhala code point analysis

7. Whole Label Evaluation Rules (WLE)

The structure of written Kannada is provided in section 3.4. using the following variables or definitions:

V	→	Vowel
M	→	Matra
C	→	Consonant
H	→	Halant / Virama
B	→	Anusvara
X	→	Visarga

Rules for forming akshar or a cluster of one unit are given below:

1. Rule 1: H must be preceded by C
2. Rule 2: M must be preceded by C
3. Rule 3: B must be preceded by C, V or M
4. Rule 4: X must be preceded by C, V or M
5. Rule 5: V cannot be preceded by H

8. Contributors

1. Dr. U.B. Pavanaja, pavanaja@vishvakannada.com
2. Dhanalakshmi K.T., lakshmikt96@gmail.com
3. Tejas Jain, teju2friends@gmail.com
4. NBGP Members

9. References

[MSR] Integration Panel, "Maximal Starting Repertoire — MSR-3 Overview and Rationale", 28 March2018
https://www.icann.org/sites/default/files/packages/lgr/msr/msr-3-wle-rules-28mar18-en.html
[NBGP] Neo-Brahmi Generation Panel
101. A Comparative Grammar of the Dravidian or South Indian Family of Languages by Robert Caldwell, Trubner & Co, London, Second Edition, 1875
102. Evolution of Kannada script - https://karnatakaitihasaacademy.org/karnataka-history/evolution-of-kannada-script/
103. History of the Kannada Script and Language - https://bookstalkist.com/history-of-the-kannada-script-and-language/
104. History of the Kannada Literature -
http://kamat.com/kalranga/kar/literature/history1.htm
105. Kannada alphabet - https://en.wikipedia.org/wiki/Kannada_alphabet

106. About Kannada language - https://en.wikipedia.org/wiki/Kannada
107. Ethnologue entry about Kannada - http://www.ethnologue.com/19/language/kan/
108. OLAC resources in and about the Kannada language - http://www.language-archives.org/language/kan
109. Encyclopaedia Britannica entry about Kannada - https://www.britannica.com/topic/Kannada-language
110. ಕನ್ನಡ ಮಧ್ಯಮ ವ್ಯಾಕರಣ, ತೀ.ನಂ. ಶ್ರೀಕಂಠಯ್ಯ, ಗೀತಾ ಬುಕ್ ಹೌಸ್, ಮೈಸೂರು, ೨೦೦೧

(*Kannada Madhyama Vyakarana* (means An Intermediate Kannada Grammar), T. N. Sreekantaiya, Geetha Book House, Mysore, 2001.)

Appendix I : Confusable Kannada and Telugu Analysis

Some code points for Kannada and Telugu script were discussed and the NBGP concluded that these is not confusable code points. Table below lists all discussed code points and their resolution.

Ser. No.	Kannada Character	Unicode codepoint	Telugu character	Unicode codepoint	NBGP Resolution
1	ಎ	0C8E	ఎ	0C0E	Distinguishable
2	ಘ	0C98	ఘ	0C18	Distinguishable
3	ಙ	0C99	ఙ	0C19	Distinguishable
4	ಚ	0C9A	చ	0C1A	Distinguishable
5	ಛ	0C9B	ఛ	0C1B	Distinguishable
6	ಪ	0CAA	ప	0C2A	Distinguishable
7	ಫ	0CAB	ఫ	0C2B	Distinguishable
8	ವ	0CB5	వ	0C35	Similar
9	ಶ	0CB6	శ	0C36	Similar
110	ಷ	0CB7	ష	0C37	Distinguishable
11	ಸ	0CB8	స	0C38	Similar
12	◌ೌ	0CCC	◌ౌ	0C4C	Distinguishable

Table 10: NBGP resolution of Telugu and Kannada code points

Appendix II : Zero Width Joiner and Zero Width Non Joiner in Kannada domain names

Zero Width Joiner (ZWJ) and Zero Width Non Joiner (ZWNJ) have special roles in Kannada. ZWNJ is used in sequences like Consonant (C) + Halant (U+0CCD) + Consonant kind where the second C should not take the vattu form after (below) the first consonant.

Example:
1.ಜ (U+0C9C) + ◌್ + (U+0CCD) + ಕ (U+0C95) = ಜ್ಕ – without using ZWNJ
2. ಜ (U+0C9C) + ◌್ + (U+0CCD) + ZWNJ (U+200C) + ಕ (U+0C95) = ಜ್‌ಕ – using ZWNJ

Example of the word not using ZWNJ – ರಾಜ್ಕುಮಾರ್ (Rajkumar)
Example of word using ZWNJ – ರಾಜ್‌ಕುಮಾರ್ (Rajkumar)

Both words mean the same. But their pronunciations are slightly different. The second form does not originally belong to Kannada, but nowadays due to the influence of English and Hindi, some people are using the form. But this form is needed to write many English words in Kannada like software (ಸಾಫ್ಟ್‌ವೇರ್, using ZWNJ). The word software will become ಸಾಫ್ಟ್ವೇರ್ if ZWNJ is not used.

