LGR Proposal Telugu 2017 Sep 30

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

*LGR Version: 2.0*

*Date:* 30th Sep 2017

*Document version:* 1.1

*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-Deva-20170323.xml". – delete for now]

# 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: to add

Native name of the script: తెలుగు

Maximal Starting Repertoire [MSR] version: 2

# Background on Script and Principal Languages Using It

The script called Telugu is written from left to right. Historically it derives from the Brahmi alphabet of the Ashokan inscriptions. Telugu is one of the 22 scheduled languages of India. The script is also used for the 8 Languages *viz.,* Gondi, Banjara, Konda, Lambadi, Sora [Shabri community] Kui, Yerukala and also Sanskrit [for publishing books]. Telugu is closely associated with Kannada.

## The Evolution of the Script

The origins of the Telugu alphabet can be traced by to the [Brahmi](http://www.omniglot.com/writing/brahmi.htm) alphabet of ancient India, which developed into an alphabet used for both Telugu and Kannada, which in turn split into two separate alphabets between the 12th and 15th centuries AD.

The earliest known inscriptions containing Telugu words appear on coins 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 an archaic style very different from the everyday spoken language. During the the second half of the 20th century, a new written standard emerged based on the modern spoken language. In 2008 Telugu was designated as a classical language by the Indian government.

Notable features

* Type of writing system: syllabic alphabet in which all consonants have an inherent vowel. Diacritics, which can appear above, below, before or after the consonant they belong to, are used to change the inherent vowel.
* When they appear the the beginning of a syllable, vowels are written as independent letters.
* When certain consonants occur together, special conjunct symbols are used which combine the essential parts of each letter.
* Direction of writing: left to right in horizontal lines

**Telugu (తెలుగు),** a Dravidian language spoken by about 75 million people mainly in the southern Indian state of Andhra Pradesh, where it is the official language. It is also spoken in such neighbouring 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, Malaysia, Mauritius, Myanmar and Réunion.

Table : Main languages considered under Telugu LGR

## Gondi, Banjara, Konda, Lambadi, Sora [Shabri community] Kui, Yerukala and also Sanskrit [publishing books]

## The structure of written Telugu

The writing system of Telugu could be summed up as composed of the following:

### The Consonants

Telugu consonants have an implicit schwa /ə/ included in them. As per traditional classification they are categorized according to their phonetic properties. There are 5 Varga groups (classes) and one non-Varga group. Each Varga, which corresponds to Stops, contains five consonants classified as per their properties. The first four consonants are classified on the basis of Voicing and Aspiration and the last is the corresponding nasal.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Varga** | **Unvoiced** | | **Voiced** | | **Nasal** |
|  | -Asp | +Asp | -Asp | +Asp |  |
| **Velar** | క | ఖ | గ | ఘ | ఙ |
| **Palatal** | చ | ఛ | జ | ఝ | ఞ |
| **Retroflex** | ట | ఠ | డ | ఢ | ణ |
| **Dental** | త | థ | ద | థ | న |
| **Bi-labial** | ప | ఫ | బ | భ | మ |

Table : Varga classification of consonants

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

Table : Non-Varga consonants

Did upto this on 30 Sep 2017.

Then decided to do only Code Point repertoire, hence skip to there.

### The Implicit Vowel Killer: Halanta[[1]](#footnote-1)

All consonants have an implicit vowel sign (schwa) within them. A special sign is needed to denote that this implicit vowel is stripped off. This is known as the Halanta (्). The Halanta thus joins two consonants and creates conjuncts, which can be generally from 2 to 4 consonant combinations. In rare cases it can join up to 5 consonants. However the notion of maximum number of consonants joining to form one akshar is not empirical. It is just an observation drawn from the words that have been observed till date. Given the confluence of languages happening in the Internet age, the possibility that one may want a generic Top Level Domain [gTLD] which may have more than the observed maximum cannot be ruled out. Hence, in the LGR work, this limit will not be enforced[[2]](#footnote-2).

### Vowels

Separate symbols exist for all Vowels, which are pronounced independently either at the beginning or after a vowel sound. To indicate a Vowel sound other than the implicit one, a Vowel modifier (Matra) is attached to the consonant. Since the consonant has a built in schwa, there are equivalent Matras for all vowels excepting the अ.

