L2/19-401

Proposal to Encode Telugu Sign Nukta

Vinodh Rajan vinodh@virtualvinodh.com
Shriramana Sharma jamadagni@gmail.com
Suresh Kolichala suresh.kolichala@gmail.com

This document proposes the encoding of the Telugu Sign Nukta in the Telugu block of the UCS.

1. Introduction

Nukta is a consonantal diacritic used in Indic scripts to extend the native character repertoire and denote non-native phonemes. It has been traditionally used to represent Perso-Arabic phonemes, and lately, English phonemes (particularly, /f/ and /z/). Many of the encoded Indic scripts including a Nukta-like character. These include historic scripts like Grantha and Siddham, where Nukta is a modern innovation.

While Nukta is a common feature of North Indic scripts (due to significant lexical borrowing from Persian and Arabic), Kannada is the only South Indic script to have a wide-spread use of Nukta to represent /f/ and /z/. In Tamil, ஃ U+0B83 seemingly takes a Nukta-like role to represent those phonemes by prepending itself to consonants, as in ஃப /f/ and ஃஜ /z/ respectively. It would appear that among the major South Indic scripts only Telugu and Malayalam do not have any characters to fulfill the role of Nukta.

2. Nukta Proposals in Telugu

Desikacharyulu (2000?) while elaborating the design of one of the earliest Telugu Unicode fonts Pothana2000 laments the lack of Nukta in the UCS and makes the following statements:

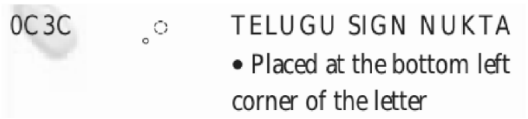
"[…] Nukta are not yet defined in the Telugu Unicode range, which they must be […]"

"[…] it is still a good idea to have Nukta in Telugu to provide extendability [sic] of the script to foreign sounds"

He, therefore, included the Devanagari Nukta in his font as a placeholder, fully aware that it won't be rendered properly and hoping "*it will be supported in the future*".

One specific Nukta character has been quasi-proposed in the past for Telugu. A Government of India (GoI) publication (Rao, U.G, 2002) refers to a Nukta to extend the character set. However, there seems to be no evidence to its actual usage apart from its inclusion in the publication and a (passing) presentation to the Unicode consortium. There was no further formal follow up from the Government of India regarding the character[1].

[1] https://unicode.org/mail-arch/unicode-ml/y2009-m09/0006.html

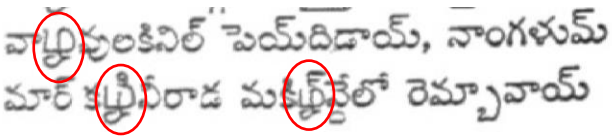


Proposed Nukta from GoI (Rao, U.G, 2002)

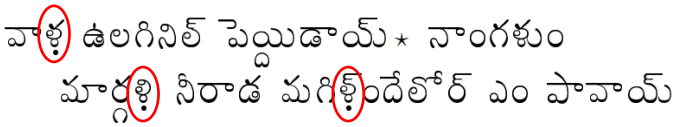
3. Actual Nukta Usage in Telugu

Since Tamil is one of the liturgical languages of Vaishnavism in the South, Tamil texts have long been rendered in the Telugu script for the purpose of recitation in the Telugu-speaking areas of Southern India. However, Modern Telugu lacks an equivalent of the Tamil Consonant LLLA ழ and couldn't faithfully represent the Tamil texts. This was overcome in several ways ranging from importing the Tamil character and treating it like a Telugu character and, quite relevant to our context, using a Nukta alongside Telugu LLA. (There is a Telugu LLLA ఴ U+0C34 that was encoded recently but it is an archaic character that was not known beyond epigraphic/historical circles and the common public was/is oblivious to its existence).

The below is an example from the Tamil text Tiruppāvai using both the conventions.

