

Comments on L2/09-141, Proposal to Encode the Grantha Script...

Kent Karlsson
2009-08-06

Document L2/09-141, *Proposal to Encode the Grantha Script in the Basic Multilingual Plane (BMP) of ISO/IEC 10646*, (despite the title, the suggested code points Grantha characters are in plane 1, not the BMP), suggests to encode the Grantha characters as a separate script, and with several precomposed characters. However, there is no longer any need to encode precomposed characters. Furthermore, many of the proposed characters are identical to already encoded Tamil characters. Double encoding poses security issues. This identity has an extent that is much more than coincidental or “far removed by history” (as for Greek, Latin, Cyrillic). The conclusion here is that Grantha character should *not* be encoded as a separate script, but only as additions to the Tamil script. This note does not include an encoding proposal, just an analysis of the given proposal, and a conclusion from that.

1 Within Grantha decompositions

Looking through the proposed code chart, it is clear that several of the proposed characters can, and should, be represented by sequences of characters. The list below gives what would be canonical decomposition mappings. The decomposable characters should not be encoded, and if encoded they should have canonical decomposition as per below.

11306 = 11305, 1133E

GRANTHA LETTER AA = GRANTHA LETTER A, GRANTHA VOWEL SIGN AA

This is a vowel letter with a vowel sign to “lengthen” it. This is a ligature in display, but that is common for Tamil and Grantha vowel signs.

1130A = 11309, 11357

GRANTHA LETTER UU = GRANTHA LETTER U, GRANTHA AU LENGTH MARK

This is a vowel letter with a vowel sign to “lengthen” it. This is a ligature in display, but that is common for Tamil and Grantha vowel signs.

1130E = 1130F, dot above

GRANTHA LETTER E = GRANTHA LETTER EE, dot above

This is a vowel letter with a dot above to modify it. U+0307, COMBINING DOT ABOVE, can be used for the dot.

11310 = 1130F, 11347

GRANTHA LETTER AI = GRANTHA LETTER EE, GRANTHA VOWEL SIGN EE

This is a vowel letter with a vowel sign to modify it.

11312 = 11313, dot above

GRANTHA LETTER O = GRANTHA LETTER OO, dot above

This is a vowel letter with a dot above to modify it. U+0307, COMBINING DOT ABOVE, can be used for the dot.

11314 = 11313, 11357

GRANTHA LETTER AU = GRANTHA LETTER OO, GRANTHA AU LENGTH MARK

This is a vowel letter with a vowel sign to “lengthen” it.

11346 = 11347, dot above

GRANTHA VOWEL SIGN E = GRANTHA VOWEL SIGN EE, dot above

This is a vowel letter with a dot above to modify it. U+0307, COMBINING DOT ABOVE, can be used for the dot.

11348 = 11347, 11347

GRANTHA VOWEL SIGN AI = GRANTHA VOWEL SIGN EE, GRANTHA VOWEL SIGN EE

This is a doubled vowel sign. The doubling signifies a modification.

1134A = 11346, 1133E

GRANTHA VOWEL SIGN O = GRANTHA VOWEL SIGN E, GRANTHA VOWEL SIGN AA

This is a split vowel sign. It's best to represent it using the components.

1134B = 11347, 1133E

GRANTHA VOWEL SIGN OO = GRANTHA VOWEL SIGN EE, GRANTHA VOWEL SIGN AA

This is a split vowel sign. It's best to represent it using the components.

1134C = 11347, 11357

GRANTHA VOWEL SIGN AU = GRANTHA VOWEL SIGN EE, GRANTHA AU LENGTH MARK

This is a split vowel sign. It's best to represent it using the components.

11360 = 1130B, 3-like squiggle

GRANTHA LETTER VOCALIC RR = GRANTHA LETTER VOCALIC R, 3-like squiggle

This is a letter modified with an explicit mark (a 3-like squiggle). That mark has not been proposed for encoding, but it should be encoded as a separate combining character.

11361 = 11362, 3-like squiggle

GRANTHA LETTER VOCALIC LL = GRANTHA VOWEL SIGN VOCALIC L, 3-like squiggle

This is a letter modified with an explicit mark (a 3-like squiggle). That mark has not been proposed for encoding, but it should be encoded as a separate combining character. Note that the first component is called, in part, “vowel sign”, but it is NOT a combining mark (see next item). Indeed the decomposition mapping for 11361 should be to <1130C, 3-like squiggle> = <GRANTHA LETTER VOCALIC L, 3-like squiggle>. The 3-like squiggle is the same as for the preceding character.

11362 = 1130C

GRANTHA VOWEL SIGN VOCALIC L = GRANTHA LETTER VOCALIC L

This is a simple one-to-one character equivalence. Note that this is called, in part, “vowel sign”, but it is NOT a combining mark.

11363 = 11362, ...

GRANTHA VOWEL SIGN VOCALIC LL = GRANTHA VOWEL SIGN VOCALIC L, ...

This seems to be a vowel letter with a vowel sign to modify it. This is a ligature in display, but that is common for Tamil and Grantha vowel signs; but the ligation makes it hard to see exactly which vowel sign is used as modifier. Note that 11363 is called, in part, “vowel sign”, but it is NOT a combining mark. Indeed the decomposition mapping for 11363 should be to <1130C, ...> = <GRANTHA LETTER VOCALIC L, ...>.

2 Grantha to Tamil decompositions

Comparing the proposed Grantha chart with the Tamil code chart indicates that Grantha is far from a separate script from Tamil. Indeed the commonality is so great that one need only encode a few additions to the Tamil script to cover Grantha (Tamil Grantha).

In the list below, the code points to the left are those proposed in L2/09-141.

11309 = 0B89 உ, TAMIL LETTER U

1130A = 0B8A ஊ, TAMIL LETTER UU; should have been = <TAMIL LETTER U, TAMIL AU LENGTH MARK>

11313 = 0B92 ஒ, TAMIL LETTER O **(used for OO in Grantha)**

11314 = 0B94 ஔ, TAMIL LETTER AU; canonically equivalent to <TAMIL LETTER O, TAMIL AU LENGTH MARK> (ligates)

11315 = 0B95 க, TAMIL LETTER KA

1131C = 0B9C ங, TAMIL LETTER JA

1131F = 0B9F ட, TAMIL LETTER TTA

11323 = 0BA3 ண, TAMIL LETTER NNA

11324 = 0BA4 த, TAMIL LETTER TA

11328 = 0BA8 ற, TAMIL LETTER NA

11329 = 0BA9 ன, TAMIL LETTER NNNA

1132F = 0BAF ய, TAMIL LETTER YA

11331 = 0BB1 ர, TAMIL LETTER RRA

11333 = 0BB3 ள, TAMIL LETTER LLA
 11334 = 0BB4 ழ, TAMIL LETTER LLLA
 11335 = 0BB5 வ, TAMIL LETTER VA
 11336 = 0BB6 ஶ, TAMIL LETTER SHA
 11337 = 0BB7 ஷ, TAMIL LETTER SSA
 11338 = 0BB8 ஸ, TAMIL LETTER SA
 11339 = 0BB9 ஹ, TAMIL LETTER HA
 1133E = 0BBE ஈ, TAMIL VOWEL SIGN AA
 1133F = 0BBF ி, TAMIL VOWEL SIGN I
 11340 = 0BC0 ீ, TAMIL VOWEL SIGN II
 11341 = 0BC1 ு, TAMIL VOWEL SIGN U (the Unicode 5.0 chart glyph is not representative)
 11342 = 0BC2 ூ, TAMIL VOWEL SIGN UU (the Unicode 5.0 chart glyph is not representative)
 11347 = 0BC6 ெ, TAMIL VOWEL SIGN E **(used for EE in Grantha)**
 1134B = 0BCA ொ, TAMIL VOWEL SIGN O **(used for OO in Grantha)**, has canonical decomposition
 1134C = 0BCC ோ, TAMIL VOWEL SIGN AU, has canonical decomposition
 11357 = 0BD7 ள, TAMIL AU LENGTH MARK (“length mark” is a misnomer, this is a vowel sign)
 11366 = 0BE6 ௦, TAMIL DIGIT ZERO
 11367 = 0BE7 ௧, TAMIL DIGIT ONE
 11368 = 0BE8 ௨, TAMIL DIGIT TWO
 11369 = 0BE9 ௩, TAMIL DIGIT THREE
 1136A = 0BEA ௪, TAMIL DIGIT FOUR
 1136B = 0BEB ௫, TAMIL DIGIT FIVE
 1136C = 0BEC ௬, TAMIL DIGIT SIX
 1136D = 0BED ௭, TAMIL DIGIT SEVEN
 1136E = 0BEE ௮, TAMIL DIGIT EIGHT
 1136F = 0BEF ௯, TAMIL DIGIT NINE
 11370 = 0BF0 ௩, TAMIL NUMBER TEN
 11371 = 0BF1 ஈ, TAMIL NUMBER ONE HUNDRED
 11372 = 0BF2 ஐ, TAMIL NUMBER ONE THOUSAND

Maybe even more characters can be equated (avagraha, om, ...), seeing the differences in the charts as only font differences. The ones given in the list here are only the obviously equivalent ones.

However, 1134D, GRANTHA SIGN VIRAMA, should probably not be equated with 0BCD, TAMIL SIGN VIRAMA. They look different, and the former induces a number of conjuncts, or subjoined letters. The Tamil virama does that only to a very limited extent (ksa and shr(ii)). However, it would be better to disunify the control function and the graphic mark function of virama (as done for Khmer).

3 Conclusion

Many (at least 42 out of 99) of the proposed GRANTHA characters should simply be equated with the corresponding TAMIL characters. Some additional proposed characters should be decomposed and always represented by the decomposition (not encoding the precomposed form). Only a smaller number of characters remain to be encoded as a Tamil Grantha addition to the Tamil script.