The correlation is shown as under:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| अ | आ | इ | ई | उ | ऊ | ऋ | ए | ऐ | ओ | औ |
|  | ा | ि | ी | ु | ू | ृ | े | ै | ो | ौ |

Table : Vowels with corresponding Matras

In addition to show sounds borrowed from English, some languages using Telugu such as Hindi, Marathi, and Konkani also admit 2 vowels and their corresponding Matras as in

ऍ ऑ

ऍण्ड /and/ ऑर /or/

Marathi replaces the ऍ by ॲ.

### The Anusvara (ं)

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

सन्त vs. संत /sənt/ saint चम्पा vs. चंपा /tʃəmpa/

### Nasalization: Chandrabindu (ँ)

Chandrabindu/Anunasika denotes nasalization of the preceding vowel as in आँख (eye) /ãkh/ eye. Present-day Hindi users tend to replace the chandrabindu by the anusvara

### Nukta (़)

Mainly used in Hindi, the nukta sign is placed below a certain number of consonants to represent words borrowed from Perso-Arabic. It can be adjoined to क ख ग ज फ to show that words having these consonants with a nukta are to be pronounced in the Perso-Arabic style.

e.g. फ़िरोज़ /firoz/

It is also placed under ड and ढ in Hindi to indicate flapped sounds

बढ़ /bədh/

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

### Visarga (ः) and Avagraha (ऽ)

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

The Avagraha (ऽ) creates an extra stress on the preceding vowel and is used in Sanskrit texts. It is rarely used in other languages using Telugu. In case of LGR, the Avagraha is not part of the repertoire as it is barred in the Maximal Starting Repertoire.

# Overall Development Process and Methodology

Under the Neo-Brahmi Generation Panel, there are many different scripts belonging to separate Unicode blocks. Each of these scripts will be assigned a separate LGR; however Neo-Brahmi GP will ensure that the fundamental philosophy behind building those LGRs are all in sync with all other Brahmi derived scripts. This is the Telugu LGR, which caters to multiple languages written using Telugu belonging to EGIDS scale 1 to 4.

## 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.

The main principle is that of Acknowledgement to Environmental Limitations. These comprise protocols or standards. All further principles are in fact subsumed under these limitations but have been spelt out separately for the sake of clarity.

### Acknowledgement to Environment Limitations:

The code point repertoire for root zone being a very special case, up the ladder in the protocol hierarchies, the canvas of available characters for selection as a part of the Root Zone code point repertoire is already constrained by various protocol layers beneath it. Following three main protocols/standards act as successive filters:

*i. The Unicode Chart:*

Out of all the characters that are needed by the given script, if the character in question is not encoded in Unicode, it cannot be incorporated in the code point repertoire. Such cases are quite rare, given the elaborate and exhaustive character inclusion efforts made by Unicode consortium.

*ii. IDNA Protocol:*

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

Example: Telugu Letter Qa (क़) is not allowed to be a part of domain name. Its decomposed form, i.e. Telugu Letter Ka followed by Telugu Sign Nukta (क+़) can be used instead.

*iii. Maximal Starting Repertoire:*

The Root-zone LGR being a repertoire of the characters which are going to be used for creation of the root zone TLDs, which in turn are an even more specialized case of domain names, the ROOT LGR procedure introduces additional exclusions on IDNA allowed set of characters.

Example: Telugu Sign Avagraha (ऽ) even if allowed by IDNA protocol, is not permitted in the Root Zone Repertoire as per the MSR.

To sum up, the restrictions start off with admitting only such characters as are part of the code-block of the given script/language. This is further narrowed down by the IDNA Protocol and finally an additional filter in the form of Maximal Starting Repertoire restricts the character set associated with the given language even more.

### No Punctuation Marks:

The TLDs being identifiers, punctuation markers present in brahmi based languages such as Danda (। ) and double Danda ( ॥ ) will not be included.

### No Symbols and Abbreviations:

Abbreviations, weights and measures and other such iconic characters like Isshar (৺), Abbreviation sign ( ॰ ) etc. will not be included.

### No Rare and Obsolete Characters:

There are characters which have been added to Unicode to accommodate rare forms especially like TELUGU LETTER VOCALIC RR (ॠ) and TELUGU LETTER VOCALIC LL (ॡ) as well as their matra forms (ॄ) and (ॣ). All such characters will not be included. This is in consonance with the Letter principle as laid down in the Root Zone LGR procedure.