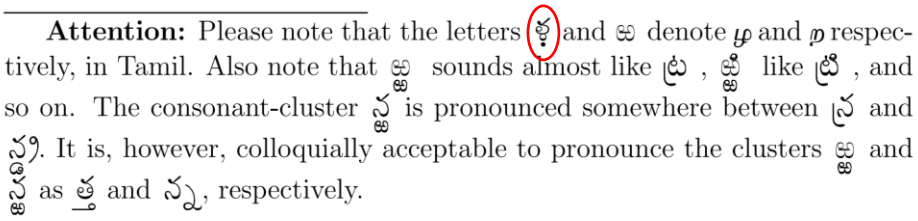


ழ as a Telugu letter (Link to the text)



ళ with Nukta as ళ (Kidambi, n.d-a)

Kidambi (n.d-a) clearly informs the readers on the Nukta convention used in the beginning of the document.



Attention: Please note that the letters ళ and ఱ denote ழ and ற respectively, in Tamil. Also note that ట్ర sounds almost like ట్రు , ట్రి like ట్రీ , and so on. The consonant-cluster న్ఱ is pronounced somewhere between న్ర and న్ఱ్ర. It is, however, colloquially acceptable to pronounce the clusters ట్ఱ and ట్ఱ్ర as ట్ర and న్ఱ, respectively.

From Kidambi, S (n.d-a)

వాళ్ఞాట్ పట్టు నిన్నీరుళ్ళిరేల్⋆ వందు మణ్ణుం మణముం కొణ్మిన్⋆
కూళ్ఞాట్ పట్టు నిన్నీర్గళై⋆ ఎంగళ్ కుళ్ళువినిల్ పుగుదలొట్టోం⋆
ఏళ్ఞాట్ కాలుం పళ్ళిప్పిలోం నాంగళ్⋆ ఇరాక్కదర్వాళ్⋆ ఇలంగై–
పాళ్ఞాగ ప్పడై పొరుదానుక్కు⋆ ప్పల్లాండు కూఱుదుమే (3)

Tiruppallāṇṭu *(Kidambi, S, n.d-b)*

வா**ழ**ாட்பட்டுநின்றீருள்ளீரேல் வந்துமண்ணும் மணமும் கொண்மின்
கூ**ழ**ாட்பட்டுநின்றீர்களை எங்கள் கு**ழு**வினில் புகுதலொட்டோம்
ஏ**ழ**ாட்காலும் ப**ழி**ப்பிலோம் நாங்களிராக்கதர்வா**ழ்** இலங்கை
பா**ழ**ாளாக ப்படைபொருதானுக்குப் பல்லாண்டு கூறுதுமே

Corresponding Tamil text.

It is the only case (we are aware of), where a dot-like Nukta sign is actually attested alongside a Telugu character. This form of Nukta as a dot and the GoI's Nukta as a small circle can very well be considered as glyphic variants. This is similar to the appearance of Virama in Tamil, which can appear as a dot or as a circle based on stylistic preferences. The text repository Prapatti: http://prapatti.com/slokas/slokasbyname.html contains numerous Telugu documents using the Nukta sign (for this purpose).

Based on the actual attestation and its ability to impart extendibility to the Telugu script, we propose to include a Telugu Sign Nukta in the UCS.

4. Possible Objections

The attested LLA + dot below ళ్ are not glyphic variants of the already encoded Telugu LLLA ఌ but rather individual characters with different graphemic identities. The latter is an archaic character unknown to the majority of the populace and the former is a clear attempt to extend the alphabet and represent a non-native phoneme in a modern context. Orthographically speaking, both are distinct practices and users must be given a choice between the two, rather making it haphazardly font dependent.

It could also be said there are not any further attestations. The usage of Nukta has a consistent albeit restricted practice in at least one specific religious context/group to be considered as minority usage. Even in Siddham, the Nukta character was encoded based one modern usage from a single site committed to writing Japanese in Siddham, and in Grantha, it was entirely a modern innovation specifically requested by the user community. Also, as noted in the previous section, there is a user community that would readily consume the Nukta character if encoded.