ZWJ has a unique purpose in Kannada where conjuncts are formed with consonants where the first consonant is ರ (Ra) (U+0CB0). When conjunct is formed with Ra + Halant + <consonant>, the default form is second <consonant> followed by arkavattu, which is Kannada equivalent of the reph of Devanagari.

Example: ರ (U+0CB0) + ◌್ (U+0CCD) + ಕ (U+0C95) = ರ್ಕ

Original Kannada did not have this arkavattu or reph. It was like any other C1+ Halant + C2 where the C2 takes the vattu form, even when C1 = ರ (Ra). To get the vattu form for ರ + Halant + <consonant>, ZWJ is used.

Example:
ರ (U+0CB0) + ◌್ (U+0CCD) + ZWJ (U+200D) + ಕ (U+0C95) = ರ್‍ಕ –Microsoft's implementation
ರ (U+0CB0) + ZWJ (U+200D) + ◌್ (U+0CCD) + ಕ (U+0C95) = ರ‍್ಕ – as per Unicode's definition. Browsers Firefox and Chrome use this rendering rule.

The original Kannada form (not using arkavattu or reph) is needed when the first letter of the word itself is ರ (Ra, U+0CB0) followed by Halant and a consonant.
Example- ರ (U+0CB0) + Halant (U+0CCD) + ಯ (U+0CAF) = ರ್ಯ as in ರ್ಯಾಲಿ (rally written in Kannada). This is wrong as per the writing style followed in Kannada. Here the

second consonant ಯ (Ya, U+0CAF) must take the vattu form. To get that one must use ZWJ as explained above. Correct form of rally written in Kannada will be ರ್‍ಯಾಲಿ making use of ZWJ. One such word I can think of will be the company Rallis India. When they go for registering domain name in Kannada, if ZWJ is not allowed in domain name, they will have to go for ರ್ಯಾಲೀಸ್.ಭಾರತ, which is the wrong form in Kannada. They should write it as ರ್‍ಯಾಲೀಸ್.ಭಾರತ by making use of ZWJ.

Thus both ZWJ and ZWNJ are needed for having proper domain names in Kannada.

How to avoid duplicate domain names involving ZWJ and ZWNJ?

ZWJ and ZWNJ are used mainly to have two display forms of what is linguistically same word or combination of characters in Kannada. When ZWJ and ZWNJs are allowed in domain names for Kannada, it will create two domain names which have two display forms but linguistically they are same. To make the browsers and DNSs treat them as equal, we have to ignore ZWJ and ZWNJs for comparing two words. This methodology is followed by the spell-check used in Microsoft Word. The same philosophy can be applied here also.

Accepting ZWJ and ZWNJ in domain names creates confusion to a majority of the linguistic community and joiner characters are prohibited for the Root Zone, hence this is explicitly prohibited by the NBGP.

Appendix III : NBGP Cross-script Variant Inclusion Policy

If, in any two given scripts, all the potential cross-script variants consist of dependent (e.g. Vowel Signs, Anusvara, Visarga, Chandrabindu etc.) characters **ONLY**, then that entire set can be ignored and no cross-script variants be proposed between those two scripts.

If, in any two given scripts, there is **AT LEAST ONE** non-dependent (e.g. Consonant, Vowel etc.) cross-script variant character/sequence present, all the potential cross-script variants be considered and proposed between the two scripts.

This cross-script analysis has been restricted to the scripts that have descended from the Brahmi as most of them share similar usage patterns. By and large, all of these scripts have a common set of characters that existed in Brahmi script and bear the same identities. However, as the scripts branched out from the Brahmi, depending on various factors, the shapes of the characters changed. This change in the shape was not uniform across all the characters and the scripts. Some characters shapes did change significantly whereas some of them still retained similarity. The cross-script similarity analysis also aims to identify such cases where the same character retained almost the same shape despite being part of the different scripts. These set of characters are variants of each other in true sense than merely of co-incidental visual similarity.

Since, having such labels is a realistic possibility and the corresponding labels look almost exactly alike, NBGP has proposed them as blocked variants.

NBGP acknowledges the concern that this shape is quite generic and may have parallels in other scripts not under its ambit. However, as NBGP does not have any exposure about actual usage of those characters in those particular scripts, NBGP desisted from including them in the analysis. As NBGP has already considered all the related scripts under the cross-script variant analysis, the similarity of the characters belonging to NBGP scripts with other scripts not under the NBGP ambit, may be of a mere co-incidental visual nature.

Additionally, this concern is not limited to these two characters but for all the characters in all the scripts under the scope of the Root LGR procedure. Carrying out this analysis can practically be done only with the Generation Panels that exist while the NBGP is active. This still leaves out those scripts out of the scope which may not have a Generation Panel established yet. Hence, carrying out this exercise in entirety is quite impracticable. This conundrum can be resolved if all the such cases are handled by the "String Similarity Assessment Panel" of ICANN.