### No Stress Markers of Classical Sanskrit and Vedic:

Stress markers for classical Sanskrit e.g. (॑) TELUGU STRESS SIGN UDATTA and (॒) TELUGU STRESS SIGN ANUDATTA will not be included. This is also in consonance with the Letter principle as laid down in the Root Zone LGR procedure.

# Repertoire

This section details the code-point repertoire that the Neo-Brahmi Generation Panel [NBGP] proposes to be included in the Telugu LGR.

One of the major sources of reference to the justification for inclusion of the code-point is the Indian National Standard 'Enhanced Inscript Keyboard layouts' [INSCRIPT]" laying down the language specific keyboard layouts for all the scheduled languages of India. It is officially published and notified in the Gazette of India. The standard specifies key-layouts for each of the scheduled languages of India. The standard among other things provides a comprehensive language-wise list of various characters as used by the scheduled (of which the set of languages under the ambit of the NBGP is a sub-set) languages of the India. The [INSCRIPT] standard carves out a sub-set of characters applicable to each of the languages out of the respective code-page of the script used by that language.

## Code Point Repertoire:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Sr. No.** | **Unicode Code Point** | **Character** | **Character Name** | **Unicode General Category (gc)** | **Indic Syllabic Category** | **Reference** |
| 1. | 0901 | ँ | TELUGU SIGN CANDRABINDU | Mn | Chandrabindu | [INSCRIPT] |
| 2. | 0902 | ं | TELUGU SIGN ANUSVARA | Mn | Anusvara (Bindu) | [INSCRIPT] |
| 3. | 0903 | ः | TELUGU SIGN VISARGA | Mc | Visarga | [INSCRIPT] |
| 4. | 0905 | अ | TELUGU LETTER A | Lo | Vowel | [INSCRIPT] |
| 5. | 0906 | आ | TELUGU LETTER AA | Lo | Vowel | [INSCRIPT] |
| 6. | 0907 | इ | TELUGU LETTER I | Lo | Vowel | [INSCRIPT] |
| 7. | 0908 | ई | TELUGU LETTER II | Lo | Vowel | [INSCRIPT] |
| 8. | 0909 | उ | TELUGU LETTER U | Lo | Vowel | [INSCRIPT] |
| 9. | 090A | ऊ | TELUGU LETTER UU | Lo | Vowel | [INSCRIPT] |
| 10. | 090B | ऋ | TELUGU LETTER VOCALIC R | Lo | Vowel | [INSCRIPT] |
| 11. | 090D | ऍ | TELUGU LETTER CANDRA E | Lo | Vowel | [INSCRIPT] |
| 12. | 090F | ए | TELUGU LETTER E | Lo | Vowel | [INSCRIPT] |
| 13. | 0910 | ऐ | TELUGU LETTER AI | Lo | Vowel | [INSCRIPT] |
| 14. | 0911 | ऑ | TELUGU LETTER CANDRA O | Lo | Vowel | [INSCRIPT] |
| 15. | 0913 | ओ | TELUGU LETTER O | Lo | Vowel | [INSCRIPT] |
| 16. | 0914 | औ | TELUGU LETTER AU | Lo | Vowel | [INSCRIPT] |
| 17. | 0915 | क | TELUGU LETTER KA | Lo | Consonant | [INSCRIPT] |
| 18. | 0916 | ख | TELUGU LETTER KHA | Lo | Consonant | [INSCRIPT] |
| 19. | 0917 | ग | TELUGU LETTER GA | Lo | Consonant | [INSCRIPT] |
| 20. | 0918 | घ | TELUGU LETTER GHA | Lo | Consonant | [INSCRIPT] |
| 21. | 0919 | ङ | TELUGU LETTER NGA | Lo | Consonant | [INSCRIPT] |
| 22. | 091A | च | TELUGU LETTER CA | Lo | Consonant | [INSCRIPT] |
| 23. | 091B | छ | TELUGU LETTER CHA | Lo | Consonant | [INSCRIPT] |
| 24. | 091C | ज | TELUGU LETTER JA | Lo | Consonant | [INSCRIPT] |
| 25. | 091D | झ | TELUGU LETTER JHA | Lo | Consonant | [INSCRIPT] |
| 26. | 091E | ञ | TELUGU LETTER NYA | Lo | Consonant | [INSCRIPT] |
| 27. | 091F | ट | TELUGU LETTER TTA | Lo | Consonant | [INSCRIPT] |
| 28. | 0920 | ठ | TELUGU LETTER TTHA | Lo | Consonant | [INSCRIPT] |
| 29. | 0921 | ड | TELUGU LETTER DDA | Lo | Consonant | [INSCRIPT] |
| 30. | 0922 | ढ | TELUGU LETTER DDHA | Lo | Consonant | [INSCRIPT] |
| 31. | 0923 | ण | TELUGU LETTER NNA | Lo | Consonant | [INSCRIPT] |
