L2/10-085

Feedback to Dr Anderson’s Grantha Summary dt 2010-Jan-28

Shriramana Sharma, 2010-Mar-12

Miscellanea

Erratum in my Grantha proposal L2/09-372

On page 40, the first line under §6.4.2 should read “anunasika sign” instead of “anusvara”. Now I shall proceed with my feedback to Dr Anderson’s document. All page and sections numbers hereafter refer to that document.

“Errors in character names”

On page 6, § IV 1 a, it is suggested that the use of the word LETTER in the names of the anusvara and visarga characters in my proposal is in error, since in all other Indic scripts they have been named SIGN. However, these words were used intentionally, as per page 40, §6.4.1 of my proposal. These characters are more appropriately analysed as independent letters, and giving them GC=Mn/Mc results in unnecessary problems in implementations due to implementors misunderstanding how a combining mark to be processed as has been discussed very many times in conversations in which I participated.

I also point out that Dr Peter Scharf of Brown University whom the UTC asked for feedback to my various documents has agreed in his mail dated 2009-Nov-01 that these characters should not be combining characters. (He does not mention the anusvara in his mail, but that is probably because he was writing in feedback to my L2/09-343 about ardhavisarga. The argument for the anusvara is the same, in Dr Scharf’s words “they are independent signs for consonantal sounds”.)

I however observe that I have indeed committed an error in that while I have given GC=Mc to be in accordance with other Indic scripts, I should have changed the character name accordingly to use the word SIGN rather than LETTER. If these characters are given GC=Mc then they should be called SIGN and if GC=Lo then LETTER. The decision of which GC to allot and hence which word to use in the name is of course up to the UTC and I agree to abide by their decision on this matter.

Other issues

On page 10, § IV 5 j, it is mentioned that I have asked for the characters to be separately encoded “because they have different CCs”. However, on page 8 of my proposal under §3.5.1, I have carefully stated that “they would differ in their GC (Mc against Mn). If it were not for the requirement for Indic vowel signs to have CCC=0, they would differ in that also". Therefore the correct reason is the difference in GC and not CC.

Below that, in § IV 5 k, it is pointed out that the names for the characters for vowel sign AU used by Ganesan and GoI follows that seen in Malayalam whereas mine differ. My opinion is just that the name should properly reflect the actual use. For scripts which may have already been encoded it may be impossible to change the names or unnecessary to add aliases. However, for a script that is yet to be encoded, what is there to prevent us from giving appropriate names for characters? The word "Length Mark" itself has little meaning except for characters like 0CD5 Kannada Length Mark which are actually used to indicate that the vowel represented by the previous character is to be lengthened. I request that some imaginary requirement for names to be common throughout Indic scripts not affect the way in which the names reflect the actual usage of the characters.

Next, in § IV 5 m, it is mentioned that "The Ligating Virama is proposed as a means of handling the repha: the repha can occur between a consonant and a spacing virama, but it cannot be placed between a consonant and the ligating virama." A small correction is required here. The repha cannot be placed between a consonant and *either the ligating or touching virama*. Thus there is a clear behavioural distinction between the spacing and the other two forms and even between the other two forms, there is a clear glyphic distinction. Summarily, there is a three-fold distinction between the virama forms of Grantha, and since using ZWJ/ZWNJ is only sufficient to unambiguously represent a two-fold distinction, and the usage of combinations of the above is also inappropriate (for details see L2/09-375) I have proposed the additional character called ligating virama.

Later, on page 11 in § V 4, a recommendation has been made to use the word Grantha Au Length Mark for 11357. I reiterate that I am not satisfied by this. I do not see the urgent need to keep the name in line with other Indic scripts *even if it is a bad representation of the actual usage of the character*.

Below, in § V 6, the Vedic characters have been shelved. I need not hide from the UTC my disappointment at this, as this seems to suggest that all the hard work I put into documenting the shape and usage of these characters is not given importance. Neither Ganesan nor GoI have considered that the strongest *existing* use of the Grantha script is among the Vedic scholars and priests of Tamil Nadu, who need these characters. I and others on the unicode@ mailing list have repeatedly iterated that it is very important that Unicode Grantha support Vedic if it is to be quickly adopted into the user community.

While the Tamil/Grantha fractions (especially the minor ones) are not in frequent use today, and hence it may be justified to postpone their encoding, the Vedic svara markers are not so. Today, the use of Grantha for Vedic is also on the decline because of absence of availability of technology to produce new Grantha books (for Vedic). If then a Grantha encoding is to be meaningfully useful to the *real* Grantha user community, and play a role in rejuvenating the usage of the script in *real* life, it should support Vedic. Perhaps there is a perception that that Grantha is only for research scholars handling palm leaf manuscripts, however that perception results from not paying attention to the real world usage of the script. If the UTC requires, and if my request alone is insufficient, I can try to get signed support for my request to the UTC to attach importance to the Vedic characters from the heads of various traditional Vedic schools in and around Tamil Nadu.

On page 12, § V 10, an option is proposed that the virama ligature ("chillu") formation is identified as the default (for the sequence CONSONANT + VIRAMA) and conjunct forms are produced by other mechanisms such as ZWJ. I wish to point out that if this is done, then Grantha will fall out of line with other Indic scripts. When a tendency to align with other Indic scripts is exhibited in the superficial matter of naming characters (w.r.t. "LENGTH MARK" etc) I daresay that even more importance should be attached to ensuring that the behaviour of the virama in Grantha is maintained in line with that in other Indic scripts. I have repeated myself that the usage of the virama is to produce the *default representation* of consonant clusters across Indic scripts. I need not teach the UTC that that default representation is in the order of ligature, conjoining forms and overt virama forms. If now the default representation is made to be virama ligatures, then any (Sanskrit) text being converted from other Indic texts on an one-to-one basis will not achieve its default representation in Grantha, where the regular order of precedence still stands. Therefore the virama should be handled in Grantha in the same way as in other Indic scripts.

Consonants RRA, NNNA and LLLA

The matter of the consonants RRA, NNNA and LLLA is mentioned on page 9 § IV 5 i and page 11 § V 5. In my proposal I originally expressed the view that these characters may be used as-is from the Tamil block. I have found reason to revise my position, however, not for the reasons Mr Ganesan has argued.

First, I have already expressed in page 20 of my document L2/09-316 "Comments on Mr Ganesan's Grantha Proposal" my doubts as to the authenticity of the samples provided by Mr Ganesan as attestation for use of these characters in Grantha. I have repeatedly and politely requested Mr Ganesan as to contact information for the Samskrita Granthalipi Sabha he mentions as the source of this sample, yet received none. Therefore, when it is mentioned on page 11, "Since evidence is attested for the three consonants", I am forced to restate my previous expression of doubt.

However, as I have expressed in page 47 of my proposal §8.4, there is potential for use of LLLA and RRA in Grantha text. Thus for that potential of use, these characters may be encoded separately for Grantha. Having said this, and seeing that characters like Combining Devanagari Digit Eight/Nine were encoded "just to complete the set", I find that I cannot successfully (and perhaps meaningfully) object to the encoding of NNNA as well. However, I request that these characters receive annotations such as 0929, 0931 and 0934 Devanagari Letter NNNA/RRA/LLLA that these characters are used for Dravidian transcription.

Extended Tamil

In pages 43, 44 § 7 of my proposal, I have given attestation samples for and spoken about "pseudo-Manipravalam". This word is a term invented by me to refer to that form of writing Sanskrit using the Tamil script where Grantha characters are imported as a cure for the insufficiency of the Tamil character repertoire to unambiguously represent Sanskrit. Considering that importing Grantha characters into Tamil essentially creates an extended Tamil script which may be accurately referred to as Grantha-supplemented Tamil, for the sake of brevity we shall refer to this as Extended Tamil.

In the same passage of my proposal I have noted that this mixed script form has scope for real-world use just as Grantha, and hence deserves to be accorded serious discussion, as it is not the product of someone's fancy but has been used in books published by respected publishing houses guided by well-learned scholars in Tamil Nadu. Therefore any Unicode model for Grantha should consider and provide for this script form as well.

Characters required for Extended Tamil

These are the additional characters required at a minimum for Extended Tamil and not currently present in the Tamil block:

1) Independent and dependent vowels Vocalic R/RR/L/LL
2) Anusvara, Visarga, Avagraha, Danda-s
3) Second, third and fourth members of the five-member consonant classes, i.e. KHA, GA, GHA, CHA, JHA, TTHA, DDA, DDHA, THA, DA, DHA, PHA, BA, BHA

The following list marks the "new" characters in square brackets:

அ ஆ இ ஈ உ ஊ [𑌋 𑍠 𑌌 𑍡] ஏ ஐ ஓ ஔ

◌ா ◌ி ◌ீ ◌ு ◌ூ [◌𑍃 ◌𑍄 ◌𑍢 ◌𑍣] ◌ே ◌ை ◌ோ ◌ௌ ◌்

[◌𑌂 ◌𑌃 𑌽] । ॥

க [𑌖 𑌗 𑌘] ங - ச [𑌛] ஜ [𑌝] ஞ - ட [𑌠 𑌡 𑌢] ண

த [𑌥 𑌦 𑌧] ந - ப [𑌫 𑌬 𑌭] ம

ய ர ல வ ஶ ஷ ஸ ஹ

The originally-Grantha consonants JA, SHA, SSA, SA, HA have already been imported into the Tamil script in contemporary usage and hence are already encoded in the Tamil block. The danda-s for punctuation are to be used, as always, from the Devanagari block. Therefore, it is necessary to develop a model by which the remaining characters mentioned above may be used within Tamil text so as to form Extended Tamil.

Options for implementing Extended Tamil

There are various options for supporting Extended Tamil. I shall attempt to enumerate these without going into too picky fine details and without omitting the important ones.

Option 1: Add characters into the Tamil block

Since this form of writing is really an extended form of Tamil, an obvious approach would be that it should be supported by encoding additional characters in the existing Tamil block to fill in the empty spaces in that block corresponding to the equivalent characters in other Sanskrit-supporting Indic scripts.

Pros: The Tamil block is full of empty spaces which are not going to be used for anything else. This option would make good use of those spaces. Placing the new characters in the Tamil block would also ensure their visibility and promote their use.

Cons: There are some people overly concerned about the so-called purity of the Tamil script asking for the deprecation (!) of the “Grantha” characters that are already present in the Tamil block with attested usage in regular Tamil books. This being the case, the addition of further Grantha characters to support an obscure form of writing such as Extended Tamil will not be received well by the politicians in the Tamil community. While politics should not decide technical issues, I do not want (myself or the UTC) to be involved in such politics since there are far better things to do than dousing a political fire.

Option 2: Use characters from the Grantha block

If then new characters are not to be encoded in the Tamil block, then the next obvious solution is to use the characters (codepoints) from the Grantha block interspersed within Tamil text.

Pros: No new characters need to be encoded except those which are separately being encoded in the Grantha block.

Cons: Using characters with different script properties in such an interspersed manner would create severe problems with identifying word boundaries (which is one of the uses of the script property if I am not mistaken). Rendering engines do not normally support cross-script rendering. This would create problems since in Extended Tamil Grantha consonants will need to be used with Tamil vowel signs and Tamil consonants with Grantha vowel signs. Further, a policy should be formulated as to what should be done in the case of characters which are identical between the two scripts, such as the vowels U etc, vowel signs AA etc and consonants NNA etc. That would involve several murky issues.

For one, if it is decided to use Grantha codepoints only when equivalent Tamil codepoints do not exist, then those vowel signs such as -AA which are identical between the two scripts would be used with Grantha consonants as well. Then GA + TAMIL -AA would be indistinguishable from GA + GRANTHA -AA which leads to the question, which I raised at the end of page 44 of my proposal, of whether GRANTHA -AA etc should be decomposed to TAMIL -AA etc or not. Similarly one should consider the fully identical consonants NNA etc as well and the almost identical ones like KA, JA, TA etc as well.

Even if it is decided to use Grantha vowel signs with Grantha consonants and Tamil vowel signs with Tamil consonants, one cannot avoid using *some* Grantha vowel signs (vocalic R etc) with Tamil consonants and *some* Tamil vowel signs (EE, AI and OO) with Grantha consonants, which still leads to the script property problems mentioned above.

Option 3: Encode separate characters in a Tamil Supplementary block

Therefore I believe that the best solution is to separately encode those additional characters that are required for Extended Tamil in a Tamil Supplementary block, such as the one requested by me in L2/09-317. These characters number 25 in all and should carry the property script=tamil to enable their painless use among Tamil characters. They are:

1) TAMIL EXTENDED LETTER VOCALIC R
2) TAMIL EXTENDED LETTER VOCALIC RR
3) TAMIL EXTENDED LETTER VOCALIC L

4) TAMIL EXTENDED LETTER VOCALIC LL
5) TAMIL EXTENDED VOWEL SIGN VOCALIC R
6) TAMIL EXTENDED VOWEL SIGN VOCALIC RR
7) TAMIL EXTENDED VOWEL SIGN VOCALIC L
8) TAMIL EXTENDED VOWEL SIGN VOCALIC LL
9) TAMIL EXTENDED SIGN ANUSVARA
10) TAMIL EXTENDED SIGN VISARGA
11) TAMIL EXTENDED SIGN AVAGRAHA
12) TAMIL EXTENDED LETTER KHA
13) TAMIL EXTENDED LETTER GA
14) TAMIL EXTENDED LETTER GHA
15) TAMIL EXTENDED LETTER CHA
16) TAMIL EXTENDED LETTER JHA
17) TAMIL EXTENDED LETTER TTHA
18) TAMIL EXTENDED LETTER DDA
19) TAMIL EXTENDED LETTER DDHA
20) TAMIL EXTENDED LETTER THA
21) TAMIL EXTENDED LETTER DA
22) TAMIL EXTENDED LETTER DHA
23) TAMIL EXTENDED LETTER PHA
24) TAMIL EXTENDED LETTER BA
25) TAMIL EXTENDED LETTER BHA

Here it may be suggested that we avoid duplicating the avagraha and instead use it from the Grantha block, since it is not going to combine with any other character. However, it forms parts of words and hence the word-boundary problem still exists. Therefore I think it is better to encode it.

Similarly it is also not possible to avoid encoding the anusvara citing the existence of a (spurious) Tamil anusvara at 0B82 as the reason, because: 1) the so-called Tamil anusvara has GC=Mn whereas the desired Grantha-style anusvara has GC=Mc. 2) the existing spurious Tamil anusvara (spurious it is not at all used in Tamil as acknowledge in the code chart) looks like a glyphic variant of the Tamil virama (pulli) at 0BCD. (I have previously drawn the UTC's attention to this at the bottom of page 4 of L2/09-324.) So it cannot be suggested that this character be used for the anusvara as it would be confounded with the pulli. Therefore it is necessary to use (reëncode) a Grantha-style spacing anusvara.

I hope that the UTC favours this third option of supporting Extended Tamil, and after the UTC allocates a Tamil Supplementary block, I will submit a separate proposal if necessary for these characters. They can be placed at the end of the block to avoid conflict with the Tamil fractions and other symbols desired to be encoded in such a block.

Pros:

1) This option avoids the need to mix characters of two scripts (i.e. cause cross-script rendering). Therefore it avoids all the cons of the previous model, including having to decide which selection of characters from each block should be used for Extended Tamil and the dilemma of whether to decompose or not, etc.

2) Since the (original) Tamil block is not touched, political issues are unlikely to arise. Saying this, I am obviously hoping that no Tamil politicians argue that these characters needed for Sanskrit should not be encoded in even the Tamil Supplementary block.

3) As I will detail below, this option also supports the implementation of the two different forms of Extended Tamil, as also provides a better way of handling the currently prevalent way of using the superscript digits 2, 3 and 4 with Tamil consonants to represent Sanskrit.

Cons: The only con to this model is the cost of 25 characters. I daresay that this cost is far outweighed by the gains in avoiding the problems involved in either of the two previous models. The SMP is vast (for now) and 25 characters is a pittance comparatively.

Details of implementing Extended Tamil

Now having concluded that the best option for implementing Extended Tamil is Option 3, encoding separate characters in a Tamil Supplementary block, I proceed to enumerate other details regarding this model.

Selection of characters

First, it should be made clear that Extended Tamil should be composed only using characters from the Tamil block and the Tamil Supplementary block and *not* from the Grantha block. There is no decomposition from Grantha characters to Tamil characters and these are treated as two distinct scripts just like Gujarati/Kaithi etc, with alike-looking characters *not* being treated as identical. Therefore all implementations of Extended Tamil beginning with input methods will use only characters from the Tamil block and from the Tamil Supplementary block (which all have script=tamil).

By the way, I should say that I think there is no necessity to give script=common for the digits and numerals (including fractions) that are common to Grantha and Tamil, just as the Devanagari digits do not have script=common despite being common to Devanagari and Kaithi (and perhaps other scripts too). However, I do not know about the other characters such as abbreviations, seeing that 0970 DEVANAGARI ABBREVIATION SIGN has script=common.

Two kinds of Extended Tamil

In my proposal page 44 § 7 I have hinted at but have not fully and exactly described some variations seen in Extended Tamil writing. I attempt to make a better description now.

In some texts, only Grantha-style consonants have been used even in the presence of Tamil equivalents when the Grantha vowel signs for Vocalic R etc need to be attached and the Grantha-style virama has been used for Grantha consonants ("system A"). However, other texts use only Tamil-style characters in these cases even with the Grantha vowel signs and use the Tamil-style virama even with Grantha consonants ("system B").

It is proposed that these two systems be handled with the same set of characters as indicated above using smart font technologies as follows: If desired, a codepoint sequence such as TAMIL LETTER KA + TAMIL EXTENDED VOWEL SIGN VOCALIC R should be displayed equivalent to GRANTHA LETTER KA + GRANTHA VOWEL SIGN VOCALIC R. Similarly if desired, a codepoint sequence such as TAMIL EXTENDED LETTER KHA + TAMIL VIRAMA should be displayed equivalent to GRANTHA LETTER KHA + GRANTHA VIRAMA. In general in System A:

TAMIL CONSONANT + TAMIL EXTENDED VOWEL SIGN → GRANTHA CONSONANT + GRANTHA VOWEL SIGN
TAMIL EXTENDED CONSONANT + TAMIL VIRAMA → GRANTHA CONSONANT + GRANTHA VIRAMA

System B would turn off both these rules. Theoretically there are two more possible writing systems where only one of these above rules is active. All these four systems can be handled by appropriate smart fonts or a single smart font with selectable features [as in Graphite parlance]. Since they are merely variant surface representations of the same content, they must be handled at the font/rendering level and not at the encoding level.

It also should be noted here that in the case of consonant clusters written in Extended Tamil there is only one ligature K·SSA (which may be written in Tamil or Grantha styles) and there is no stacking at all. Therefore even a sequence like D·DHA where both consonants are to be written in the Grantha style (i.e. represented by TAMIL EXTENDED LETTER codepoints) there is no ligature or stacking and there may even be more than one Grantha virama-s displayed such as for D·DH·YA. All this must be handled correctly by smart fonts.

Important: A user should *not* attempt to manually force these writing systems in the absence of appropriate smart font technologies by using appropriate assortment of Grantha and Tamil codepoints. *A portable and compliant implementation or text composed in Extended Tamil should never contain any Grantha codepoints and only codepoints from the Tamil and Tamil Supplementary blocks must be used.* Thus input methods will have to input only the Tamil Virama codepoint in an Extended Tamil text and it will automatically be displayed as a Grantha Virama glyph when it is used with Tamil Extended consonants. Similarly only the Tamil codepoints for the first and fifth (and in the case of JA, third) members of the consonant classes are to be used and they will automatically be displayed as the Grantha equivalents when used with Tamil Extended vowel signs.

In passing and since it is somewhat relevant here, I mention that on page 11 of my proposal, I have mentioned that the letters TA and NA in Tamil are identical in Grantha and Tamil, and the same is true for JA and HA. On later reflection, I found that this is not strictly true. Many printings distinguish between these characters in the two scripts by showing descenders in Tamil and in contrast limiting the characters to the baseline in Grantha, so:

த 𑌤 ந 𑌨 ஜ 𑌜 ஹ 𑌹

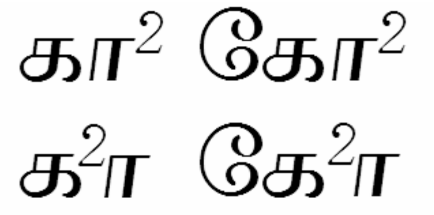
The obvious reason for this is that in Grantha one has stacking requirements and descenders are a hindrance to that. While this contrast is not consistent in the written forms of these scripts, it does exist in authoritative printings, and since it would be appropriate for a computer implementation of a Grantha font to follow printed material, we must respect this distinction. While it is very slight and easy to overlook, it does exist consistently. Therefore it is not entirely true that the consonants TA, NA, JA and HA are identical between the scripts. These characters should be handled appropriately when Grantha vowel signs (represented by Tamil Extended codepoints) are applied to them.

Returning to the present topic, I conclude that there exist small variants in Extended Tamil writing and these should be handled as appropriate by fonts.

Characters with superscript digits

The Unicode chapter on Tamil currently (as of v5.2) and correctly notes (in page 289, §9.6) that the superscript digits 2, 3 and 4 are used with the non-nasal class consonants in Tamil KA, CA, TTA, TA and PA to represent the missing characters required for Tamil. It is further, however, prescribed that the appropriate superscript digits encoded as separate characters in Unicode should be used for this purpose. However, I submit that this does not permit of one-to-one conversion between Indic scripts and this form of writing Sanskrit using Tamil.

There is also a rendering issue. Currently if the superscript digits are to be used, those digits would be inserted after any vowel signs. Therefore they would be displayed after the vowel sign. This is a problem, especially when the vowel sign (or part thereof) stands to the right of the consonant. The 'problem' is that the digit semantically gravitates to the consonant glyph since it qualifies the meaning of that glyph, but current rendering systems would not reorder the superscript digits to appear before the vowel signs. Thus the actual display is as shown right-above whereas the desired one is as shown right-below.



Categorically changing the behaviour of rendering engines in this case (where superscript digits follow vowel signs) would also not be advisable, since someone may actually require the digits to appear after the vowel sign for some other purpose. Therefore I suggest that a particular font implementing Extended Tamil may provide, for the TAMIL EXTENDED consonants in the Tamil Supplementary block, glyphs showing the regular Tamil letters with appropriate superscript digits. Since rendering engines supporting Extended Tamil will anyhow have to support the placement of Tamil vowel signs with TAMIL EXTENDED consonant codepoints, the desired rendering can be achieved effortlessly this way.

Here I note that in this system of not importing Grantha-style characters at all to denote Sanskrit using the Tamil script, the usage of the superscript digits is not limited to the class consonants but is also required for characters like the anusvara, visarga, vocalic R etc which have no Tamil equivalents. There is a current implementation of this system (as of 2010-Mar-12) at http://tamilcc.org/thoorihai/thoorihai.php. The author of that website also discusses some alternatives in http://tamilcc.org/thoorihai/Manual.pdf. All these alternatives can be satisfactorily implemented in Extended Tamil using appropriate fonts.

Conclusion

I have attempted to provide a better description of Extended Tamil (previously called pseudo-Manipravalam) than that found in my Grantha proposal. Some minor implementation details may possibly need to be reviewed and revised. It is however clear that Extended Tamil *is* a valid writing system and *must* be supported by Unicode.

-o-o-o-