Encoding a Telugu Nukta would especially enable the transliteration of Perso-Arabic consonants such as *qa kha ġa za fa wa* etc, the North Indic *ṛa* and *ṛha*, the South Indic *ṉa*, the East Indic ya and so on when used in combination with appropriate existing characters. In a more modern context, it would especially allow Telugu to represent *fa* and *za* unambiguously.

5. Suggested Transliteration Mapping

Based on the generic Indic model, we provide below suggested mappings to the existing Indic characters in Telugu using the proposed Nukta. Kannada is also shown as an existing cognate equivalent.

ISO 15919	qa	kha	ġa	za	ṛa	ṛha	fa
Devanagari	क़	ख़	ग़	ज़	ड़	ढ़	फ़
Telugu	క్	ఖ్	గ్	జ్	డ్	ఢ్	ఫ్
Kannada	ಕ್	ಖ್	ಗ್	ಜ್	ಡ್	ಢ್	ಫ್

ISO 15919	ẏa
Devanagari	य़
Bengali	য়
Oriya	ୟ
Telugu	య్
Kannada	ಯ್

ISO 15919	l̤a	n̤a
Devanagari	ऴ	ऩ
Tamil	ழ	ன
Malayalam	ഴ	ൻ
Telugu	ళ్ ఌ	న్
Kannada	ಌ	ನ್

Currently, the lack of a Telugu Nukta either leads to inaccurate transliteration or requires ad-hoc strategies to preserve one-to-one equivalence. For instance, *Aksharamukha* (http://aksharamukha.appspot.com/) uses an ad-hoc spacing dot (U+00B7) as a Nukta equivalent

for Telugu, which is awkwardly placed next to the characters and does not blend well with the script. For instance, the word /zafār/ when expressed as జఫార్ is inaccurate, when using U+00B7 జ·ఫా·ర్ looks awkward and unnatural. But the sequence జ఼ఫార్, using a Nukta looks more harmonious and native-looking. It would greatly aid transliteration if Telugu could have a native Nukta sign encoded. Some other sample words are showed below:

chattīsgaṛh	छत्तीसगढ़	ఛత్తీస్‌గఢ్
tēṉi	தேனி	తేని
āḻvār	ஆழ்வார்	ఆళ్వార్ (Archaic: ఆఴ్వార్)
malampuḻa	മലമ്പുഴ	మలమ్పుళ (Archaic: మలమ్పుఴ)

6. Placement of Nukta

In general, it is to position itself in the middle below a consonant. But it is recommended to move it a bit left with some consonants as shown below to avoid collision with the aspiration marker) and confusion (with existing characters). For the vowel sign /ai/, it is recommended that the second part of the sign is moved a bit lower to avoid collision with Nukta.

The below shows the Nukta sign ◌఼ combining with క ఖ గ జ డ ఢ ఫ య ళ న alongside other vowel signs.

క్ క఼ కా కి కీ కు కూ కె కే కై కొ కో కౌ (and not కై to avoid collision)

ఖ్ ఖ఼ ఖా ఖి ఖీ ఖు ఖూ ఖె ఖే ఖై ఖొ ఖౌ (and not ఖు to avoid collision)

గ్ గ఼ గా గి గీ గు గూ గె గే గై గొ గో గౌ

జ్ జ఼ జా జి జీ జు జూ జె జే జై జొ జో జౌ

డ్ డ఼ డా డి డీ డు డూ డె డే డై డొ డో డౌ (and not డ఼ to avoid confusion with ఢ)

ఢ్ ఢ఼ ఢా ఢి ఢీ ఢు ఢూ ఢె ఢే ఢై ఢొ ఢో ఢౌ

ఫ్ ఫ఼ ఫా ఫి ఫీ ఫు ఫూ ఫె ఫే ఫై ఫొ ఫో ఫౌ

య్ య఼ యా యి యీ యు యూ యె యే యై యొ యో యౌ

ళ్ ళ఼ ళా ళి ళీ ళు ళూ ళె ళే ళై ళొ ళో ళౌ

న్ న఼ నా ని నీ ను నూ నె నే నై నొ నో నౌ

7. Character to be encoded

It is proposed that Telugu Sign Nukta be encoded in the Telugu block of the UCS with the associated character properties.

