Proposal for a Telugu Script Root Zone Label Generation Ruleset [LGR]

*LGR Version:* 3.0

*Date:* 2018-01-15

*Document version:* 1.5

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

# General Information/ Overview/ Abstract

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

# 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: telɯgɯ
Native name of the script: తెలుగు
Maximal Starting Repertoire [MSR] version: 2
The Unicode Standard, Version 6.3
Telugu Range: 0C00–0C7F

# Background on Script and Principal Languages Using It

The Telugu language uses the script called 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 or their variants 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 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 for the Telugu and Kannada scripts. The earliest known inscriptions containing Telugu words appear on the bilingual coins of Satavahanas that date back to 400 BC. 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. Until the 20th century Telugu was written in granthic style very different from the colloquial language. During the 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.

 



## 3.2 Notable features

The Telugu orthography superficial resembles 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 have 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 the secondary vowel. A Consonant cluster may be formed with a single standalone character followed by one or more secondary forms of consonants. The direction of writing is from left to right. The order of composition of syllabaries do 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 languages is a Dravidian language spoken by about 75 million (ca. 2011) 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 [for publishing books]. Particularly with reference to these, except for Sanskrit, primary school text books, some reading material and dictionaries are published for these languages in this script.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 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 |
| 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 writing system of Telugu consists of sixteen character signs representing vowels that can stand alone and fifteen dependent signs corresponding to sixteen vowels excepting /a/ అ which does not exist as an explicit symbol but can be found as an inherent sound with the consonants. Besides these, there are four more dependent symbols which always occur with the vowels as extensions. This could be summed up as in the following:

### 3.5.1 The vowels and vowel modifiers

There are fourteen vowel characters viz. అ, ఆ, ఇ, ఈ, ఉ, ఊ, ఋ, ఌ, ఎ, ఏ, ఐ, ఒ, ఓ, ఔ, in the common inventory and two (ౠ, ౡ) which are obsolete. Each member of the common inventory has one to many secondary variants depending on the size and shape of the consonant that functions as an anchor. There are six modifiers that concern with vowels, viz., [~ ] ఁ, ం [ṁ], ః [ḥ], ँ [ M] a special symbol not common in standard Telugu writings, the symbol avagraha ఽ [:.] is used as a common symbol to indicate doubling the vowel length follows only long vowels, and finally a symbol ్ [H] when appended to consonants it deducts the inherent vowel /a/ from the consonant. The Halanta sign has the the same characteristic as that of a secondary vowel sign in that both of them are added to consonants and they delete the inherent vowel [a].

R1. Inherent vowel deletion rule:

C[ca] + M [ా, ి … ] | ్ [H] -> C[cా, ి] | ్

C[ca] + M [0C3E-3F, 0C40-44, 0C62-63, 0C46-48, 0C4A-4C]|[0C4D] ->

C[c]M[0C3E-3F, 0C40-44, 0C62-63, 0C46-48, 0C4A-4C]|[0C4D]

C = Consonant, ca= a consonant with an inherent ‘a’, M =Secondary vowel;

| No. | Independent vowelsPrimary allographsWith code points | Dependent vowelsSecondary allographsWith code points |
| --- | --- | --- |
| 1.  | అ 0C05 | No explicit signrecognized or encoded |
| 2.  | ఆ 0C06 | ా 0C3E |
| 3.  | ఇ 0C07 | ి 0C3F |
| 4.  | ఈ 0C08 | ీ 0C40 |
| 5.  | ఉ 0C09 | ు 0C41 |
| 6.  | ఊ 0C0A | ూ 0C42 |
| 7.  | ఋ 0C0B | ృ 0C43 |
| 8.  | ౠ 0C60 | ౄ 0C44 |
| 9.  | ఌ 0C0F | ౢ 0C62 |
| 10.  | ౡ 0C61 | ౣ 0C63 |
| 11.  | ఎ 0C0E | ె 0C46 |
| 12.  | ఏ 0C0F | ే 0C47 |
| 13.  | ఐ 0C10 | ై 0C48 |
| 14.  | ఒ 0C12 | ొ 0C4A |
| 15.  | ఓ 0C13 | ో 0C4B |
| 16.  | ఔ 0C14 | ౌ 0C4C |

Table 2: Vowels and the corresponding dependent signs

|  |  |
| --- | --- |
| No. | Modifier signs With code points |
| 1. | ँ 0C00 |
| 2. | ఁ 0C01 |
| 3. | ం 0C02 |
| 4. | ః 0C03 |
| 5. | ఽ 0C3D |
| 6. | ్ 0C4D |

Table 3: Vowel modifiers and the consonantal modifiers

### 3.5.2 The Anusvara or sunna (ం - U+0C02)

The Anusvara 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 occur 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 /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 |
| 8. | కలం /kalaM/ | కలము /kalamu/ ‘pen’ |

Table 4: Homorganic nasal and Homorganic nasal + Halant

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

Candrabindu denotes nasalization of the preceding vowel 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 characterised 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]. 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 are again divided into two subsets, each is characterized by absence or presence of sonority i.e. [+/- son]. The obstruents which are characterized by [-cont] i.e. non-continuous nature are further characterized by [–son] are fricatives, viz. శ[ś], ష [ṣs], స [s], హ [h] while the remaining carry the feature of sonority i.e. [+son].

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| No. | *Varga or*Place of Articulation | -asp-vd-nas | +asp-vd-nas | -asp+vd-nas | +asp+vd-nas | -asp+vd+nas |
| 1. | Velar | క | ఖ | గ | ఘ | ఙ |
| 2. | Palatal | చ | ఛ | జ | ఝ | ఞ |
| 3. | Retroflex | ట | ఠ | డ | ఢ | ణ |
| 4. | Dental | త | థ | ద | థ | న |
| 5. | Bilabial | ప | ఫ | బ | భ | మ |

Table 5: Varga classification of consonants

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Non-vargaconsonants | య | ర | ల | ళ | వ |
| శ | ష | స | - | హ |

Table 6: Non-Varga 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 which does not allow each and every character that is recognized in the Unicode chart. Another special layer that is less restrictive is the Domain Name System which is governed by an additional protocol known as IDNA (Internationalized Domain Names in Applications). This domain may exclude some characters from the Unicode repertoire for the concerned language. However, Telugu does not have many such characters that shall be restricted as per this principle. One such character for example is, the Telugu Sign Avagraha " ఽ" (U+oc3d) even if allowed by IDNA protocol, may not be permitted in the Root Zone Repertoire as per the MSR. 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 the following will not be permitted in the root zone LGR.

# 5. The Repertoire

In this section we describes the Maximal Starting Repertoire (MSR) for the Label Generation Rules (LGR) described in “Procedure to Develop and Maintain the Label Generation Rules for the Root Zone in Respect of IDNA Labels”. The Procedure involves a two-stage process, the Generation Panels (GP) propose LGRs for a given script, which are then reviewed and integrated by the Integration Panel (IP).
 As a starting point for the GPs, the IP establishes the maximal set of code points as well as the default whole label variant evaluation rules for the Root Zone. Collectively these are called as the MSR.

The essential goal of the Language Generation Panel is to recommend an acceptable list of code points to be included in the Root Zone Repertoire. Therefore, it is necessary, as part of the generation panel to recognize the relevant code points that shall be included in the root zone repertoire.

## 5.1 Code Point Repertoire Includes:



Fig. 1. The Unicode Standard 6.3 (1991-2017 Unicode, Inc .)

In the following, the Telugu language Unicode Code points have been presented and discussed with reference to the Principles that constrain the label generation rules. Issues Related to the Management of IDN Variant TLDs (IIR) and the Principles for Unicode Code Point Inclusion or exclusion in Labels in the DNS [IABCP] have several principles that constrain the development of the label generation rules for any zone, including the root zone. It’s 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 start from the 67 code points in MSR-2 and there are 63 code points to be included.

| No. | Unicode Code Point |  Glyph | Character Name | GC | EGIDS status | Indic Syllabic Category | Ref. |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 1.  | 0C02 | ం |  TELUGU SIGN ANUSVARA | Mc | 2 Tel4 San5 Others[[1]](#footnote-1) | NIMDUSUNNA | 2, 3 |
| 2.  | 0C03 | ః |  TELUGU SIGN VISARGA | Mc | 2 Tel4 San5 Others | VISARGA | 2, 3 |
| 3.  | 0C05 | అ | TELUGU LETTER A | Lo | 2 Tel5 Others | Vowel | 2, 3 |
| 4.  | 0C06 | ఆ | TELUGU LETTER AA | Lo | 2 Tel5 Others | Vowel | 2,3 |
| 5.  | 0C07 | ఇ | TELUGU LETTER I | Lo | 2 Tel5 Others | Vowel | 2, 3 |
| 6.  | 0C08 | ఈ | TELUGU LETTER II | Lo | 2 Tel5 Others | Vowel | 2, 3 |
| 7.  | 0C09 | ఉ | TELUGU LETTER U | Lo | 2 Tel5 Others | Vowel | 2, 3 |
| 8.  | 0C0A | ఊ | TELUGU LETTER UU | Lo | 2 Tel5 Others | Vowel | 2, 3 |
| 9.  | 0C0B | ఋ | TELUGU LETTER VOCALIC R | Lo | 2 Tel5 Others | Vowel | 2,3 |
| 10  | 0C0E | ఎ | TELUGU LETTER E | Lo | 2 Tel5 Others | Vowel | 2,3 |
| 11.  | 0C0F | ఏ | TELUGU LETTER EE | Lo | 2 Tel5 Others | Vowel | 2, 3 |
| 12.  | 0C10 | ఐ | TELUGU LETTER AI | Lo | 2 Tel5 Others | Vowel | 2, 3 |
| 13.  | 0C12 | ఒ | TELUGU LETTER O | Lo | 2 Tel5 Others | Vowel | 2, 3 |
| 14.  | 0C13 | ఓ | TELUGU LETTER OO | Lo | 2 Tel5 Others | Vowel | 2, 3 |
| 15.  | 0C14 | ఔ | TELUGU LETTER AU | Lo | 2 Tel5 Others | Vowel | 2, 3 |
| 16.  | 0C15 | క | TELUGU LETTER KA | Lo | 2 Tel5 Others | Consonant | 2, 3 |
| 17.  | 0C16 | ఖ | TELUGU LETTER KHA | Lo | 2 Tel5 Others | Consonant | 2, 3 |
| 18.  | 0C17 | గ | TELUGU LETTER GA | Lo | 2 Tel5 Others | Consonant | 2, 3 |
| 19.  | 0C18 | ఘ | TELUGU LETTER GHA | Lo | 2 Tel5 Others | Consonant | 2, 3 |
| 20.  | 0C19 | ఙ | TELUGU LETTER NGA | Lo | 2 Tel5 Others | Consonant | 2, 3 |
| 21.  | 0C1A | చ | TELUGU LETTER CA | Lo | 2 Tel5 Others | Consonant | 2, 3 |
| 22.  | 0C1B | ఛ | TELUGU LETTER CHA | Lo | 2 Tel5 Others | Consonant | 2, 3 |
| 23.  | 0C1C | జ | TELUGU LETTER JA | Lo | 2 Tel5 Others | Consonant | 2, 3 |
| 24.  | 0C1D | ఝ | TELUGU LETTER JHA | Lo | 2 Tel5 Others | Consonant | 2, 3 |
| 25.  | 0C1E | ఞ | TELUGU LETTER NYA | Lo | 2 Tel5 Others | Consonant | 2, 3 |
| 26.  | 0C1F | ట | TELUGU LETTER TTA | Lo | 2 Tel5 Others | Consonant | 2, 3 |
| 27.  | 0C20 | ఠ | TELUGU LETTER TTHA | Lo | 2 Tel5 Others | Consonant | 2, 3 |
| 28.  | 0C21 | డ | TELUGU LETTER DDA | Lo | 2 Tel5 Others | Consonant | 2, 3 |
| 29.  | 0C22 | ఢ | TELUGU LETTER DDHA | Lo | 2 Tel5 Others | Consonant | 2, 3 |
| 30.  | 0C23 | ణ | TELUGU LETTER NNA | Lo | 2 Tel5 Others | Consonant | 2, 3 |
| 31.  | 0C24 | త | TELUGU LETTER TA | Lo | 2 Tel5 Others | Consonant | 2, 3 |
| 32.  | 0C25 | థ | TELUGU LETTER THA | Lo | 2 Tel5 Others | Consonant | 2, 3 |
| 33.  | 0C26 | ద | TELUGU LETTER DA | Lo | 2 Tel5 Others | Consonant | 2, 3 |
| 34.  | 0C27 | ధ | TELUGU LETTER DHA | Lo | 2 Tel5 Others | Consonant | 2, 3 |
| 35.  | 0C28 | న | TELUGU LETTER NA | Lo | 2 Tel5 Others | Consonant | 2, 3 |
| 36.  | 0C2A | ప | TELUGU LETTER PA | Lo | 2 Tel5 Others | Consonant | 2, 3 |
| 37.  | 0C2B | ఫ | TELUGU LETTER PHA | Lo | 2 Tel5 Others | Consonant | 2, 3 |
| 38.  | 0C2C | బ | TELUGU LETTER BA | Lo | 2 Tel5 Others | Consonant | 2, 3 |
| 39.  | 0C2D | భ | TELUGU LETTER BHA | Lo | 2 Tel5 Others | Consonant | 2, 3 |
| 40.  | 0C2E | మ | TELUGU LETTER MA | Lo | 2 Tel5 Others | Consonant | 2, 3 |
| 41.  | 0C2F | య | TELUGU LETTER YA | Lo | 2 Tel5 Others | Consonant | 2, 3 |
| 42.  | 0C30 | ర | TELUGU LETTER RA | Lo | 2 Tel5 Others | Consonant | 2, 3 |
| 43.  | 0C32 | ల | TELUGU LETTER LA | Lo | 2 Tel5 Others | Consonant | 2, 3 |
| 44.  | 0C33 | ళ | TELUGU LETTER LLA | Lo | 2 Tel5 Others | Consonant | 2, 3 |
| 45.  | 0C35 | వ | TELUGU LETTER VA | Lo | 2 Tel5 Others | Consonant | 2, 3 |
| 46.  | 0C36 | శ | TELUGU LETTER SHA | Lo | 2 Tel5 Others | Consonant | 2, 3 |
| 47.  | 0C37 | ష | TELUGU LETTER SSA | Lo | 2 Tel5 Others | Consonant | 2, 3 |
| 48.  | 0C38 | స | TELUGU LETTER SA | Lo | 2 Tel5 Others | Consonant | 2, 3 |
| 49.  | 0C39 | హ | TELUGU LETTER HA | Lo | 2 Tel5 Others | Consonant | 2, 3 |
| 50.  | 0C3E | ా | TELUGU VOWEL SIGN AA | Mn | 2 Tel5 Others | Matra | 2, 3 |
| 51.  | 0C3F | ి | TELUGU VOWEL SIGN I | Mn | 2 Tel5 Others | Matra | 2, 3 |
| 52.  | 0C40 | ీ |  TELUGU VOWEL SIGN II | Mn | 2 Tel5 Others | Matra | 2, 3 |
| 53.  | 0C41 | ు |  TELUGU VOWEL SIGN U | Mc | 2 Tel5 Others | Matra | 2, 3 |
| 54.  | 0C42 | ూ | TELUGU VOWEL SIGN UU | Mc | 2 Tel5 Others | Matra | 2, 3 |
| 55.  | 0C43 | ృ |  TELUGU VOWEL SIGN VOCALIC R | Mc | 2 Tel5 Others | Matra | 2, 3 |
| 56.  | 0C44 | ౄ |  TELUGU VOWEL SIGN VOCALIC RR | Mc | 2 Tel5 Others | Matra | 2, 3 |
| 57.  | 0C46 | ె |  TELUGU VOWEL SIGN E | Mn | 2 Tel5 Others | Matra | 2, 3 |
| 58.  | 0C47 | ే |  TELUGU VOWEL SIGN EE | Mn | 2 Tel5 Others | Matra | 2, 3 |
| 59.  | 0C48 | ై |  TELUGU VOWEL SIGN AI | Mn | 2 Tel5 Others | Matra | 2, 3 |
| 60.  | 0C4A | ొ |  TELUGU VOWEL SIGN O | Mn | 2 Tel5 Others | Matra | 2, 3 |
| 61.  | 0C4B | ో |  TELUGU VOWEL SIGN OO | Mn | 2 Tel5 Others | Matra | 2, 3 |
| 62.  | 0C4C | ౌ |  TELUGU VOWEL SIGN AU | Mn | 2 Tel5 Others | Matra | 2, 3 |
| 63.  | 0C4D | ్ |  TELUGU SIGN VIRAMA | Mn | 2 Tel5 Others | Matra | 2, 3 |

Table 7: Included code points

## 5.3 Code points not included:

Refers to the principle in section 4, the code points to be excluded from the Repertoire are:

**The following two code points are not in widespread use and may not be permitted.**

* 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/. They need to be blocked.**

* 0C55 ౕ TELUGU LENGTH MARK
* 0C56 ౖ TELUGU AI LENGTH MARK

**Historic phonetic variants: Phonological variants shall not be permitted. They are not in MSR-2.**

* 0C58 ౘ TELUGU LETTER TSA
* 0C59 ౙ TELUGU LETTER DZA
* 0C5A ~~ద~~ TELUGU LETTER RRRA (letter for an alveolar consonant whose exact phonetic value is not known).

**The two additional vowels listed below to transcribe Sanskrit may not be permitted. They are not in MSR-2.**

* 0C60 ౠ TELUGU LETTER VOCALIC RR
* 0C61 ౡ TELUGU LETTER VOCALIC LL

**The following two Dependent vowels of the vowel signs to transcribe Sanskrit sounds may not be permitted. They are not in MSR-2.**

* 0C62 ౢ TELUGU VOWEL SIGN VOCALIC L
* 0C63 ౣ TELUGU VOWEL SIGN VOCALIC LL

Starting from the MSR-2, There are four code points to be excluded.

| No. | Unicode Code Point |  Glyph | Character Name | GC | EGIDS status | Indic Syllabic Category | Ref. | Current and wide-spread use?[Y/N] |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1.  | 0C0C | ఌ | TELUGU LETTER VOCALIC L | Lo | 2 Telu5 Gon6b other | Vowel | 3, 8, 9 | No |
| 2.  | 0C31 | ఱ | TELUGU LETTER RRA | Lo | 2 Telu5 Gon6b other | Consonant | 3, 8, 9 | No |
| 3.  | 0C55 | ౕ |  TELUGU LENGTH MARK | Mn | 2 Telu5 Gon6b other | Matra | 3, 8, 9 | Yes, but to be excluded by principle |
| 4.  | 0C56 | ౖ |  TELUGU AI LENGTH MARK | Mn | 2 Telu5 Gon6b other | Matra | 3, 8, 9 | No |

Table 8: Excluded code points

# 6. Variants

Standard Unicode encoding of the Telugu code points represented basic simple stand alone Characters and some dependent characters which enter into different combinations in syllable formation. There are no characters in the Telugu Unicode chart that are the result of combinations of the characters permitted as per the [MSR] or at least have formal similarity. However, Telugu has a small number of variants that have identical values but derive from different character combinations. The other variants occur due to dissimilarity The NBGP categorizes these confusingly similar variants in two groups.

## Type 1: Similarity within the script

Shared sound value but formally different due to the combinations of different code points
 Ex. i. Ca +e+u -> ko
 ii. Ca+o: -> ko

Variation due to display and rendering differently of identical code points which are often confusing and of variable acceptance.

|  |  |  |  |
| --- | --- | --- | --- |
| Sr.No. | Character seq. [Ca+e+u] | -> :: Co <-  | Ca+o  |
| 1 | [ క +ె +ు] -> 0C15+0C46+0C41 | కెు :: కొ  | క+ొ0C15+0C4A |
| 2 | [ మ +ె +ు] -> 0C2E+0C46+0C41 |  మొ :: మొ  | మ+ ొ0C2E +0C4A |
| 3 | [య +ె +ు] -> 0C2E+0C46+0C41 | యొ :: యొ | య+ ొ0C2E +0C4A |
| 4 | [ ఝ +ె +ు] -> 0C1D+0C46+0C41 | ఝొ :: ఝొ | ఝ+ ొ0C1D +0C4A |
| 5 | [ ఘ +ె +ు] -> 0C18+0C46+0C41 | ఘొ :: ఘొ | ఘ+ ొ0C18+0C4A |

Table 9: Similarity within the script

Type 1 cases are proposed to be considered as variants. These cases are interesting in that they present no similarity in the form but have similar phonetic output. This is not unusual to find such regional variations and they are regularly used by the Telugu users. These may not cause confusion but become annoying to learners and particularly problematic to the analyzers and generators. Foregoing is the brief description of these variants followed by variants in Table 9.

## 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 in the South East Asia. Some of the characters of these scripts display formal similarity with each other. For example a number of, characters of the Kannada script are almost similar except for the flattened head stroke in Kannada whereas in Telugu it is like a tick mark on the top of the character. Out of the total there are 32 variant sets as shown in the following table.

### Type2-1 Cross-Script Variants for Telugu and Kannada

| Variant Set | Telugu | Kannada |
| --- | --- | --- |
| CP | Glyph | CP | Glyph |
| 1 | 0C02 |  ం | 0C82 | ಂ |
| 2 | 0C03 |  ః | 0C83 | ಃ |
| 3 | 0C05 |  అ | 0C85 | ಅ |
| 4 | 0C06  |  ఆ | 0C86 | ಆ |
| 5 | 0C07 |  ఇ | 0C87 | ಇ |
| 6 | 0C08 |  ఈ | 0C88 | ಈ |
| 7 | 0C10 | ఐ | 0C90 | ಐ |
| 8 | 0C12 |  ఒ | 0C92 | ಒ |
| 9 | 0C13 |  ఓ | 0C93 | ಓ |
| 10 | 0C14 |  ఔ | 0C94 | ಔ |
| 11 | 0C16 |  ఖ | 0C96 | ಖ |
| 12 | 0C17 |  గ | 0C97 | ಗ |
| 13 | 0C1C |  జ | 0C9C | ಜ |
| 14 | 0C1D |  ఝ | 0C9D | ಝ |
| 15 | 0C20 |  ఠ | 0CA0 | ಠ |
| 16 | 0C21 |  డ | 0CA1 | ಡ |
| 17 | 0C22 |  ఢ | 0CA2 | ಢ |
| 18 | 0C23 |  ణ | 0CA3 | ಣ |
| 19 | 0C25 |  థ | 0CA5 | ಥ |
| 20 | 0C26 |  ద | 0CA6 | ದ |
| 21 | 0C27 |  ధ | 0CA7 | ಧ |
| 22 | 0C28 |  న | 0CA8 | ನ |
| 23 | 0C2C | బ | 0CAC | ಬ |
| 24 | 0C2D | భ | 0CAD | ಭ |
| 25 | 0C2E | మ | 0CAE | ಮ |
| 26 | 0C2F | య | 0CAF | ಯ |
| 27 | 0C30 |  ర | 0CB0 | ರ |
| 28 | 0C32 |  ల | 0CB2 | ಲ |
| 29 | 0C33 |  ళ | 0CB3 | ಳ |
| 30 | 0CBF | ಿ | 0C3F | ి |
| 31 | 0CC1 |  ು | 0C41 | ు  |
| 32 | 0CC3 | ೃ  | 0C43 | ృ  |

Table 10: Cross-Script Variant for Telugu and Kannada

### Type2-2 Cross-Script Variants for Telugu and Devanagari

| Variant Set | Telugu | Devanagari |
| --- | --- | --- |
| CP | Glyph | CP | Glyph |
| 2 | 0C03 |  ః | 0903 | ः |

Table 11: Cross-Script Variant for Telugu and Devanagari

### Type2-3 Cross-Script Variants for Telugu and Gujarati

| Variant Set | Telugu | Gujarati |
| --- | --- | --- |
| CP | Glyph | CP | Glyph |
| 2 | 0C03 |  ః | 0A83 | ઃ |

Table 12: Cross-Script Variant for Telugu and Gujarati

### Type2-4 Cross-Script Variants for Telugu and Malayalam

| Variant Set | Telugu | Malayalam |
| --- | --- | --- |
| CP | Glyph | CP | Glyph |
| 1 | 0C02 |  ం | 0D02 | ം |
| 2 | 0C03 |  ః | 0D03 | ം |

Table 11: Cross-Script Variant for Telugu and Malayalam

### Type2-5 Cross-Script Variants for Telugu and Sinhala

| Variant Set | Telugu | Sinhala |
| --- | --- | --- |
| CP | Glyph | CP | Glyph |
| 1 | 0C02 |  ం | 0D82 | ം |
| 2 | 0C03 |  ః | 0D82 | ം |

Table 12: Cross-Script Variant for Telugu and Sinhala

[TBD –The following temporary table show the set of code points that need to be discussed by the NBGP]

| # | Telugu | Kannada | Current Views | NBGP decision |
| --- | --- | --- | --- | --- |
| 1 | ఞ (0C1E) | ಞ (0C9E) | (RAO) Green, (UBP) Red | TBD |
| 2 | ట (0C1F) | ಟ (0C9F) | (RAO) Green, (UBP) Red | TBD |

| # | Telugu | Sinhala | Current Views | NBGP decision |
| --- | --- | --- | --- | --- |
| 3 | ర (0C30) | ර් (0DBB+0DCA) | Newly Added  | TBD |

# 7. Whole Label Evaluation Rules (WLE)

In this section we provide the WLEs that are required by the language. The 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. The details of syllable formation, see Appendix B.

C → Consonant
M → Matra
V → Vowel
B → Anusvara (Bindu)
D → Candrabindu
X → Visarga

H → Halant / Virama

Rule 1. The sign for H i.e. the vowel deleting halanta must be preceded only by the C and not any thing.

Rule 2. The sign for M, i.e. the secondary vowel signs must always be preceded only by the C and not any thing.

Rule 3. X, the visarga sign must always be preceded only by the V or M.

Rule 4. B, the anusvara sign must always be preceded only by the V or M.

Rule 5. D, the chandrabindu sign must always be preceded only by the V or M.

Rule 6. The vowel V shall not be preceded by the sign H.

8. Contributors

Gangadhar Panday
Uma Maheshwara Rao

NBGP members

# 9. References

1. Disanayaka, J.B. 2017. Encyclopedia of Sinhala Language and Culture. Colombo: Sumitha Publishers. (first edition 2012).
2. Krishnamurti, Bhadriraju (Ed.). 2000. Telugu bhaashaa charitra. Hyderabad: P.S. Telugu University. (first edition 1974).
3. Krishnamurti, Bhadriraju and [J P L Gwynn](https://en.wikipedia.org/wiki/J_P_L_Gwynn). 1985.  A Grammar of Modern Telugu. New Delhi: Oxford University Press. [*ISBN*](https://en.wikipedia.org/wiki/International_Standard_Book_Number) [*978-0-19-561664-4*](https://en.wikipedia.org/wiki/Special%3ABookSources/978-0-19-561664-4).
4. Sridhar, S.N. 1980. *Kannada*. New York: Routledge.
5. Suresh, Kolichala.2012. Proposal to encode Telugu LLLA (Telugu ೞ): http:// eemaata.com/ unicode-proposal/ telugu-llla-proposal.pdf
6. Suresh, Kolichala. 2012. Divergent developments of alveolar stop \*ṯ in Telugu http:// kolichala.com/ dravidian/ Divergent\_developments\_of\_alveolar\_stop\_in\_Telugu.pdf
7. Telugu Unicode Chart, Telugu Range: 0C00–0C7F. The Unicode Standard, Version 10.0 http: //www.unicode.org/Public/ 10.0.0/ charts.
8. Uma Maheshwara Rao, G. 2012. Telugu bhaasha-saMgaNanaM. Hyderabad: P.S.Telugu University. ISBN: 81-86073-372-9.
9. Uma Maheshwara Rao, G. 2003. Standard Telugu Written Language. VIDYULLIPI-4. pp. 1-14. Hyderabad: SCIL.
10. 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.

#

# Appendix A: Code Points Similarity Analysis

## A-1. Telugu and Kannada

The following table defines Telugu and Malayalam code points which are visually similar.

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

## A-2. Telugu and Malayalam

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

## A-3. Telugu and Sinhala

Beside those identical code point defined as variants in Section 6, there is no 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 ఘ ఙ చ ఛ జ ఝ 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);

Ex. 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;

Ex. క్ ఖ్ గ్ ఘ్ ఙ్

* D= arthanusvara/chandrabindu = ఁ may occur with one of the C, V, and the combined CM to form CD, CMD, VD, and CHC([HC]\*)D;
* B= purnanusvara, the homorganic nasal and an Archiphoneme = ం, may occur with one of the C, V, and the combined CM to form CB, CMB, VB, and CH([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, and CH([HC]\*)X,

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 C-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, 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, 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, $=VD|B|X; ex.**

2a=V+D=$; అఁ ఆఁ ఇఁ ఈఁ ఉఁ ఊఁ ఎఁ ఏఁ ఐఁ ఒఁ ఓఁ ఔఁ;

2b=V+B=$; అం ఆం ఇం ఈం ఉం ఊం ఎం ఏం ఐం ఒం ఓం ఔం;

2c=V+X=$; అః ఆః ఇః ఈః ఉః ఊః ఎః ఏః ఐః ఒః ఓః ఔః;

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

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

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

 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, $=CB|D|X; Ex.**

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

3a. $=CD:  కఁ ఖఁ గఁ ఘఁ ఙఁ చఁ ఛఁ జఁ ఝఁ ఞఁ టఁ ఠఁ etc.

3b. $=CB:  కం ఖం గం ఘం ఙం చం ఛం జం ఝం ఞం టం ఠం etc.

3c. $=CX:  కః ఖః గః ఘః ఙః చః ఛః జః ఝః ఞః టః ఠః etc.

There are 3\*35=105 graphic consonant modifier syllables are possible.

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

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

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, $=CMB|D|X;**

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

కాం కిం కీం కుం కూం కెం కేం కైం కొం కోం కౌం

 కాఁ కిఁ కీఁ కుఁ కూఁ కెఁ కేఁ కైఁ కొఁ కోఁ కౌఁ

 కాః కిః కీః కుః కూః కెః కేః కైః కొః కోః కౌః

A total of 35\*12\*3 consonant plus a dependent vowel and one of the three modifiers derive 1260 possible graphic syllables in Telugu.

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

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

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

3 consonant clusters: ర్ద్ర,  ష్ట్ర, స్త్ర, న్ధ్ర, ఙ్ఘ్ర ష్ట్ల, త్ర్య, 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, $=CH(CH[CH])CM;**

Any consonant followed by the halanta 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; Ex.

క్కు ఖ్ఖి గ్గొ, ఘ్ఘి, ఙ్ఙా, చ్చా, ఛ్ఛూ,  జ్జే, ఝ్ఝె, ఞ్ఞో, ట్టీ, ఠ్ఠూ, డ్డా, ఢ్ఢూ, ణ్ణా, etc.

ర్ద్రు,  ష్ట్రి, స్త్రీ, న్ధ్రి, ఙ్ఘ్రి ష్ట్లా, త్ర్యా, త్స్ను  etc.

త్స్న్యా

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|D|X])|CM([B|D|X])|CH(CH[C])M([B|D|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|VD|VX|

C|CB|CD|CX|CM|CH|

CHC|CHCB|CHCD|CHCX|CHCM| CHCH|CHCHC|CHCHCB|CHCHCD|CHCHCX|CHCHCM

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

1. Others are the EGIDS 5 languages, listed in Table 1: Main languages considered under Telugu LGR [↑](#footnote-ref-1)