L2/16-084

Proposal to Encode Single-dot Sign at U+1133B

Comment on M. Hosken's Badaga dot sign Proposal (L2/15-256)

Dr. N. Ganesan (naa.ganesan@gmail.com)

1.0 Introduction: Martin Hosken, in L2/15-256, has proposed for encoding a new Nukta sign of a dot below for Badaga consonants in Tamil script block. This is a novel method and just beginning to be used by few people and not widespread practice. Badaga language is traditionally written using Kannada script and some Badaga books in Kannada script are given at the end. In Tamil script, the voiced consonants are usually written using 2, 3, 4 as superscript or subscript diacritic upon the consonants and this method of using numeral digits in Tamil script is documented in The Unicode Standard. For the same purpose, if it is desired to use a dot below diacritic, that dot-below sign can be brought from the related script of Grantha block. A code point and the character name is suggested for the dot below sign in this document. This is in parallel with the two-dot below sign being brought upon Tamil letters from the Grantha block.

2.0 Nukta signs from Grantha block: In November 2015 UTC meeting, it is decided to use Script Extensions property to 1133C GRANTHA SIGN NUKTA to enable the two-dot nukta usage on Tamil letters as requested in L2/15-256. Instead of using single-dot nukta from North Indian scripts such as Devanagari, it is requested to use a script that is much closer to Tamil script for single-dot nukta sign also. Refer to the mail by Dr. Ken Whistler, dated 2/1/2016, suggesting similar approach for bringing in a one-dot nukta upon Tamil characters. Dr. Ken Whister's email to Unicode members' list is attached at the end. The reason for both nukta signs (one-dot and two-dot signs below) in the Grantha block is consistency and to use a closely related script for both of the nukta signs needed for Badaga and other Nilgiris hill languages. *Historically, Tamil script is related to Grantha script and not to Devanagari script.* There are even five Grantha consonants in Tamil code chart encoded in the Unicode standard for transliteration purposes. Hence it will be very appropriate to employ the Grantha block for addition of nukta signs to Tamil script and bring them via the Script Extension technique.

3.0 Character Name for Single-dot Nukta at U+1133B in Grantha block:

As suggested in Ken Whistler's original mail (1/29/2016), U+1133B can be assigned for the single dot below sign in Grantha block. In the Grantha derived scripts, the name

used for dot/drop is the Sanskrit word, VINDU (later also called BINDU as V- > B- over time). This can be seen in Grantha related scripts: (1) BALINESE WINDU U+1B5C (2) JAVANESE PADA WINDU U+ A9C6 (3) THAI CHARACTER PHINTHU U+0E3A etc., Hence, our request is to encode the single-dot below sign at **U+1133B GRANTHA SIGN VINDU**. This new single-dot nukta sign will be just above U+1133C GRANTHA SIGN NUKTA in the Grantha block in the Standard. Dr. Ken Whistler's mail (2/1/2016) is attached below for reference.

4.0 Ken Whistler's mail on the single-dot below sign on Tamil letters:

On 1/29/2016 8:57 PM, John Hudson wrote:
<<<
Presuming the normalised ordering of a nukta within the Tamil script should follow the typical pattern, then yes, CCC=7 would make sense, and would exclude both U+0323 and U+1CDD, which are not subject to the same normalisation ordering. Taken in consideration with my previous point regarding the benefits to developers of encoding marks alongside the bases to which they are applied, this seems to me the strongest technical case in favour of encoding a new nukta character within the Tamil script block.
>>>

That does seem to be the consideration that led to the initial UTC decision.

But let's make another assessment from scratch.

The problem which needs to be addressed is how to represent the diacritic dot (or dots) below that are manifestly present in Badaga written text (and in several other minority language orthographies) written with the Tamil script in Tamil Nadu.

The engineering requirements that I see falling out from this are:

1. The diacritic must be a combining mark with ccc=7.

This comes from the general Indic rendering system implementation requirements, as suggested by Martin. It takes 0323 COMBINING DOT BELOW (ccc=220) off the table.