0C3C  Telugu Sign Nukta

0C3C;TELUGU SIGN NUKTA;Mn;7;NSM;;;;;N;;;;;

8. Indic Syllabic Category

The following addition should be made to the IndicSyllabicCategory.txt file under:

```
# Indic_Syllabic_Category=Nukta

0C3C          ; Nukta # Mn       TELUGU SIGN NUKTA
```

9. Indic Positional Category

The following addition should be made to the IndicPositionCategory.txt file under:

```
# Indic_Positional_Category=Bottom


0C3C          ; Bottom # Mn       TELUGU SIGN NUKTA
```

10. Collation

As the Nukta is meant for transcribing sounds which are not native to Telugu, it is recommended that consonants with the Nukta are collated after consonants without the Nukta so as to not disturb the existing Telugu collation order.

References

1. Desikacharyulu, T.K. (2000?). Creating Unicode Compatible OpenType Telugu Fonts https://upload.wikimedia.org/wikipedia/te/c/c2/Pothanapaper.PDF
2. Rao, U. G. (2002). Telugu Script. Vishwabharat@TDIL. Issue 5. http://tdil.meity.gov.in/pdf/Vishwabharat/tdil-april-2002.zip
3. Kidambi, S. (n.d-a). ఆండాళ్ అరుళిచ్చెయ్దు తిరుప్పావై (Āṇḍāḷ aruḷicceyda Tiruppāvai). http://prapatti.com/slokas/telugu/naalaayiram/aandaal/tiruppaavai.pdf
4. Kidambi, S. (n.d-b). పెరియాళ్వార్ అరుళిచ్చెయ్దు తిరుప్పల్లాండు (Periyāḻvār aruḷicceyda Tiruppallāṁḍu). http://prapatti.com/slokas/telugu/naalaayiram/periyaazvaar/tiruppallaandu.pdf

ISO/IEC JTC 1/SC 2/WG 2
PROPOSAL SUMMARY FORM TO ACCOMPANY SUBMISSIONS
FOR ADDITIONS TO THE REPERTOIRE OF ISO/IEC 10646[2]
Please fill all the sections A, B and C below.
Please read Principles and Procedures Document (P & P) from http://std.dkuug.dk/JTC1/SC2/WG2/docs/principles.html **for guidelines and details before filling this form.**
Please ensure you are using the latest Form from http://std.dkuug.dk/JTC1/SC2/WG2/docs/summaryform.html **.**
See also http://std.dkuug.dk/JTC1/SC2/WG2/docs/roadmaps.html **for latest *Roadmaps*.**

A. Administrative

1. **Title:** ***Proposal to Encode Telugu Sign Nukta***
2. Requester's name: *Vinodh Rajan, Shriramana Sharma, Suresh Kolichala*
3. Requester type (Member body/Liaison/Individual contribution): *Individual*
4. Submission date: *11/12/2019*
5. Requester's reference (if applicable):
6. Choose one of the following:
   This is a complete proposal: *Yes*
   (or) More information will be provided later:

