L2/20-085

18/03/2020

Revised Proposal to Encode Telugu Sign Nukta

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

Introduction

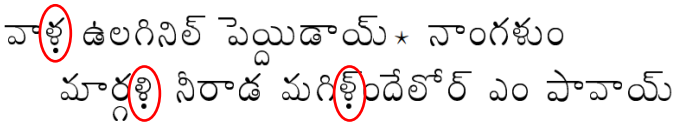
This is a revised proposal of L2/19-401 that consolidates various attestations (from L2/405 and additional new ones) and attempts to provide a streamlined proposal for its inclusion in TUS, alongside resolving confusability issues associated with the previous proposal.

Nukta is a consonantal diacritic used in Indic scripts to extend the native character repertoire and denote non-native phonemes. 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.

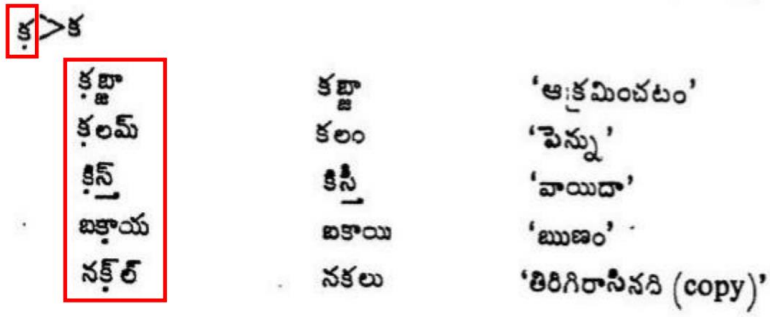
While Nukta is a common feature of North Indic scripts (due to significant lexical borrowing from Persian and Arabic), Kannada is the only South Indic script to have a wide-spread use of Nukta to represent the phonemes, /f/ and /z/. In Tamil, ஃ U+0B83 seemingly takes a Nukta-like role to represent those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

Attestations for Telugu Sign Nukta

Nukta with Telugu Letter LLA (from L2/19-401):



Nukta with Telugu Letter KA & GA (From L2/19-405):

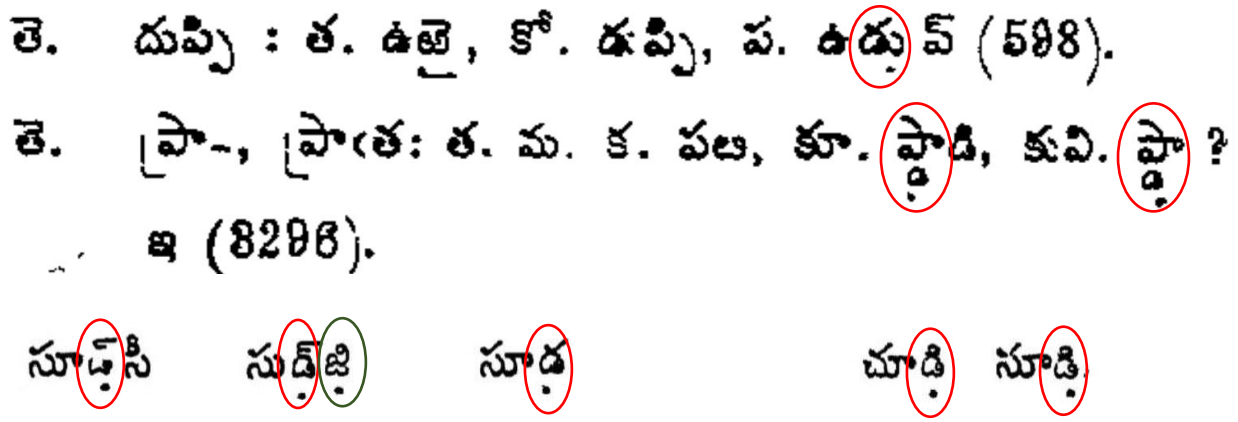


18/03/2020

గ > గ

గ़ాయిబ్	గాయెబు	'కనబడకపోవటం'
దగ़ా	దగా	'మోసం'
వగ़ైరా	వగైరా	'మొదలైనవి'
గ़లత్	గలతు	'తప్పు'

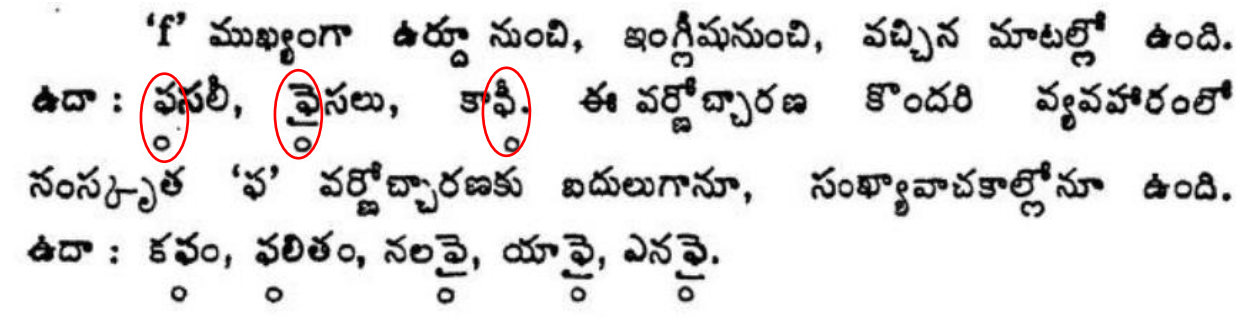
Nukta with Telugu Letter DDA & JA (Krishnamurti, 1979):



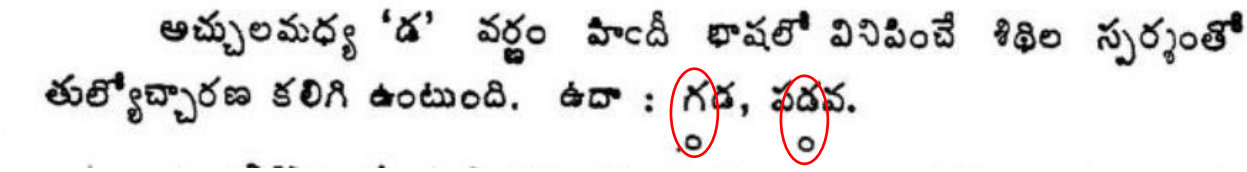
All of the above show Nukta as a dot.

However, below are further instances that use a 'circle'-shaped Nukta.

Nukta with Telugu Letter PHA (Krishnamurti, 1979):



Nukta with Telugu Letter GA and DDA (Krishnamurti, 1979):

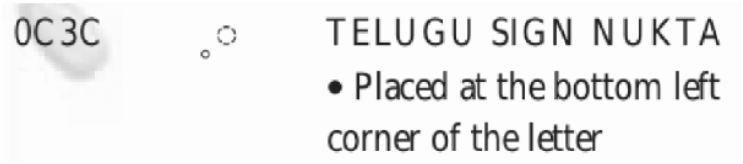


18/03/2020

Unifying ‘Dot’ and ‘Circular’ Nuktas

It is quite common in handwriting to replace a dot with a small circle. For instance, Latin /i/ can also appear in handwriting with a circle on top. Even with Indic scripts, the Tamil Virama is sometimes stylistically expressed as a circle.

The Telugu Nukta sign appearing in a Government of India publication (Rao, 2002) (without any attestations) is also denoted as a circle (apart from the curious ‘placed at the bottom left corner of the letter’ annotation).



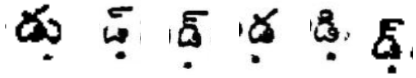
Both the ‘dot’ and ‘circular’ Nukta are used to denote non-native phonemes and, hence, carry the same semantics. We, therefore, consider them to be just glyphic variants of the same underlying character.

Resolving Confusion with Aspirated Consonants

Aspirated Telugu consonants usually have a ‘tear-drop’ as a component of their graphemic structure. Specifically, for the four consonants: CHA ఛ, DDHA ఢ, DHA ధ & PHA ఫ, the ‘tear-drop’ is the only factor that differentiates from their unaspirated counterparts: CA చ, DDA డ, DA ద & PA ప.

It may appear that a combination of the above four unaspirated consonants and Nukta might look deceptively similar to the corresponding aspirated consonants. However, it can be seen (as shown below) that this is overcome by placing Nukta quite far from the base of the consonant, in contrast to the ‘tear-drop’ that usually occurs in close proximity to the base.

DDA vs DDHA



DDA with Nukta (as a dot)

DDA with Nukta (as a circle)

It can be clearly noted that the *Nukta-ted DDA* appears quite distinct from the *aspirate DDHA*. In fact, the most distinctive form of *DDA + Nukta* is with the 'circular' variant. Below table shows the specific aspirate and unaspirated consonants along with the two glyphic variants of Nukta.

Consonants without Teardrop (CWoT)	చ డ ద ప
Consonants with Teardrop (CWT)	ఛ ఢ ధ ఫ
CWoT + Nukta (Dot)	చ఼ డ఼ ద఼ ప఼
CWoT + Nukta (Circular)	చ఼ డ఼ ద఼ ప఼
CWT + Nukta (Dot)	ఛ఼ ఢ఼ ధ఼ ఫ఼
CWT + Nukta (Circular)	ఛ఼ ఢ఼ ధ఼ ఫ఼

(*The attested Nukta-ted forms have been highlighted in red*)

Nevertheless**,** to avoid any confusion (even if minimal) that may arise due to the use of a dot-shaped Nukta with the above four consonants, we propose to use the 'circle'-shaped Nukta as the representative shape to be shown in the code charts along with an annotation about its alternate appearance.

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 and recommended annotations.

0C3C ◌ Telugu Sign Nukta

- Can also appear as a big dot
- It must be placed sufficiently below the baseline of a consonant to avoid confusion/collision with the aspiration marker

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

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
```

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
```

Collation

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

References

1. Krishnamurti, B (Ed.) (1979). Telugu Bhāṣācaritra. Andhara Pradesh Sahitya Academy, Hyderabad. https://archive.org/details/in.ernet.dli.2015.392597
2. Rao, U. G. (2002). Telugu Script. Vishwabharat@TDIL. Issue 5. http://tdil.meity.gov.in/pdf/Vishwabharat/tdil-april-2002.zip

18/03/2020

ISO/IEC JTC 1/SC 2/WG 2
PROPOSAL SUMMARY FORM TO ACCOMPANY SUBMISSIONS
FOR ADDITIONS TO THE REPERTOIRE OF ISO/IEC 10646[1]
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: *18/03/2020*
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

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.

[1] 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)

18/03/2020

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?	*Yes*
If YES, reference:	*See Proposal*
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:	