2. The diacritic must not unnecessarily break script runs in Tamil.

That follows from the fact that the *rest* of the text in question is going to be using Tamil characters. Encoding this with a character that breaks script runs will do nobody any favors.

3. The diacritic must be easily available on Badaga language keyboards and display correctly with fonts that support Badaga text display.

4. The single dot form and the double dot form should not be unified as a single character for representation.

Unification of those two would be an unusual step and lead to confusion in use and data representation, I think. In that respect I agree with Michael Everson -- so the requirement from the data is for *two* distinct characters that meet

requirements #1 and #2, although there is only a single *function* involved here, and no single orthography seems to need to distinguish a single dot diacritic from a double dot diacritic systematically.

O.k. so much for requirements. Here are the *non*-requirements.

A. The diacritic (or actually pair of them) does not need to be called a "nukta".

B. The diacritic does not need to have a name starting with "TAMIL".

C. The diacritic does not even need to formally be Script=Tamil.

D. The diacritic does not need to be encoded in the Tamil block.

E. The diacritic does not have to *not* occur in fonts that support the Tamil language, because such a font can have language-specific extensions that are simply not seen in normal Tamil language data in the Tamil orthography.

F. A distinction between a dot form glyph and a small circle form glyph does not require an encoding distinction for this set of writing systems.

This last has been amply demonstrated in this general area for the pulli, for example. It is simply a stylistic variation, and not a contrastive character distinction requiring encoding. (On this point , I *disagree* with Michael Everson.)

O.k., now with that set of requirements (and *non*-requirements), what are our options?

First, let's take a census of ccc=7 combining marks that already exist in the standard and whose basic shape is a dot below (or a double dot below):

Single dot below shape:

```
093C        ; 7 # Mn      DEVANAGARI SIGN NUKTA
09BC        ; 7 # Mn      BENGALI SIGN NUKTA
0A3C        ; 7 # Mn      GURMUKHI SIGN NUKTA
0ABC        ; 7 # Mn      GUJARATI SIGN NUKTA
0B3C        ; 7 # Mn      ORIYA SIGN NUKTA
1037        ; 7 # Mn      MYANMAR SIGN DOT BELOW
1C37        ; 7 # Mn      LEPCHA SIGN NUKTA
110BA       ; 7 # Mn      KAITHI SIGN NUKTA
11173       ; 7 # Mn      MAHAJANI SIGN NUKTA
111CA       ; 7 # Mn      SHARADA SIGN NUKTA
112E9       ; 7 # Mn      KHUDAWADI SIGN NUKTA
114C3       ; 7 # Mn      TIRHUTA SIGN NUKTA
115C0       ; 7 # Mn      SIDDHAM SIGN NUKTA
116B7       ; 7 # Mn      TAKRI SIGN NUKTA
```

Double dot below shape:

```
0CBC        ; 7 # Mn      KANNADA SIGN NUKTA
1133C       ; 7 # Mn      GRANTHA SIGN NUKTA
```

Now if we intersect that list with the set consisting
of Indic_Syllabic_Category=Nukta, that will remove
the Myanmar dot below from the list. (Note that there
are also other ISC=Nukta characters which are not ccc=7
and which are not called "nukta", for whatever historical
reasons.)

Looking at the candidate list, there is no good reason to
favor one of the historic North India scripts over any of
the widely implemented modern scripts. And Bengali,
Gurmukhi, and Gujarati would seem to have no particular
advantage, compared to Devanagari, either -- other than
their not *being* Devanagari. At least Odia is more southern --
spoken and written widely in Andhra Pradesh, as well as Odisha.
That would then further pare down the list to:

Single dot below shape:

093C ; 7 # Mn DEVANAGARI SIGN NUKTA
0B3C ; 7 # Mn ORIYA SIGN NUKTA

Double dot below shape:

0CBC ; 7 # Mn KANNADA SIGN NUKTA
1133C ; 7 # Mn GRANTHA SIGN NUKTA

The solution favored for the double dot below diacritic for
the minority languages written in Tamil script orthographies
is to use U+1133C, the Grantha nukta. That has the correct
combining class, the correct shape, and does not break
script runs (in implementations that support Script Extensions)
because it is already scx={Gran Taml}. So to meet all the
requirements, all that is still needed is to make sure that
U+1133C is added to keyboards and fonts supporting any
minority language that use the double dot below letter
diacritic.

Now for the single dot below that has caused all this ruckus.
There are three classes of solutions:

1. Encode a new, dedicated character in the correct script.

That is what the UTC chose originally in response to Martin
Hosken's proposal: add U+0BBC TAMIL SIGN NUKTA. That
has the correct combining class, the correct appearance,
and correct behavior for script runs (assuming it was also
assigned Script=Tamil). It would be easy to add to keyboards
and fonts.

2. Use one of the existing single dot nuktas from a different
script. Choosing from
the above, pared-down list, the likeliest candidates are
either the Devanagari nukta (093C) or the Oriya nukta (0B3C).

Either of those would have the correct combining class and
the correct shape. To address the script run issue, an appropriate
entry for Script_Extensions would have to be added. Devanagari
has lots of precedents for being shared around this way -- there

are characters with shared usage with Grantha, and even one
shared with Tamil already:

A8F3 ; Deva Taml # Lo DEVANAGARI SIGN CANDRABINDU VIRAMA

And that is in addition to the use of the Devanagari dandas for
the occasional instances of straight line dandas in various
South Indian scripts, for example.

Getting either a Devanagari or an Oriya nukta into keyboards and
fonts would also be straightforward, but would require more
people to pay attention than for the first case. People would need
to know about the additional use for Badaga, etc., written in Tamil.

3. Encode a new character but don't *call* it TAMIL SIGN NUKTA.

What people may be missing here is that there is a possible third
way. All discussions of Tamil seem inevitably to end up mesmerized
by names and code charts and phonetic functions, instead of
focusing on the real issue: We need to be able to reliably represent
a dot below used as a diacritic in Badaga when that language is
written using the Tamil script. We could accomplish essentially
that with something that triggers none of the irrelevant discussion.

For example, something like:

XXXX SOUTH INDIAN LETTER-FORMING DOT BELOW

Make that gc=Mn, ccc=7. Then give it Script=Inherited. The latter
will take care of the script run issue automatically. Make it
ISC=Nukta, so the rendering engines understand how it is
supposed to function. The requirement
to make it into Badaga keyboards and fonts is no more onerous than
that for ensuring that a Devanagari or an Oriya nukta would make it there.

As for chart position, because Tamilians clearly detest the idea
of adding it to the Tamil chart, just make use of a code chart
hole elsewhere. I would suggest the Grantha block as the obvious
candidate, since we are already suggesting use of the Granta nukta
for the two-dot forms. (Obvious code position: U+1133B, but it
could go after the cantillation marks, just as well.)

Take a look at the list of non-requirements I listed above. By not
calling this a "nukta", not using "TAMIL" in its name, not giving
it the Tamil script property, and not putting it in the Tamil code
chart, we are avoiding all of the obvious problems, while having
a technically equivalent solution that will give Badaga writers
the character they need for their orthographies.

Think about it.
-Ken

References:

Badaga language is traditionally printed using Kannada script:

(1) L. D. Barnett. *A Catalogue of the Kannada, Badaga, and Kurg Books in the Library of the British Museum.* London: Longmans & Co., 1910
.http://dsal.uchicago.edu/bibliographic/bmcatalogs/Z7049.I3B86.pdf
(2) Using Kannada script, the Gospel of Mark: 1896
https://archive.org/stream/gospelofmarkinba00sike#page/n0/mode/2up
(3) Gospel of Luke, 1890
http://gospelgo.com/q/Badaga%20Bible%20-%20Gospel%20of%20Luke.pdf