B. Technical – General

1. Choose one of the following:
   a. This proposal is for a new script (set of characters):
      Proposed name of script:
   b. The proposal is for addition of character(s) to an existing block: *Yes*
      Name of the existing block: *Telugu*
2. Number of characters in proposal: *1*
3. Proposed category (select one from below - see section 2.2 of P&P document):
   A-Contemporary A | B.1-Specialized (small collection) | B.2-Specialized (large collection)
   C-Major extinct | D-Attested extinct | E-Minor extinct
   F-Archaic Hieroglyphic or Ideographic | G-Obscure or questionable usage symbols
4. Is a repertoire including character names provided?
   a. If YES, are the names in accordance with the "character naming guidelines" in Annex L of P&P document? *Yes*
   b. Are the character shapes attached in a legible form suitable for review? *Yes*
5. Fonts related:
   a. Who will provide the appropriate computerized font to the Project Editor of 10646 for publishing the standard?
   Vinodh Rajan
   b. Identify the party granting a license for use of the font by the editors (include address, e-mail, ftp-site, etc.):
   Vinodh Rajan, vinodh@virtualvinodh.com
6. References:
   a. Are references (to other character sets, dictionaries, descriptive texts etc.) provided? *Yes*
   b. Are published examples of use (such as samples from newspapers, magazines, or other sources) of proposed characters attached? *Yes*
7. Special encoding issues:
   Does the proposal address other aspects of character data processing (if applicable) such as input, presentation, sorting, searching, indexing, transliteration etc. (if yes please enclose information)? *Yes*
   Sorting, Transliteration

8. Additional Information:

Submitters are invited to provide any additional information about Properties of the proposed Character(s) or Script that will assist in correct understanding of and correct linguistic processing of the proposed character(s) or script. Examples of such properties are: Casing information, Numeric information, Currency information, Display behaviour information such as line breaks, widths etc., Combining behaviour, Spacing behaviour, Directional behaviour, Default Collation behaviour, relevance in Mark Up contexts, Compatibility equivalence and other Unicode normalization related information. See the Unicode standard at http://www.unicode.org for such information on other scripts. Also see Unicode Character Database (http://www.unicode.org/reports/tr44/) and associated Unicode Technical Reports for information needed for consideration by the Unicode Technical Committee for inclusion in the Unicode Standard.

[2] Form number: N4502-F (Original 1994-10-14; Revised 1995-01, 1995-04, 1996-04, 1996-08, 1999-03, 2001-05, 2001-09, 2003-11, 2005-01, 2005-09, 2005-10, 2007-03, 2008-05, 2009-11, 2011-03, 2012-01)

C. Technical - Justification

1. Has this proposal for addition of character(s) been submitted before? *No*
   If YES explain
2. Has contact been made to members of the user community (for example: National Body, user groups of the script or characters, other experts, etc.)? *Yes*
   If YES, with whom? *Suresh Kolichala & Vinodh Rajan*
   If YES, available relevant documents: *-*
3. Information on the user community for the proposed characters (for example: size, demographics, information technology use, or publishing use) is included? *Yes*
   Reference: *See Proposal*
4. The context of use for the proposed characters (type of use; common or rare) *Common*
   Reference: *See Proposal*
5. Are the proposed characters in current use by the user community? *Yes*
   If YES, where? Reference: *See Proposal*
6. After giving due considerations to the principles in the P&P document must the proposed characters be entirely in the BMP? *Yes*
   If YES, is a rationale provided? *Yes*
   If YES, reference: *Telugu is in BMP*
7. Should the proposed characters be kept together in a contiguous range (rather than being scattered)? *Yes*
8. Can any of the proposed characters be considered a presentation form of an existing character or character sequence? *No*
   If YES, is a rationale for its inclusion provided?
   If YES, reference:
9. Can any of the proposed characters be encoded using a composed character sequence of either existing characters or other proposed characters? *No*
   If YES, is a rationale for its inclusion provided?
   If YES, reference:
10. Can any of the proposed character(s) be considered to be similar (in appearance or function) to, or could be confused with, an existing character? *No*
    If YES, is a rationale for its inclusion provided?
    If YES, reference:
11. Does the proposal include use of combining characters and/or use of composite sequences? *Yes*
    If YES, is a rationale for such use provided? *Yes*
    If YES, reference: *See Proposal*
    Is a list of composite sequences and their corresponding glyph images (graphic symbols) provided?
    If YES, reference:
12. Does the proposal contain characters with any special properties such as control function or similar semantics? *No*
    If YES, describe in detail (include attachment if necessary)
13. Does the proposal contain any Ideographic compatibility characters? *No*
    If YES, are the equivalent corresponding unified ideographic characters identified?
    If YES, reference: