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

LGR Version: 3.0
Date: 2019-06-07
Document version: 2.8
Authors: Neo-Brahmi Generation Panel [NBGP]

1. General Information/ Overview/ Abstract

This document lays down the Label Generation Rule Set for the Telugu script. Three main components of the Telugu Script LGR, viz. 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: "proposal-telugu-lgr-07jun19-en.xml".

In addition, a list of test labels has been provided in the following file, which covers the repertoire, variant code points and the whole label evaluation rules, providing examples for valid and invalid labels: “telugu-test-labels-07jun19-en.txt”.

2. Script for which the LGR is proposed

ISO 15924 Code: Telu
ISO 15924 Key N°: 340
ISO 15924 English Name: Telugu
Latin transliteration of native script name: telugu
Native name of the script: తెలుగు
Maximal Starting Repertoire [MSR] version: 4
The Unicode Standard, Version: 6.3
Telugu Unicode Range: 0C00–0C7F

3. Background of the Script and Principal Languages Using It

The Telugu language uses the Telugu script which is written in the form of sequences of orthographic syllables. Each orthographic syllable is formed of one or more Telugu characters placed from left to right and top to bottom. Telugu is one of the 22 scheduled languages of India. The Telugu script is immediately related to Kannada and closely related to the Sinhala script.

3.1 The Evolution of the Script

The origins of the Telugu script can be traced to the Brahmi alphabet of ancient India, often known as Asokan Brahmi. Historically the script is derived from the Southern Brahmi or Bhattiprolu Brahmi alternatively known as the Telugu Brahmi alphabet of 3rd century BCE. Later, by 5th century during the *Chalukyan* period, it developed into a common alphabet used for Telugu and Kannada. The Telugu-Kannada common alphabet split into two separate alphabets during the 12th and 13th centuries AD to be called the Telugu and Kannada scripts. In addition to the common origin, a longer period of shared political and cultural confederation of the Telugu and Kannada speaking regions has ultimately resulted in the considerable proportion of the shared identical character signs between the two scripts (34 out of 63 characters, see Table 10) .

The earliest known inscriptions containing Telugu words appear on the bilingual coins of *Satavahanas* that date back to 2nd century AD [104]. The first inscription entirely in Telugu was made in 575 AD and was probably made by *Renati Cholas*, who started writing royal proclamations in Telugu instead of Sanskrit. Telugu developed as a poetical and literary language during the 11th century AD. Until the 20th century Telugu was written in *Granthic* style very different from the colloquial language. During the second half of the 20th century, a modern written style emerged based on the modern colloquial language. In 2008 Telugu was designated as a classical language by the Indian government.

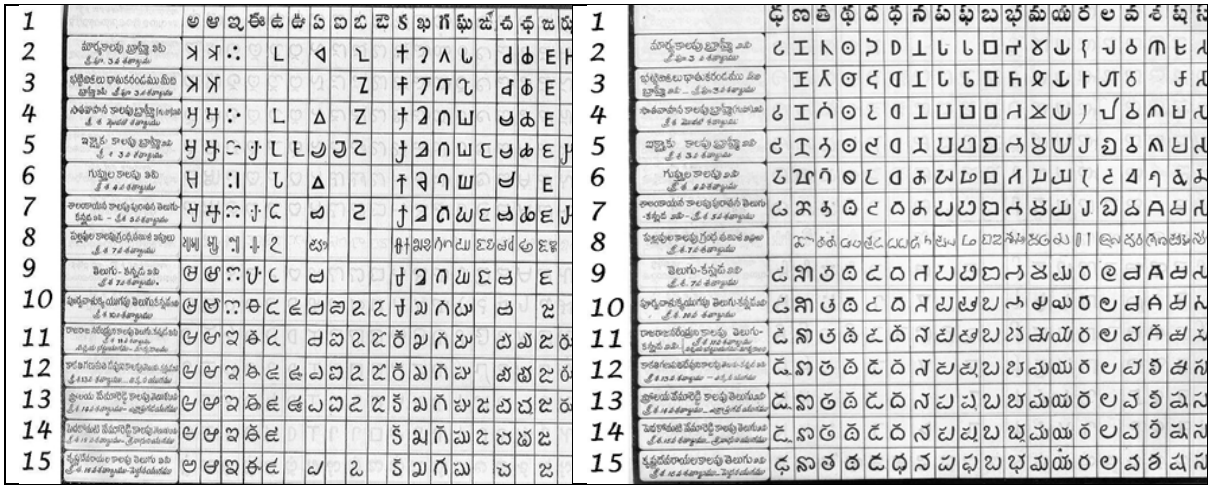

Figure 1: Evolution of Telugu script

3.2 Notable Features

The Telugu orthography superficially appears as a series of circles and semi-circles. Most consonants carry a tick mark called *Talakattu*. The writing system is classified as abugida type that employs alpha-syllabaries. The alphabet consists of vowels, consonants and modifiers. Each of these vowels and consonants has one or more secondary allographs. The secondary allographs always appear as dependent symbols on the first character of a syllable. Each syllable is formed of a single standalone vowel or one or more consonants. Each of these consonants may occur with an inherent vowel or modified by a secondary vowel. A Consonant cluster may be formed with a single standalone character followed

by one or more secondary forms of consonants. The order of composition of syllabaries does not match with the reading order. There are rules to learn to read orthographic sequences into phonetic sequences whether simple or complex syllables.

3.3 The Telugu (తెలుగు) Language

The Telugu language is a Dravidian language spoken by about 75 million (ca. 2001) people mainly in the southern Indian states of Andhra Pradesh and Telangana where it is the official language. It is also spoken in such neighboring states as Karnataka, Tamil Nadu, Orissa, Maharashtra and Chattisgarh, and is one of the 22 scheduled languages of India. There are also quite a few Telugu speakers in Canada, the USA, South Africa, Malaysia, Mauritius, Myanmar, Sri Lanka and Réunion

3.4 Languages that Use the Telugu Script

The script is also used for ten other languages, viz. Gondi, Koya, Konda, Kuvi, Kolavar or Kolami, Yerukala, Banjara or Lambadi, Savara or Sora, Adivasi Odiya and also Sanskrit.
In the Telugu speaking region, the tradition of writing Sanskrit in the Telugu script has remained a common practice. During the last few decades, a considerable number of publications in the form of text books, dictionaries and other reading material has been produced in the Telugu script in Gondi, Koya, Konda, Kuvi, Kolami, Yerukala, Banjara, Savara and Adivasi Odiya.

no.	Name of the language (ISO639 Code)	Language family	Status	EGIDS Scale
1	Telugu (tel)	Dravidian	Scheduled and Classical	2
2	Gondi (gon)	Dravidian	Modern Tribal	5
3	Koya (kff)	Dravidian	Modern Tribal	5
4	Konda (knd)	Dravidian	Modern Tribal	6b
5	Kuvi (kxv)	Dravidian	Modern Tribal	5
6	Kolavar or Kolami (kfb)	Dravidian	Modern Tribal	5
7	Yerukala (yeu)	Dravidian	Modern Tribal	6
8	Banjara or Lambadi (lmn)	Indo-Aryan	Modern Tribal	5
9	Savara or Sora (srb)	Austro-Asiatic	Modern Tribal	5
10	Adivasi Odiya (ort)	Indo-Aryan	Modern Tribal	5

no.	Name of the language (ISO639 Code)	Language family	Status	EGIDS Scale
11	Sanskrit (san)	Indo-Aryan	Scheduled and Classical	4

Table 1: Main languages considered under Telugu LGR

3.5 The Structure of Written Telugu

The Telugu script as it is used for the Telugu language consists of a total of 72 characters [102] comprising 40 consonants, 16 characters representing vowels that can stand alone and 16 dependent signs, each corresponding one of the sixteen vowels excepting /a/ అ; no explicit dependent symbol exists for that sound, instead it is inherent with the consonants in the absence of a dependent sign. Besides these, there are six additional dependent symbols, of which five always occur with the vowels, as extensions. The sixth, the halant sign ్ U+0C4D, occurs with consonants. The following subsections give further details.

3.5.1 The vowels and vowel modifiers

There are fourteen vowel characters viz. అ [a], ఆ [ā], ఇ [i], ఈ [ī], ఉ [u], ఊ [ū], ఋ [r̥], ఌ [l̥], ఎ [e], ఏ [ē], ఐ [ai], ఒ [o], ఓ [ō], ఔ [au], in the common inventory [103] for all the languages using Telugu script [111] specified above and two (ౠ [r̥̄], ౡ [l̥̄]) to write Sanskrit loan words. For these vowels, there are corresponding fifteen marks, except for అ [a] (which is inherent). These are listed in Table 2 below. There are six modifiers for vowels: ఁ [~], ం [ṃ], ః [ḥ], ఀ [~] (a special symbol not common in standard Telugu writings), ౼ [:.] (the *avagraha* sign, commonly used to indicate doubling the vowel length and follows only long vowels), and ్ [H] (the *halant* sign, when appended to a consonant, deducts the inherent vowel /a/ from it). The *halant* sign has similar characteristic as that of a secondary vowel sign in that both of them delete the inherent vowel [a] when added to consonants.

R1. Inherent vowel deletion rule: An inherent vowel of a consonant gets deleted either before a *matra* sign or before the *halant* sign.

No.	Independent vowels primary allographs with code points	Dependent vowels secondary allographs with code points
1.	అ U+0C05	No explicit sign recognized or encoded
2.	ఆ U+0C06	ా U+0C3E
3.	ఇ U+0C07	ి U+0C3F

No.	Independent vowels primary allographs with code points	Dependent vowels secondary allographs with code points
4.	ఈ U+0C08	ీ U+0C40
5.	ఉ U+0C09	ు U+0C41
6.	ఊ U+0C0A	ూ U+0C42
7.	ఋ U+0C0B	ృ U+0C43
8.	ౠ U+0C60	ౄ U+0C44
9.	ఌ U+0C0C	ౢ U+0C62
10.	ౡ U+0C61	ౣ U+0C63
11.	ఎ U+0C0E	ె U+0C46
12.	ఏ U+0C0F	ే U+0C47
13.	ఐ U+0C10	ై U+0C48
14.	ఒ U+0C12	ొ U+0C4A
15.	ఓ U+0C13	ో U+0C4B
16.	ఔ U+0C14	ౌ U+0C4C

Table 2: Vowels and the corresponding dependent signs

No.	Modifier signs	Code Points	Common name
1.	ఀ	U+0C00	Candrabindu
2.	ఁ	U+0C01	Ardhānusvāra or Arasunna
3.	ం	U+0C02	Pūrṇanusvāra or Sunna
4.	ః	U+0C03	Visarga
5.	ఽ	U+0C3D	Avagraha
6.	్	U+0C4D	Halant

Table 3: Vowel modifiers and the consonantal modifiers

3.5.2 The Anusvāra or sunna (ం - U+0C02)

The Anusvāra or *sunna* represents a homorganic nasal before the corresponding consonant and as a substitute to transcribe word final /mu/. Essentially it substitutes a cluster of a Nasal Consonant + Halant before a consonant. Writing alternatively with a nasal consonant + Halant + Consonant is rare and often occurs while transcribing

Sanskrit words. Otherwise the writing practice with nasal consonant + Halant + Consonant of the later type is virtually absent in Telugu.

No.	Homorganic nasal = Archiphoneme /M/	Homorganic nasal + Halant
1.	లంక /laMka/	లఙ్క /laṅka/ 'island'
2.	కంచె /kaMce/	కఞ్చె [kaɲce] 'fence'
3.	పంట /paMTa/	పణ్ట /pa ṇ Ta/ 'harvest'
4.	కంత /kaMta/	కన్త /kanta/ 'hole'
5.	కంప /kaMpa/	కమ్ప /kampa/ 'thornybush'
6.	కంస /kaMsa/	కమ్స /kansa/ 'king Kansa'
7.	సింహ /siMha/	సిమ్హ /simha/ 'lion'

Table 4: Homorganic nasal and Homorganic nasal + Halant

3.5.3 Nasalization: Candrabindu (ఁ U+0C00) or arasunna (ఀ U+0C01)

Candrabindu, which denotes nasalization of the preceding vowel, is used in the Prakrit texts transcribed in the Telugu script and the *arasunna* as in old Telugu తెలుఁగు /telũgu/ 'telugu'. Present-day Telugu users do not use the candrabindu frequently unless to bring special emphasis as in hãã, hũũ, etc.

3.5.4 The Consonants

The Telugu consonants have an implicit vowel /a/ included in them. As per the traditional classification, they are categorized according to their phonetic properties. There are 5 *varga* groups (classes) and one non-*varga* group. Each *varga* corresponds to a particular set of stops characterized by particular place of articulation. Each *varga* contains four oral stops and one nasal stop ordered by the complexity of their manner from left to right as [-vd,-asp, -nas], [-vd, +asp, -nas], [+vd, -asp, -nas], [+vd, +asp, -nas], [+vd, -asp, +nas] (where, vd = voiced, asp = aspirated, nas = nasal). Each feature set defines the character by the *varga*. Each *varga* from top to bottom are defined by an additional place feature of articulation. The non-*varga* set is again divided into two subsets, each is characterized by absence or presence of sonority; i.e. [+/- son]. The obstruents characterized by [–son] are fricatives, viz. శ[ś], ష [ṣ], స [s], హ [h], while the remaining carry the feature of sonority i.e. [+son].

No.	Place of Articulation	-asp -vd -nas	ISO	+asp -vd -nas	ISO	-asp +vd -nas	ISO	+asp +vd -nas	ISO	-asp +vd +nas	ISO

1.	Velar	క	k	ఖ	kh	గ	g	ఘ	gh	ఙ	ṅ
2.	Palatal	చ	c	ఛ	ch	జ	j	ఝ	jh	ఞ	ñ
3.	Retroflex	ట	ṭ	ఠ	ṭh	డ	ḍ	ఢ	ḍh	ణ	ṇ
4.	Dental	త	t	థ	th	ద	d	ధ	dh	న	n
5.	Bilabial	ప	p	ఫ	ph	బ	b	భ	bh	మ	m

Table 5: Classification of stop consonants

Sonorants	య	y	ర	r	ఱ	ṛ	ల	l	ళ	ḷ	వ	v
Fricatives	శ	ś	ష	ṣ	స	s	హ	h				

Table 6: Non-stop consonants

4. The Development Process and Methodology

The Neo-Brahmi Generation Panel involves a number of different scripts with distinct Unicode blocks. Each of these scripts usually will have a separate LGR. However, a common thread runs through the neo-Brahmi scripts in the process of LGR development.

A number of guiding principles that are laid out will be used in the development of the scheme. As specified elsewhere, the NBGP adopts the following principles in the selection of code-points from the code-point repertoire for the Telugu language script. A principle, like the Inclusion principle, deals with whether the character is regularly used in the language, besides its unambiguous nature.

The second important principle, the exclusion principle, deals with the use of the code point repertoire for root zone and does not allow every character that is tabulated in the Unicode chart. A baseline layer of restriction is set for the Domain Name System by the protocol known as IDNA (Internationalized Domain Names in Applications). IDNA excludes some characters from the Unicode repertoire for the concerned script. An additional layer is added for the root zone, called the Maximal Starting Repertoire (MSR). Telugu does not have many such characters that are restricted. One such character for example is, the Avagraha " ఽ" (U+0C3D), which is restricted by MSR even if allowed by the IDNA protocol.

Similarly, certain punctuation marks that were used in the traditional texts are not assigned any code points and hence not necessary to be included here. Other cases such as symbols and abbreviations are not permitted. In addition to the above, rare and

obsolete characters though recognized in the Unicode chart of Telugu will not be permitted in the root zone LGR.

4.1 How to Avoid Duplicate Domain Names Involving ZWJ and ZWNJ?

ZWJ and ZWNJ are used mainly to write two distinct displays of the same consonant cluster or sequence which do not have any semantic and phonetic significance (see 4.2).

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.

If ZWJ and ZWNJs are allowed in domain names for Telugu, they create two distinct forms of the same domain name. To make the browsers and DNS to treat them as equal, we have to ignore ZWJ and ZWNJs for comparing two words. The same procedure is usually followed by the spell-checkers of the language.

4.2 Zero Width Joiner and Zero Width Non-Joiner in Telugu Domain Names

MSR excludes invisible characters like Zero Width Non-Joiner (U+200C) and Zero Width Joiner (U+200D), as they require ad hoc representation in different ways. These are required in certain cases where a typical visual shape of an akshar is desired.

There are contrastive usages of written forms derived from the use of Zero Width Joiner (ZWJ) and Zero Width Non-Joiner (ZWNJ). They have special roles in the writing system of Telugu.

ZWNJ is used in sequences like Consonant (C) + Halant (U+0C4D) + Consonant, where the second C is prevented from taking the usual dependent allograph (vattu) form after (below) the first consonant, as in the following example:

1. క (U+0C15) + ్ (U+0C4D) +స (U+0C38) + ్ (U+0C4D) +వ (U+0C35) +ా (0C3E) = క్స్వా– without using ZWNJ
   Example: వాక్స్వతంత్ర్యం

2. క (U+0C15) + ్ (U+0C4D) + ZWNJ (U+200C) +స (U+0C38) + ్ (U+0C4D) + వ (U+0C35) +ా (0C3E) = క్‌స్వా– using the ZWNJ
   Example: వాక్‌స్వాతంత్ర్యం

Both forms of the words though written with different graphic signs may mean the same and they are also same even in their pronunciation. Though the second form was not previously common, its usage is gaining ground due to the influence of English and Hindi. It is frequently used in transcribing many English words into Telugu, such as 'software' (సాఫ్ట్‌వేర్, using ZWNJ). The word 'software' will become సాఫ్ట్వేర్ if ZWNJ is not used.

4.3 Methodology to incorporate the feedback received through Public Comment process

The Telugu script LGR proposal was published for public comment to allow those who had not participated in the NBGP to make their views known. The NBGP analyzed all comments received to finalize the proposal. The analysis of public comments can be accessed online given at [112].

5. The Repertoire

In this section, we present the discussion on the code points that would form the repertoire of code points licensed by the [MSR] to be validated and used in the root zone label generation rules. Section 5.1 provides the section of the [MSR] applicable to the Telugu script on which the Telugu code point repertoire is based. Section 5.2 details the code point repertoire that the Neo-Brahmi Generation Panel proposes to be included in the Telugu LGR.

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

Figure 2: Telugu Code Page from [MSR]

Color convention[1]:

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

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

Not PVALID in IDNA2008 are in White background

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

5.2 Code Points Repertoire

In the following, the Telugu Script Unicode Code points have been presented and discussed with reference to the Principles that constrain the label generation rules. It is important to note that the purpose of this document is to state unambiguously the Telugu code points that can be used in the root zone repertoire.
The following table lists 63 code points for the Telugu LGR, out of a total number of 67 code points listed in MSR, excluding four code points which are obsolete.

No.	Unicode Code Point	Glyph	Character Name	EGIDS status	Indic Syllabic Category	Reference
1.	0C02	◌ం	TELUGU SIGN A NUSVARA	2 Tel 4 San 5 Others[2]	ANUSVĀRA	102, 103
2.	0C03	◌ః	TELUGU SIGN VISARGA	2 Tel 4 San 5 Others	VISARGA	102, 103
3.	0C05	అ	TELUGU LETTER A	2 Tel 5 Others	Vowel	102, 103
4.	0C06	ఆ	TELUGU LETTER AA	2 Tel 5 Others	Vowel	102, 103
5.	0C07	ఇ	TELUGU LETTER I	2 Tel 5 Others	Vowel	102, 103
6.	0C08	ఈ	TELUGU LETTER II	2 Tel 5 Others	Vowel	102, 103
7.	0C09	ఉ	TELUGU LETTER U	2 Tel 5 Others	Vowel	102, 103
8.	0C0A	ఊ	TELUGU LETTER UU	2 Tel 5 Others	Vowel	102, 103
9.	0C0B	ఋ	TELUGU LETTER VOCALIC R	2 Tel 5 Others	Vowel	102, 103
10	0C0E	ఎ	TELUGU LETTER E	2 Tel 5 Others	Vowel	102, 103
11.	0C0F	ఏ	TELUGU LETTER EE	2 Tel 5 Others	Vowel	102, 103

[2] Others are the EGIDS 5 languages, listed in Table 1: Main languages considered under Telugu LGR

No.	Unicode Code Point	Glyph	Character Name	EGIDS status	Indic Syllabic Category	Reference
12.	0C10	ఐ	TELUGU LETTER AI	2 Tel 5 Others	Vowel	102, 103
13.	0C12	ఒ	TELUGU LETTER O	2 Tel 5 Others	Vowel	102, 103
14.	0C13	ఓ	TELUGU LETTER OO	2 Tel 5 Others	Vowel	102, 103
15.	0C14	ఔ	TELUGU LETTER AU	2 Tel 5 Others	Vowel	102, 103
16.	0C15	క	TELUGU LETTER KA	2 Tel 5 Others	Consonant	102, 103
17.	0C16	ఖ	TELUGU LETTER KHA	2 Tel 5 Others	Consonant	102, 103
18.	0C17	గ	TELUGU LETTER GA	2 Tel 5 Others	Consonant	102, 103
19.	0C18	ఘ	TELUGU LETTER GHA	2 Tel 5 Others	Consonant	102, 103
20.	0C19	ఙ	TELUGU LETTER NGA	2 Tel 5 Others	Consonant, Nasal-Consonant	102, 103
21.	0C1A	చ	TELUGU LETTER CA	2 Tel 5 Others	Consonant	102, 103
22.	0C1B	ఛ	TELUGU LETTER CHA	2 Tel 5 Others	Consonant	102, 103
23.	0C1C	జ	TELUGU LETTER JA	2 Tel 5 Others	Consonant	102, 103
24.	0C1D	ఝ	TELUGU LETTER JHA	2 Tel 5 Others	Consonant	102, 103
25.	0C1E	ఞ	TELUGU LETTER NYA	2 Tel 5 Others	Consonant, Nasal-Consonant	102, 103
26.	0C1F	ట	TELUGU LETTER TTA	2 Tel 5 Others	Consonant	102, 103

No.	Unicode Code Point	Glyph	Character Name	EGIDS status	Indic Syllabic Category	Reference
27.	0C20	ఠ	TELUGU LETTER TTHA	2 Tel 5 Others	Consonant	102, 103
28.	0C21	డ	TELUGU LETTER DDA	2 Tel 5 Others	Consonant	102, 103
29.	0C22	ఢ	TELUGU LETTER DDHA	2 Tel 5 Others	Consonant	102, 103
30.	0C23	ణ	TELUGU LETTER NNA	2 Tel 5 Others	Consonant, Nasal-Consonant	102, 103
31.	0C24	త	TELUGU LETTER TA	2 Tel 5 Others	Consonant	102, 103
32.	0C25	థ	TELUGU LETTER THA	2 Tel 5 Others	Consonant	102, 103
33.	0C26	ద	TELUGU LETTER DA	2 Tel 5 Others	Consonant	102, 103
34.	0C27	ధ	TELUGU LETTER DHA	2 Tel 5 Others	Consonant	102, 103
35.	0C28	న	TELUGU LETTER NA	2 Tel 5 Others	Consonant, Nasal-Consonant	102, 103
36.	0C2A	ప	TELUGU LETTER PA	2 Tel 5 Others	Consonant	102, 103
37.	0C2B	ఫ	TELUGU LETTER PHA	2 Tel 5 Others	Consonant	102, 103
38.	0C2C	బ	TELUGU LETTER BA	2 Tel 5 Others	Consonant	102, 103
39.	0C2D	భ	TELUGU LETTER BHA	2 Tel 5 Others	Consonant	102, 103
40.	0C2E	మ	TELUGU LETTER MA	2 Tel 5 Others	Consonant, Nasal-Consonant	102, 103
41.	0C2F	య	TELUGU LETTER YA	2 Tel 5 Others	Consonant	102, 103

No.	Unicode Code Point	Glyph	Character Name	EGIDS status	Indic Syllabic Category	Reference
42.	0C30	ర	TELUGU LETTER RA	2 Tel 5 Others	Consonant	102, 103
43.	0C32	ల	TELUGU LETTER LA	2 Tel 5 Others	Consonant	102, 103
44.	0C33	ళ	TELUGU LETTER LLA	2 Tel 5 Others	Consonant	102, 103
45.	0C35	వ	TELUGU LETTER VA	2 Tel 5 Others	Consonant	102, 103
46.	0C36	శ	TELUGU LETTER SHA	2 Tel 5 Others	Consonant	102, 103
47.	0C37	ష	TELUGU LETTER SSA	2 Tel 5 Others	Consonant	102, 103
48.	0C38	స	TELUGU LETTER SA	2 Tel 5 Others	Consonant	102, 103
49.	0C39	హ	TELUGU LETTER HA	2 Tel 5 Others	Consonant	102, 103
50.	0C3E	ా	TELUGU VOWEL SIGN AA	2 Tel 5 Others	Matra	102, 103
51.	0C3F	ి	TELUGU VOWEL SIGN I	2 Tel 5 Others	Matra	102, 103
52.	0C40	ీ	TELUGU VOWEL SIGN II	2 Tel 5 Others	Matra	102, 103
53.	0C41	ు	TELUGU VOWEL SIGN U	2 Tel 5 Others	Matra	102, 103
54.	0C42	ూ	TELUGU VOWEL SIGN UU	2 Tel 5 Others	Matra	102, 103
55.	0C43	ృ	TELUGU VOWEL SIGN VOCALIC R	2 Tel 5 Others	Matra	102, 103
56.	0C44	ౄ	TELUGU VOWEL SIGN VOCALIC RR	2 Tel 5 Others	Matra	102, 103
57.	0C46	ె	TELUGU VOWEL SIGN E	2 Tel 5 Others	Matra	102, 103

No.	Unicode Code Point	Glyph	Character Name	EGIDS status	Indic Syllabic Category	Reference
58.	0C47	◌ే	TELUGU VOWEL SIGN EE	2 Tel 5 Others	Matra	102, 103
59.	0C48	◌ై	TELUGU VOWEL SIGN AI	2 Tel 5 Others	Matra	102, 103
60.	0C4A	◌ొ	TELUGU VOWEL SIGN O	2 Tel 5 Others	Matra	102, 103
61.	0C4B	◌ో	TELUGU VOWEL SIGN OO	2 Tel 5 Others	Matra	102, 103
62.	0C4C	◌ౌ	TELUGU VOWEL SIGN AU	2 Tel 5 Others	Matra	102, 103
63.	0C4D	◌్	TELUGU SIGN VIRAMA	2 Tel 5 Others	Matra	102, 103

Table 7: Included code points

5.3 Code Points Not Included

Referring to the principle in section 4, the code points to be excluded from the repertoire are the following, for the reasons listed.

The following code points are not in widespread use.

- 0C00 ◌ఀ TELUGU LETTER CANDRABINDU
- 0C01 ◌ఁ TELUGU LETTER ARASUNNA
- 0C0C ఌ TELUGU LETTER VOCALIC L
- 0C31 ఱ TELUGU LETTER RRA

Various signs: Allographs of vowel diacritics /a:/ and part of a diacritic specific to particular consonant /h/.

- 0C55 ◌ౕ TELUGU LENGTH MARK
- 0C56 ◌ౖ TELUGU AI LENGTH MARK

Historic phonetic variants: Phonological variants shall not be permitted. They are not in MSR.

- 0C58 చే TELUGU LETTER TSA
- 0C59 జే TELUGU LETTER DZA

The two additional vowels listed below to transcribe Sanskrit are not permitted. They are not in MSR.

- 0C60 ఋూ TELUGU LETTER VOCALIC RR
- 0C61 ఌూ TELUGU LETTER VOCALIC LL

The following two dependent vowels used to transcribe Sanskrit sounds are not permitted. They are not in MSR.

- 0C62 ◌ృ TELUGU VOWEL SIGN VOCALIC L
- 0C63 ◌ౄ TELUGU VOWEL SIGN VOCALIC LL

Starting from the MSR, There are four code points to be excluded.

No.	Unicode Code Point	Glyph	Character Name	EGIDS status	Indic Syllabic Category	Reference	Note
1.	0C0C	ఌ	TELUGU LETTER VOCALIC L	2 Telu 5 Gon 6b other	Vowel	103, 108, 109	It is not used in modern Telugu
2.	0C31	ఱ	TELUGU LETTER RRA	2 Telu 5 Gon 6b other	Consonant	103, 108, 109	It is not used in modern Telugu
3.	0C55	◌ఁ	TELUGU LENGTH MARK	2 Telu 5 Gon 6b other	Matra	103, 108, 109	It is not available on general keyboard.
4.	0C56	◌ఀ	TELUGU AI LENGTH MARK	2 Telu 5 Gon 6b other	Matra	103, 108, 109	It is not used in modern Telugu

Table 8: Excluded code points

6. Variants

Telugu code points representing the basic simple stand-alone characters and some dependent characters may enter into different combinations to form syllables. There are no characters in the Telugu Unicode chart that - either in simple form or in combined form are deemed similar by NBGP, when the restrictions of WLE (section 7) are taken into account. However, Telugu has a small number of variants that have identical values but derive from different character combinations. The NBGP categorizes these confusingly similar variants in two groups.

6.1 Type 1: Similarity within the Script

Certain vowels [o, ō] display different shapes in combination with certain consonants, though they have shared sound and code point values. For example:

i. Ca + e + u(:) -> mo(:)

ii. Ca + o(:) -> ko(:)

The variants, which are often confusing and of variable acceptance are due to the display of their rendering differently due to the identical code points.

These cases are interesting in that they present no similarity in their forms but have similar phonetic output. It is not unusual to find such regional variations and they are regularly used by Telugu users. These may not cause confusion but become annoying to learners.

However, ె + ు (U+0C46 + U+0C41) is matra + matra sequence, which is not allowed in the WLE rules in section 7. Therefore, these are not defined as variant sequences by NBGP.

Class	Character Sequence [Ca+e+u]	Glyph of [Ca+e+u]	Glyph of [Ca+o]	Character Sequence [Ca+o]
1	[క +ె +ు] -> 0C15+0C46+0C41 (This class includes other consonants like, kha, ga, nga, ca, cha, ja, nya, ta, tha, da, dha, na, ta, tha, da, dha, na, pa, pha, ba, bha, ra, la, va, Sa, sha, sa, and ha)	కు (Disallowed by WLE rule 2)	కొ	క+ొ 0C15+0C4A

Class	Character Sequence [Ca+e+u]	Glyph of [Ca+e+u]	Glyph of [Ca+o]	Character Sequence [Ca+o]
2	[మ +ె +ు] -> 0C2E+0C46+0C41	మొ (Disallowed by WLE rule 2)	మో	మ+ ో 0C2E +0C4A
	[య +ె +ు] -> 0C2F+0C46+0C41	యొ (Disallowed by WLE rule 2)	యో	య+ ో 0C2F +0C4A
	[ఱ +ె +ు] -> 0C1D+0C46+0C41	ఱొ (Disallowed by WLE rule 2)	ఱో	ఱ+ ో 0C1D +0C4A
	[ఘ +ె +ు] -> 0C18+0C46+0C41	ఘొ (Disallowed by WLE rule 2)	ఘో	ఘ+ ో 0C18+0C4A

Table 9a: Similarity within the script

6.2 Type 1: Variants within Script due to Alternative Spelling

Similar to the above, there are a set of representations in Telugu syllable formations where a homorganic nasal (anusvāra) in a syllable has alternate spelling which is represented visually different, as shown below.

No.	Homorganic nasal (anusvāra)+ consonant	Homorganic nasal consonant + halant + consonant
1.	లంక /laMka/	లఙ్క /laŋka/ ‘island’
2.	కంచె /kaMce/	కఞ్చె [kaɲce] ‘fence’
3.	పంట /paMTa/	పణ్ట /paNTa/ ‘harvest’
4.	కంత /kaMta/	కన్త /kanta/ ‘hole’
5.	కంప /kaMpa/	కమ్ప /kampa/ ‘thornybush’
6.	కంస /kaMsa/	కమ్స /kansa/ ‘king Kansa’
7.	సింహ /siMha/	సిమ్హ /simha/ ‘lion’

Table 9b: Variants with anusvāra alternating with nasal consonants

Writing alternatively with a nasal consonant + halant + consonant is rare in Telugu and often occur while transcribing Sanskrit words. Since the variants have exactly the same pronunciation, the GP considered whether the rarer representation of nasal consonant + halant + consonant should be disallowed in order to avoid the source of confusion.

Nasal Consonants are:

1. U+0C19 TELUGU LETTER NGA (ఙ)
2. U+0C1E TELUGU LETTER NYA (ఞ)
3. U+0C23 TELUGU LETTER NNA (ణ)
4. U+0C28 TELUGU LETTER NA (న)
5. U+0C2E TELUGU LETTER MA (మ)

Similarly and very frequently, the word final ము [mu] is often represented alternatively by the variant anusvāra ◌ం [M] as in the following:

కలం kalaM	కలము kalamu	'pen'
పుస్తకం pustakaM	పుస్తకము pustakamu	'book'
ఆముదం a:mudaM	ఆముదము a:mudamu	'castor oil'
దేశం deSaM	దేశము deSamu	'country'

~~In such cases, one of the confusable variants must be disallowed. This can be disallowed by the WLE rule: H cannot follow a nasal consonant.~~

Originally, the GP disallowed Halant following Nasal-C. However, in response to public comment that the rule was too restrictive to spelling rules, and the fact that homophonic variant is not in scope of Neo-Brahmi LGRs, the GP decided to remove the restriction. Therefore, H must follow any consonant.

6.3 Type 2: Shared Similarity with the Other Related Scripts.

There are many Brahmi derived scripts particularly in the Southern part of India, Sri Lanka, and South East Asia. Some of the characters of these scripts display similarity with each other. Such cases, relevant for Telugu script, are given below.

6.3.1 Type2: Cross-Script Variants for Telugu and Kannada

A number of characters of the Kannada script are almost similar to characters of Telugu script, except for the flattened head-stroke in Kannada contrasting with a tick mark on the top of the character in Telugu. Out of the total, there are 34 such cases which are categorized as variant sets, as shown in the following table.

Variant Set	Telugu Code Point	Kannada Code Point
1	◌ం (0C02)	◌ಂ (0C82)
2	◌ః (0C03)	◌ಃ (0C83)

Variant Set	Telugu Code Point	Kannada Code Point
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	఩ (0C2C)	಩ (0CAC)
26	ప (0C2D)	ಪ (0CAD)
27	మ (0C2E)	ಮ (0CAE)

Variant Set	Telugu Code Point	Kannada Code Point
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 10: Cross-script variant code points for Telugu and Kannada

The Telugu and Kannada variant sets in Table 10 are cross-script variant code points. The details of various *akshar* combinations and variant disposition can be found in section 6.4

Code points which have been analyzed and found to be similar, but not considered as variants, are listed in Appendix A.

6.3.2 Type2: Cross-Script Variants for Telugu and Devanagari

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

Devanagari Code Point	Telugu Code Point
ः (0903)	ఃు (0C03)

Table 11: Candidate cross-script variant code point for Telugu and Devanagari

6.3.3 Type2: Cross-Script Variants for Telugu and Gujarati

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

Gujarati Code Point	Telugu Code Point
ः (0A83)	ఃు (0C03)

Table 12: Candidate cross-script variant code point for Telugu and Gujarati

6.3.4 Type2 Cross-Script Variants for Telugu and Oriya

The following code points exhibit similarity between the Telugu and Oriya scripts.

Telugu Code Point	Oriya Code Point
ం (0C02) ANUSVĀRA	ଂ (0B20) LETTER TTHA
ఃం (0C03) SIGN VISARGA	ଃ (0B03) SIGN VISARGA
ఁ (0C30) LETTER RA	ଂ (0B20) LETTER TTHA

Table 13: Candidate cross-script variant code points for Telugu and Oriya

The first two (U+0C02 – U+0B20 and U+0C03 – U+0B03) are dependent signs and U+0C30 is a stand-alone character in Telugu. NBGP discussions concluded that there is no need to recognize the cross-script variant code points between the Oriya and the Telugu scripts. This is because U+0C30 and U+0B20 are distinguishable and there are not enough other variant code points in each script to form labels that look the same. Therefore, these are not defined as variant code points.

6.3.5 Type2: Cross-Script Variants for Telugu and Malayalam

The two code points, viz. the anusvāra and the visarga are the only identical signs between the Telugu and Malayalam scripts. However, as there are not enough other variant code points to form labels, they are not defined as variant code points between the two languages.

Telugu Code Point	Malayalam Code Point
◌ం (0C02)	ം (0D02)
◌ఃం (0C03)	ഃം (0D03)

Table 14: Candidate cross-script variant code points for Telugu and Malayalam

6.3.6 Type2: Cross-Script Variants for Telugu and Sinhala

The two code points, viz. the anusvāra and the visarga are the only identical signs between the Telugu and Sinhala scripts*. However, as there are not enough other variant code points to form labels, they are not defined as variant code points between the two languages.

*Note: Initially, the member of NBGP have proposed to include the code point ఁ (0C30) of Telugu and ಀ (0CB0) of Kannada as possible variant with that of Sinhala ఁ (0DBB). However, on the consideration of IP's comments, it was discussed with the members of Sinhala GP and finally agreed to drop 0C30, 0CB0, and 0DBB. Hence, none of these code points will be necessary to be recognized as cross script variants between the Telugu and

Sinhala. Similar decision is followed with the Kannada LGR proposal.

Telugu Code Point	Sinhala Code Point
◌ం (0C02)	◌ං (0D82)
◌ః (0C03)	◌ඃ (0D83)

Table 15: Candidate cross-script variant code points for Telugu and Sinhala

6.4 Cross Script Variants of Various *Akshar* Combinations

6.4.1 Conjunct Consonant Combinations

Cross script variants of various *Akshar* combinations (consonant-consonant-dependent characters) common between the Telugu and Kannada scripts include the following:

Variant Set	Telugu Code Point	Kannada Code Point
1	◌ం (0C02)	◌ಂ (0C82)
2	◌ః (0C03)	◌ಃ (0C83)
3	ఖ (0C16)	ಖ (0C96)
4	గ (0C17)	ಗ (0C97)
5	ఙ (0C1C)	ಙ (0C9C)
6	ఝ (0C1D)	ಝ (0C9D)
7	ఞ (0C1E)	ಞ (0C9E)
8	ట (0C1F)	ಟ (0C9F)
9	ఠ (0C20)	ಠ (0CA0)
10	డ (0C21)	ಡ (0CA1)
11	ఢ (0C22)	ಢ (0CA2)
12	ణ (0C23)	ಣ (0CA3)
13	థ (0C25)	ಥ (0CA5)
14	ద (0C26)	ದ (0CA6)
15	ధ (0C27)	ಧ (0CA7)
16	న (0C28)	ನ (0CA8)
17	఩ (0C2C)	಩ (0CAC)

Variant Set	Telugu Code Point	Kannada Code Point
18	భ (0C2D)	ಭ (0CAD)
19	మ (0C2E)	ಮ (0CAE)
20	య (0C2F)	ಯ (0CAF)
21	ర (0C30)	ರ (0CB0)
22	ల (0C32)	ಲ (0CB2)
23	ళ (0C33)	ಳ (0CB3)
24	ి (0C3F)	ಿ (0CBF)
25	ు (0C41)	ು (0CC1)
26	ృ (0C43)	ೃ (0CC3)

Table 16: Cross-script variants between Telugu and Kannada for conjunct consonant combination analysis

Table 16 includes 26 distinct Telugu code points that occur in the formation of conjunct consonant combinations in Telugu and Kannada. Excluding the stand alone vowels from the total common *Akshar* combinations of cross script variants, there are a set of 21 consonants (C), three vowel matras (M) and two vowel modifiers that enter into the formation of the following combinations:

Sl. No.	*Akshar* combinations	Number
1.	CM	= 21*3=63
2.	CB	= 21*1=21
3.	CX	= 21*1=21
4.	CHCM	=21*21*3=1323
5.	CHCB	=21*21*1=441
6.	CHCX	=21*21*1=441
7.	CHCMB	=21*21*3*1=1323
8.	CHCMX	=21*21*3*1=1323
9.	All combinations:	=4956

Table-17 total number of Akshar combinations

There occurs a total of 4956 conjunct consonant combinations modified by matras and vowel modifiers that are identical and can be labeled for variant labels between Telugu and Kannada scripts. These combinations are covered by the variant code points in Section 6, Table 10.

6.4.2 Other Combinations

NBGP created the possible combinations of Telugu code points and cross checked with other Neo-Brahmi scripts for candidate variants. The possible combinations are:

1. CHCMB, CHCMX
2. CHCM, CHCB, CHCX
3. VB, VX, V
4. CHC, CM, CB, CX, C

Where,

C	→	Consonant
M	→	Matra
V	→	Vowel
B	→	Anusvāra (Bindu)
X	→	Visarga
H	→	Halant / Virama

NBGP concludes that beside those identical code points defined as variants in Section 6, Table 10, there are no other variant code points between Telugu combinations and other scripts code points or code point combinations.

6.5 Variant disposition

As variants mentioned in Section 6, Table 10 can result in whole label variants, they may be considered for "blocked" disposition. 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.

7. Whole Label Evaluation Rules (WLE)

In this section we provide the WLEs that are required by the language. A number of rules have been formulated so that they can be adopted for LGR specification. Below are the symbols used in the WLE rules, for each of the "Indic Syllabic Category" as mentioned in the Table 7: Code point repertoire and the details of syllable formation, see Appendix B.

C	→	Consonant
M	→	Matra
V	→	Vowel
B	→	Anusvāra (Bindu)
X	→	Visarga
H	→	Halant / Virama
~~Nasal-C~~	~~→~~	~~Nasal Consonant~~ (Ref. Section 6.2 Type 1)

Rule 1. H must be preceded by C (Ref. Appendix B: Syllable formation Rule 4)
Rule 2. M must be preceded by C (Ref. Appendix B: Syllable formation Rule 6)
Rule 3. X must be preceded by V or M or C (Ref. Appendix B: syllable formation rule 3c, 5c and 7c)

Rule 4. B must be preceded by V or M or C (Ref. Appendix B: syllable formation rule 3b, 5b and 7b)
~~Rule 5. H cannot follow Nasal-C~~ (Ref. Section 6.2 Type 1)
Rule 6. V cannot be preceded by H

For Rule 6, 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 but first of which ends with a Halant and the second word begins with a Vowel. Some sections of the linguistic usage require the explicit presence of H for full representation of the sound intended. However, by and large, the form of the first word without the H is considered enough for full representation of the sound intended
as in the following examples:

Example:

'house of knowledge': దార్ అల్ఉలూమ్ da:rH alHulu:mH / దార్ అలుమ్ da:rH alulu:mH

'The Qor'an': ఖుర్ఆన్ KhurHa:nH / ఖురాన్ Khura:nH

'in Telangana Rashtra Samiti': టీఆర్ఎస్‌లో ti:a:rHesHlo / టీఆరెస్లో ti:a:rHesHlo

'Y.S. R. C. party': వైఎస్ఆర్‌సిపి vaiesHa:rHsi:pగ / వైఎసార్‌సిపి vaiesHa:rHsi:pi

'British India': బ్రిటిష్ఇండియా bHritiShHiMdiya / బ్రిటిషిండియా britiShiMdiya

The representations where there are cases with V preceded by H against where V is not preceded by H, the latter is awkward and the former is in demand in modern usage.

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 perceptually dissimilar but phonetically and semantically similarity between the two labels (with and without H) for majority of the linguistic community, hence this is explicitly prohibited by the NBGP.

8. Contributors

Gangadhar Panday
Uma Maheshwara Rao, G.
NBGP members

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 18 February 2019)

[101] Disanayaka, J.B. 2017. Encyclopedia of Sinhala Language and Culture. Colombo: Sumitha Publishers. First edition 2012.

[102] Krishnamurti, Bhadriraju, Ed., 2000. *Telugu bhaashaa charitra*. Hyderabad: P.S. Telugu University. First edition 1974.

[103] Krishnamurti, Bhadriraju and J P L Gwynn. 1985. *A Grammar of Modern Telugu.* New Delhi: Oxford University Press. *ISBN 978-0-19-561664-4*. Delhi.

[104] Sarma, I. K. 1980. *Coinage of Satavahana Empire***.** Delhi : Agam Kala Prakashan,

[105] Sridhar, S.N. 1980. *Kannada*. New York: Routledge.

[106] Suresh, Kolichala. 2012. Proposal to encode Telugu LLLA, Telugu ౷: http://eemaata.com/unicode-proposal/telugu-llla-proposal.pdf. Accessed on 9 July 2018.

[107] Suresh, Kolichala. 2012. Divergent developments of alveolar stop *t̲ in Telugu http://kolichala.com/dravidian/Divergent_developments_of_alveolar_stop_in_Telugu.pdf. Accessed on 9 July 2018.

[108] Telugu Unicode Chart, Telugu Range: 0C00–0C7F. The Unicode Standard, Version 10.0. http://www.unicode.org/Public/10.0.0/charts. Accessed on 9 July 2018.

[109] Uma Maheshwara Rao, G. 2012. Telugu bhaasha-saMgaNanaM. Hyderabad: P.S. Telugu University. ISBN: 81-86073-372-9.

[110] Uma Maheshwara Rao, G. 2003. Standard Telugu Written Language. VIDYULLIPI-4. pp. 1-14. Hyderabad: SCIL.

[111] Usha Devi, A. and Chandra Sekhara Reddy. D. 2015. Peoples Linguistic Survey of India. Andhra Pradesh and Telangana rAshtraala bhaashalu, vol.3, part 1. ISBN: 978-93-85231-05-6. Hyderabad: emesco.

[112] Public comment feedback for Kannada, Telugu, Oriya Script LGR Proposals, https://docs.google.com/document/d/1m9MbBfNBQZAFc9SOYpt0lgeeyM3N-DsUP173J4Vb948 (Accessed on 18 February 2019)

Appendix A: Confusable Code Points Analysis

A-1. Telugu and Kannada

The following table defines Telugu and Kannada code points which are confusable.

No.	Telugu		Kannada	
	CP	Glyph	CP	Glyph
1	0C35	వ	0CB5	ವ
2	0C36	శ	0CB6	ಶ
3	0C38	స	0CB8	ಸ

Table A-1: Confusable code points of Telugu and Kannada script

The following table lists other code points which have been analyzed and concluded that they are distinguishable.

No.	Telugu		Kannada		NBGP resolution
	CP	Glyph	CP	Glyph	
1	0C0E	ఎ	0C8E	ಎ	distinguishable
2	0C18	ఘ	0C98	ಘ	distinguishable
3	0C19	ఙ	0C99	ಙ	distinguishable
4	0C1A	చ	0C9A	ಚ	distinguishable
5	0C1B	ఛ	0C9B	ಛ	distinguishable
6	0C2A	ప	0CAA	ಪ	distinguishable
7	0C2B	ఫ	0CAB	ಫ	distinguishable
8	0C37	ష	0CB7	ಷ	distinguishable
9	0C4C	ౌ	0CCC	ೌ	distinguishable

Table A-2: Other NBGP resolutions on Telugu and Kannada script

A-2. Telugu and Malayalam

Beside those identical code points defined as variants in Section 6, there are no other similar code points between Telugu and Malayalam.

A-3. Telugu and Sinhala

Beside those identical code points defined as variants in Section 6, there are no other similar code points between Telugu and Sinhala.

Appendix B: Syllable formation in the Telugu Script

The Telugu script grammar allows us to state the nature and structure of the graphic syllables in the formation of words. The extended notion of syllable is often used to characterize orthographies of South-Asian scripts especially Brahmi derived scripts where words are composed of sequences of one or more orthographic *aksharas* or syllables. These aksharas are again composed of sequences of certain characters from the alphabet. The Telugu alphabet has the following types of characters (encoded into the Unicode) that either on their own or by entering larger combinations form *aksharas* as shown here. There are 12 different types of syllables possible in Telugu:

The following Variables are involved in the formation of syllable [$]:

- C = Consonants, that are standalone characters or graphemes with an inherent vowel `a' can function as syllables;
  Stops: క ఖ గ ఘు ఙ చ ఛ జ ఝు ఞ ట ఠ డ ఢ ణ త థ ద ధ న ప ఫ బ భ మ;
  Fricatives: శ ష స హ
  Sonorants: య ర ఱ ల ళ వ

- V = Vowels, that stand alone and represented by the graphic signs of the following may function as syllables;
  అ ఆ ఇ ఈ ఉ ఊ ఎ ఏ ఐ ఒ ఓ ఔ ఋ

- M = Matras or the dependent vowel signs when occur with a consonant may function as syllables (characteristically delete the inherent vowel of the consonant);
  Example. కా కి కీ కు కూ కె కే కై కొ కో కౌ; etc.

- H = Halant or virama = ్; It may occur with one of the consonants represented by C to form CH syllables;
  Example. క్ ఖ్ గ్ ఘ్ ఙ్

- B= Pūrṇānusvāra, the homorganic nasal and an Archiphoneme = ం, may occur with one of the C, V, and the combined CM to form CB, CMB, VB, and C([HC]*)B
-
- X= visarga or the glottal check= ఃు, may occur with one of the C, V, and the combined CM to form CX, CMX, VX

The operators used: The following four operators are employed to define the delimitation of the graphic syllables in Telugu.

No.	Symbol	Function;
1.	|	Alternative;
2.	[]	encloses optional elements;
3.	*	Variable occurrence;
4.	()	The sequence cluster;

Table B-1 symbols and functions

An Akshara in Telugu can be defined as any C or V and a combination of M (dependent vowels), and the vowel modifiers as in the following:

The following syllable formation rules derive all possible graphic syllables in Telugu.

1. The syllable formation rule-1, a $= V;

Every standalone vowel character can function as a syllable, Ex.

అ, ఆ, ఇ, ఈ, ఉ, ఊ, ఎ, ఏ, ఐ, ఒ, ఓ, ఔ, ఋ;

After the exclusion of obsolete vowels 13 syllables are possible.

2. The syllable formation rule-2, a $= C;

Every standalone consonant character can function as a syllable, Ex.

క ఖ గ ఘ ఙ,
చ ఛ జ ఝ ఞ,
ట ఠ డ ఢ ణ,
త థ ద ధ న,
ప ఫ బ భ మ,
య ర ఱ ల ళ వ,
శ ష స హ;

There are 35 such syllables are possible.

3. Syllable formation rule-3, $=VB|X;

Example:

3a=V+B=$; అం ఆం ఇం ఈం ఉం ఊం ఎం ఏం ఐం ఒం ఓం ఔం;
3b=V+X=$; అః ఆః ఇః ఈః ఉః ఊః ఎః ఏః ఐః ఒః ఓః ఔః;

In combination with V and one of the two B or X, a total 36 syllables are possible. Syllable combinations with vocalic R are not used.

4. Syllable formation rule-4, a $= CH;

A standalone consonant may be appended by the halant marker H to form the corresponding graphic syllables as shown here.

Example:

క్ ఖ్ గ్ ఘ్ ఙ్
చ్ ఛ్ జ్ ఝ్ ఞ్
ట్ ఠ్ డ్ ఢ్ ణ్
త్ థ్ ద్ ధ్ న్
ప్ ఫ్ బ్ భ్ మ్
య్ ర్ ఱ్ ల్ ళ్ వ్
శ్ ష్ స్ హ్

There are 35 such graphic syllables are possible.

5. Syllable formation rule-5, $=CB|X; Ex.

Standalone consonants can take one of the three vowel modifiers and form the corresponding syllables as shown below:

Example:

5a. $=CB: కం ఖం గం ఘం ఙం చం ఛం జం ఝం ఞం టం ఠం etc.

5b. $=CX: కః ఖః గః ఘః ఙః చః ఛః జః ఝః ఞః టః ఠః etc.

There are 2*35=70 graphic consonant modifier syllables are possible.

6. Syllable formation rule-6, $=CM;

A consonant may get attached with a vowel modifier or the dependent vowel diacritic to form the corresponding syllables;

Example:

కా కి కీ కు కూ కృ క కె కే కై కొ కో కౌ; etc.

A total of 35*13 consonant + vowel diacritic combinations may derive 455 graphic syllables in Telugu.

7. Syllable formation rule-7, $=CMB|X;

A consonant with a dependent vowel when followed by one of the three modifiers may derive the following graphic syllables;

Example:

7a. కాం కిం కీం కుం కూం కెం కేం కైం కొం కోం కౌం

7b. కాఁ కిఁ కీఁ కుఁ కూఁ కెఁ కేఁ కైఁ కొఁ కోఁ కౌఁ

A total of 35*12*2 consonant plus a dependent vowel and one of the three modifiers derive 840 possible graphic syllables in Telugu.

8. Syllable formation rule-8, $=CH[(C)*C];

Any consonant followed by the halant marker may combine with another consonant or consonants to form complex graphic syllables;

Example:

2 consonant clusters: ఖ్ఖ గ్గ, ఘ్ఘ, జ్జ, చ్చ, ఛ్చ, జ్జ, ఝ్ఝ, ఞ్ఞ, ట్ట, ఠ్ఠ, డ్డ, ఢ్ఢ, ణ్ణ, etc.

3 consonant clusters: ర్ద్ర, ష్ట్ర, స్త్ర, న్ద్ర, జ్ఞ్ర ష్ట్ర, త్య్, త్స్న etc.

4 consonant clusters: త్స్న్య ;

A total of 35*1*35 =1225 CHC syllables involving two consonant clusters are possible; Further, a total of 35*1*35*1*35 =42,875 CHCHC syllables involving three consonant clusters are possible; Though four consonant clusters are extremely rare but theoretically possible as shown above.

9. Syllable formation rule-9, $=CH(CH[CH])CM;

Any consonant followed by the halant marker and a consonant or consonants may be appended by one of the dependent vowels to form complex graphic syllables involving two to three consonant clusters;

Example:

క్కు ఖ్ఖి గ్గొ, మ్ఘు, జ్జా, చ్చా, చ్ఛూ, జ్జే, ఝ్ఝు, ఞ్ఞో, ట్టీ, ఠ్ఠూ, డ్డా, ఢ్ఢూ, ణ్ణా, etc.

ర్దు, ప్త్రీ, స్త్రీ, న్ద్రి, జ్ఞి ష్ట్లా, క్ర్యా, త్న్నా etc.

త్స్న్యూ

A total of 35*1*35*1*12= 14,700 complex syllables involving two consonant clusters followed by dependent vowels are possible.

A total of 35*1*35*1*35*12= 5,14,500 complex syllables involving three consonant clusters followed by dependent vowels are possible.

The following is a summary of possible syllable types with the glyphs in Telugu:

$= V([B|X])|CM([B|X])|CH(CH[C])M([B|X])

As per our definition the following 21 subtypes of graphic syllables are possible which however can be grouped under 8 rules as discussed above.

$ = V|VB|VX|
C|CB|CX|CM|CH|
CHC|CHCB|CHCX|CHCMre CHCH|CHCHC|CHCHCB|CHCHCX|CHCHCM

Therefore, typologically 8 distinct types of graphic syllables can be derived in the language.

Appendix C: 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.