| 32. | 0924 | त | TELUGU LETTER TA | Lo | Consonant | [INSCRIPT] |
| 33. | 0925 | थ | TELUGU LETTER THA | Lo | Consonant | [INSCRIPT] |
| 34. | 0926 | द | TELUGU LETTER DA | Lo | Consonant | [INSCRIPT] |
| 35. | 0927 | ध | TELUGU LETTER DHA | Lo | Consonant | [INSCRIPT] |
| 36. | 0928 | न | TELUGU LETTER NA | Lo | Consonant | [INSCRIPT] |
| 37. | 092A | प | TELUGU LETTER PA | Lo | Consonant | [INSCRIPT] |
| 38. | 092B | फ | TELUGU LETTER PHA | Lo | Consonant | [INSCRIPT] |
| 39. | 092C | ब | TELUGU LETTER BA | Lo | Consonant | [INSCRIPT] |
| 40. | 092D | भ | TELUGU LETTER BHA | Lo | Consonant | [INSCRIPT] |
| 41. | 092E | म | TELUGU LETTER MA | Lo | Consonant | [INSCRIPT] |
| 42. | 092F | य | TELUGU LETTER YA | Lo | Consonant | [INSCRIPT] |
| 43. | 0930 | र | TELUGU LETTER RA | Lo | Consonant | [INSCRIPT] |
| 44. | 0932 | ल | TELUGU LETTER LA | Lo | Consonant | [INSCRIPT] |
| 45. | 0933 | ळ | TELUGU LETTER LLA | Lo | Consonant | [INSCRIPT] |
| 46. | 0935 | व | TELUGU LETTER VA | Lo | Consonant | [INSCRIPT] |
| 47. | 0936 | श | TELUGU LETTER SHA | Lo | Consonant | [INSCRIPT] |
| 48. | 0937 | ष | TELUGU LETTER SSA | Lo | Consonant | [INSCRIPT] |
| 49. | 0938 | स | TELUGU LETTER SA | Lo | Consonant | [INSCRIPT] |
| 50. | 0939 | ह | TELUGU LETTER HA | Lo | Consonant | [INSCRIPT] |
| 51. | 093A | ऺ | TELUGU VOWEL SIGN OE | Mn | Matra | [INSCRIPT] |
| 52. | 093B | ऻ | TELUGU VOWEL SIGN OOE | Mc | Matra | [INSCRIPT] |
| 53. | 093C | ़ | TELUGU SIGN NUKTA | Mn | Nukta | [INSCRIPT] |
| 54. | 093E | ा | TELUGU VOWEL SIGN AA | Mc | Matra | [INSCRIPT] |
| 55. | 093F | ि | TELUGU VOWEL SIGN I | Mc | Matra | [INSCRIPT] |
| 56. | 0940 | ी | TELUGU VOWEL SIGN II | Mc | Matra | [INSCRIPT] |
| 57. | 0941 | ु | TELUGU VOWEL SIGN U | Mn | Matra | [INSCRIPT] |
| 58. | 0942 | ू | TELUGU VOWEL SIGN UU | Mn | Matra | [INSCRIPT] |
| 59. | 0943 | ृ | TELUGU VOWEL SIGN VOCALIC R | Mn | Matra | [INSCRIPT] |
| 60. | 0944 | ॄ | TELUGU VOWEL SIGN VOCALIC RR | Mn | Matra | [INSCRIPT] |
| 61. | 0945 | ॅ | TELUGU VOWEL SIGN CANDRA E = candra | Mn | Matra | [INSCRIPT] |
| 62. | 0947 | े | TELUGU VOWEL SIGN E | Mn | Matra | [INSCRIPT] |
| 63. | 0948 | ै | TELUGU VOWEL SIGN AI | Mn | Matra | [INSCRIPT] |
| 64. | 0949 | ॉ | TELUGU VOWEL SIGN CANDRA O | Mc | Matra | [INSCRIPT] |
| 65. | 094B | ो | TELUGU VOWEL SIGN O | Mc | Matra | [INSCRIPT] |
| 66. | 094C | ौ | TELUGU VOWEL SIGN AU | Mc | Matra | [INSCRIPT] |
| 67. | 094D | ् | TELUGU SIGN VIRAMA | Mn | Halant / Virama | [INSCRIPT] |
| 68. | 094F | ॏ | TELUGU VOWEL SIGN AW | Mc | Matra | [INSCRIPT] |
| 69. | 0956 | ॖ | TELUGU VOWEL SIGN UE | Mn | Matra | [INSCRIPT] |
| 70. | 0957 | ॗ | TELUGU VOWEL SIGN UUE | Mn | Matra | [INSCRIPT] |
| 71. | 0972 | ॲ | TELUGU LETTER CANDRA A | Lo | Consonant | [INSCRIPT] |
| 72. | 0973 | ॳ | TELUGU LETTER OE | Lo | Consonant | [INSCRIPT] |
| 73. | 0974 | ॴ | TELUGU LETTER OOE | Lo | Consonant | [INSCRIPT] |
| 74. | 0975 | ॵ | TELUGU LETTER AW | Lo | Consonant | [INSCRIPT] |
| 75. | 0976 | ॶ | TELUGU LETTER UE | Lo | Consonant | [INSCRIPT] |
| 76. | 0977 | ॷ | TELUGU LETTER UUE | Lo | Consonant | [INSCRIPT] |
| 77. | 0979 | ॹ | TELUGU LETTER ZHA | Lo | Consonant | [INSCRIPT] |
| 78. | 097A | ॺ | TELUGU LETTER HEAVY YA | Lo | Consonant | [INSCRIPT] |
| 79. | 097B | ॻ | TELUGU LETTER GGA | Lo | Consonant | [INSCRIPT] |
| 80. | 097C | ॼ | TELUGU LETTER JJA | Lo | Consonant | [INSCRIPT] |
| 81. | 097E | ॾ | TELUGU LETTER DDDA | Lo | Consonant | [INSCRIPT] |
| 82. | 097F | ॿ | TELUGU LETTER BBA | Lo | Consonant | [INSCRIPT] |

Table : Code point repertoire

Apart from the above individual code-points, the Neo-Brahmi Generation Panel also proposes some specific sequences which enable conditional inclusion of the "TELUGU LETTER RRA" in the repertoire.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Sr. No.** | **Unicode Code Points** | **Sequence** | **Character Names** | **Unicode General Category (gc)** | **Reference** |
| 1. | 0931  094D  092F | ऱ्य | TELUGU LETTER RRA  TELUGU SIGN VIRAMA  TELUGU LETTER YA | Lo  Mn  Lo | [INSCRIPT] |
| 2. | 0931  094D  0939 | ऱ्ह | TELUGU LETTER RRA  TELUGU SIGN VIRAMA  TELUGU LETTER HA | Lo  Mn  Lo | [INSCRIPT] |

Table : Sequences

## Structural Formation of Telugu:

All the languages written in Brahmi derived scripts follow a particular way of formation of its words, known as "akshar". In the next section there are detailed akshar formation rules as applicable to representation of "Hindi" language when written in Telugu Script. These rules need slight changes for different languages written in Telugu in terms of

- Character addition/deletion (e.g. Nukta [U+093C] character is applicable for Hindi but not Marathi)

- Presence or absence of a particular rule (e.g. Eyelash Ra construct is required in Marathi, Konkani and Nepali but not in Hindi).

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

## Akshar formation rules for Hindi:

This section details the Akshar formation rules as applicable to Hindi. The first section lists the categories of the characters in the form of variables. In the rules, instead of their descriptive names, the variable names are used. The second section lists four operators along with their functions which are assumed while specifying the rules. The following two sections describe the two major categories of the Akshar formations first of which begins with the vowels and the second one with the consonants.

### Variables involved

Dash → Hyphen -

Digit → Indo-Arabic digits [0-9]

C → Consonant

M → Matra

V → Vowel

B → Anusvara (Bindu)

D → Chandrabindu (Anunasika)

X → Visarga

H → Halant / Virama

N → Nukta

### Operators used:

|  |  |
| --- | --- |
| **Symbol** | **Function** |
| | | Alternative |
| [ ] | Optional |
| \* | Variable Repetition |
| ( ) | Sequence Group |

Table : Symbol functions

In what follows, the Vowel Sequence and the Consonant Sequence pertinent to Telugu, when used to write Hindi, are given.

### The Vowel Sequence

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

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

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

Examples:

|  |  |  |  |
| --- | --- | --- | --- |
| **Sequence Description** | **Sequence** | **Example** | **Example**  **Decomposition** |
| Vowel | V | अ /a/ |  |
| Vowel + Anusvara | V[D] | अं /aṁ/ | अ ं |
| Vowel + Chandrabindu | V[B] | अँ /aṃ/ | अ ँ |
| Vowel + Visarga | V[X] | अः /aḥ/ | अ ः |

Table

### Consonant Sequence

A consonant sequence begins with a consonant. It may be optionally followed by a Nukta (N), Matra (M), Anusvara (B), Chandrabindu (D), Visarga (X) or a Halanta (H). The number of instances of these characters occurring after a consonant is restricted to one. There is a possibility of further extension of the Consonant sequence after the N, M and H. Each of these has been discussed in the following sections:

1. A single consonant (C)

(The consonant shall be treated as coterminous with the Consonant along with the Nukta sign wherever such a case is pertinent.)

Examples:

|  |  |  |  |
| --- | --- | --- | --- |
| **Sequence Description** | **Sequence** | **Example** | **Example**  **Decomposition** |
| Consonant | C | क /ka/ |  |
| Consonant + Nukta | C[N] | क़ /ḳa/ | क ़ |

Table

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

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

Examples:

|  |  |  |  |
| --- | --- | --- | --- |
| **Sequence Description** | **Sequence** | **Example** | **Example**  **Decomposition** |
| Consonant + Matra | C[M] | कि /ki/ | क ि |
| Consonant + Anusvara | C[B] | कं /kaṁ/ | क ं |
| Consonant + Chandrabindu | C[D] | कँ /kaṃ/ | क ँ |
| Consonant + Visarga | C[X] | कः /kaḥ/ | क ः |
| Consonant + Halanta | C[H] | क् /k/  (Pure Consonant) | क ् |

Table

2. A. A CM sequence can be optionally followed by D, B or X

(CM)[D|B|X]

Example:

|  |  |  |  |
| --- | --- | --- | --- |
| **Sequence Description** | **Sequence** | **Example** | **Example**  **Decomposition** |
| Consonant + Matra + Anusvara | CM[B] | कीं /kīṁ/ | क ी ं |
| Consonant + Matra + Chandrabindu | CM[D] | काँ /kāṃ/ | क ा ँ |
| Consonant + Matra + Visarga | CM[X] | कीः /kīḥ/ | क ी ः |

Table

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

Example:

|  |  |  |  |
| --- | --- | --- | --- |
| **Sequence Description** | **Sequence** | **Example** | **Example**  **Decomposition** |
| Consonant + Halanta + Consonant + Halanta + Consonant + Halanta + Consonant | CHCHCHC | न्क्र्य /nkrya/ | न ् क ् र ् य |

Table

**Subsets:**

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

Example:

|  |  |  |  |
| --- | --- | --- | --- |
| **Sequence Description** | **Sequence** | **Example** | **Example**  **Decomposition** |
| Consonant + Halanta + Consonant + Matra | CHC[M] | क्की /kkī/ | क ् क ी |
| Consonant + Halanta + Consonant + Anusvara | CHC[B] | क्कं /kkaṁ/ | क ् क ं |
| Consonant + Halanta + Consonant + Chandrabindu | CHC[D] | क्कँ /kkaṃ/ | क ् क ँ |
| Consonant + Halanta + Consonant + Visarga | CHC[X] | क्कः /kkaḥ/ | क ् क ः |

Table

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

Example:

|  |  |  |  |
| --- | --- | --- | --- |
| **Sequence Description** | **Sequence** | **Example** | **Example**  **Decomposition** |
| Consonant + Halanta + Consonant + Matra + Anusvara | CHCM[B] | क्कीं /kkīṁ/ | क ् क ी ं |
| Consonant + Halanta + Consonant + Matra + Chandrabindu | CHCM[D] | क्कीँ /kkīṃ/ | क ् क ी ँ |
| Consonant + Halanta + Consonant + Matra + Visarga | CHCM[X] | क्कीः /kkīḥ/ | क ् क ी ः |

Table

These are the basic akshar rules on which the overall Telugu LGR is based. As languages other than Hindi are considered, some language specific characters and rules are introduced. There are some additional finer aspects to these rules as one takes into account the digits, punctuations and special standalone characters like Avagraha. Those aspects are not discussed here as the MSR on which the LGRs are supposed to be based, excludes those characters.

# Variants

There are no characters/character sequences in Telugu which can be created by using the characters permitted as per the [MSR] and look exactly alike.

However, Santhali language has a unique requirement for Nukta character (़ U+093C ) positioning which is not common in other Telugu based languages. Santhali requires the Nukta character to be followed after certain Vowels and Matras. Complete representation of these Santhali combinations necessitated the Whole Label Evaluation rules (given in the ‎7) to be opened up for these specific cases. A regular non-Santhali user mostly cannot even anticipate possibility of such a combination and can mistake it for something else.

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

|  |  |
| --- | --- |
| **Variant 1** | **Variant 2** |
| आ  U+0906 | आ़  U+0906 U+093C |
| ओ  U+0913 | ओ़  U+0913 U+093C |
| ा  U+093E | ा़  U+093E U+093C |
| ो  U+094B | ो़  U+094B U+093C |

# Whole Label Evaluation Rules (WLE)

This section provides the WLEs that are required by all the languages mentioned in section ‎3.2 when written in Telugu Script. The rules have been drafted in such a way that they can be easily translated into the LGR specification.

Below are the symbols used in the WLE rules, for each of the "Indic Syllabic Category" as mentioned in the .

|  |  |  |
| --- | --- | --- |
| C | → | Consonant |
| M | → | Matra |
| V | → | Vowel |
| B | → | Anusvara (Bindu) |
| D | → | Chandrabindu |
| X | → | Visarga |
| H | → | Halant / Virama |
| N | → | Nukta |
| S | → | Eyelash Reph (C1HC2)  where  C1 is 0931 (ऱ - TELUGU LETTER RRA)  H is 094D (् - TELUGU SIGN VIRAMA)  C2 is either - 092F (य - TELUGU LETTER YA)  or 0939 (ह - TELUGU LETTER HA) |

Below are the specific WLE rules:

1. N: must be preceded only by either of specific set of Cs, Vs and Ms

The specific Cs are:

* 1. क (U+0915)
  2. ख (U+0916)
  3. ग (U+0917)
  4. ज (U+091C)
  5. ड (U+0921)
  6. ढ (U+0922)
  7. फ (U+092B)

The specific Vs are:

1. आ (U+0906) (Required in Santhali language)
2. ओ (U+0913) (Required in Santhali language)

The specific Ms are:

1. ा (U+093E) (Required in Santhali language)
2. ो (U+094B) (Required in Santhali language)
3. H: must be preceded by C or CN
4. X: must be preceded by either of V, C, N or M
5. B: must be preceded by either of V, C, N or M
6. D: must be preceded by either of V, C, N or M
7. M: must be preceded either by C or CN
8. V: Can **NOT** be preceded by H

**Case of Eyelash Reph:**

In the WLE rules, there is no specific mention of the Eyelash Reph for two reasons:

1. As the U+0931 is added as a part of permissible sequences in , it gets permitted only with the specific sequences.
2. The last characters of both the sequences of which the U+0931 is part, are consonants. As the Eyelash-Reph can take all the combinations as that of a consonant, no specific handling in terms of context rule is required.

# Contributors

Neo-Brahmi Generation Panel members.

# References

[MSR] Maximal Starting Repertoire

[INSCRIPT] Bureau of Indian Standards (BIS), "Enhanced Inscript Keyboard layouts" (IS 16350: 2016)

*<This is a paid resource managed by Bureau of Indian Standards. NBGP will try to get a copy of the same and then share the same with IP>*

[NBGP] Neo-Brahmi Generation Panel

Appendix

http://www.omniglot.com/writing/telugu.htm

1. Unicode (cf. Unicode 3.0 and above) prefers the term Virama. In this report both the terms have been used to denote the character that suppresses the inherent vowel. [↑](#footnote-ref-1)
2. This can be the case when a foreign language word, which admits a large number of consonants, is transliterated into Devanāgarī [↑](#footnote-ref-2)