

Proposal to encode characters for Extended Tamil

Shriramana Sharma, jamadagni-at-gmail-dot-com, India

2010-Jul-10

This document requests the encoding of characters which are required to enable the Tamil script to support the writing of Sanskrit. It repeats some relevant sections from my Grantha proposal L2/09-372 and my feedback document L2/10-085. It also contains some more information and citations and is a formal proposal for the encoding of those characters.

§1. Introduction

It is well known that the Tamil script has an insufficient character repertoire to represent the Sanskrit language. Sanskrit can be and is written and printed quite naturally in most other (major and some minor) Indian scripts, which have the required number of characters. However, it cannot be written in plain Tamil script without some contrived extensions acting in the capacity of diacritical marks, just as it cannot be written in the Latin script without the usage of diacritical marks (as prescribed by ISO 15919 or IAST). Another option is to import written forms from another script (to be specific, Grantha, Tamil's closest Sanskrit-capable relative) to cover the remaining unsupported sounds.

We call this version of Tamil that has been extended to support Sanskrit as Extended Tamil. It should be noted that this is neither a distinct script from Tamil nor can it be considered a mixture of two scripts (Tamil and Grantha), because the 'grammar' of the script (i.e. the orthographic rules, such as those of forming consonant clusters) is still that of Tamil. Thus Extended Tamil may be likened to the IPA extensions to the Latin script to denote sounds that the Latin script does not natively denote or differentiate.

There exists a large amount of printed material in the Tamil script showing both kinds of Extended Tamil (i.e. one using diacritic marks and one importing foreign symbols) handling the Sanskrit-capability problem (while printings importing Grantha written forms are somewhat rare and mostly not of recent date). The present proposal is to encode in Unicode those characters that are needed to support Sanskrit writing in Tamil.

Since the apparent "mixture" of the Grantha script in Extended Tamil is only in that version of Extended Tamil which chooses importing over diacritics, the borrowing of Grantha written forms is optional. Therefore there should be no problem in considering Extended Tamil characters for encoding quite independent of the encoding of Grantha.

Currently (even as of 5.2) the Unicode standard (§9.6, p 289) prescribes the use of superscript characters 00B2, 00B3 and 2074 to handle Extended Tamil. However, we shall show that while the glyphs these characters supply may be indeed used as diacritics, there exist problems which necessitate the encoding of separate characters.

§2. Characters that are needed for Extended Tamil

The Tamil script does not have characters for vocalic R, RR, L and LL (dependent and independent). It has only one non-nasal consonant per class among the class consonants whereas the other Sanskrit-supporting scripts like Devanagari and Grantha have four, leaving three consonants missing per class. The exception is CA-class where two characters – CA and JA – are already present in Tamil and only CHA and JHA are missing. Tamil also does not have an anusika sign, anusvara, visarga, ardhavisarga, danda-s and avagraha.

However, since the danda-s from 0964 and 0965 and the ardhavisarga from 1CF2 (and the newly-proposed rotated version at 1CF3) are being used for all Indic scripts, these characters do not need to be duplicated for Extended Tamil.

Thus, to extend the Tamil script to represent Sanskrit, one needs to encode characters for vocalic R etc, the missing class consonants and then the anusika sign etc.

Thus those characters that need to be encoded for Extended Tamil are:

- 1) TAMIL EXTENDED LETTER VOCALIC R
- 2) TAMIL LETTER VOCALIC RR
- 3) TAMIL LETTER VOCALIC L
- 4) TAMIL LETTER VOCALIC LL
- 5) TAMIL VOWEL SIGN VOCALIC R
- 6) TAMIL VOWEL SIGN VOCALIC RR
- 7) TAMIL VOWEL SIGN VOCALIC L
- 8) TAMIL VOWEL SIGN VOCALIC LL
- 9) TAMIL LETTER KHA
- 10) TAMIL LETTER GA
- 11) TAMIL LETTER GHA
- 12) TAMIL LETTER CHA
- 13) TAMIL LETTER JHA
- 14) TAMIL LETTER TTHA
- 15) TAMIL LETTER DDA

- 16) TAMIL LETTER DDHA
- 17) TAMIL LETTER THA
- 18) TAMIL LETTER DA
- 19) TAMIL LETTER DHA
- 20) TAMIL LETTER PHA
- 21) TAMIL LETTER BA
- 22) TAMIL LETTER BHA
- 23) TAMIL SIGN ANUNASIKA
- 24) TAMIL SIGN SPACING ANUSVARA
- 25) TAMIL SIGN GRANTHA-STYLE VISARGA
- 26) TAMIL SIGN AVAGRAHA

§3. Comparison table

The following table shows the characters that are needed for the Sanskrit language in four different script systems. The first two, Devanagari and Grantha are self-evident. The third and fourth columns show one version each of the two major versions of Extended Tamil, which we will call liberal and conservative. The boxes corresponding to characters that need to be newly encoded for Extended Tamil have been shaded in.

Dev.	Gran.	ET-L	ET-C	Dev.	Gran.	ET-L	ET-C
अ	஁	அ	அ	्	ஃ	ஃ/ஃ	ஃ
आ	஁	ஆ	ஆ	ा	ா	ா	ா
इ	ஐ	இ	இ	ि	ி	ி	ி
ई	ஃ	ஈ	ஈ	ी	ீ	ீ	ீ
उ	உ	உ	உ	ु	ு	ு	ு
ऊ	ஊ	ஊ	ஊ	ू	ூ	ூ	ூ
ऋ	஋	஋	ᱠ'	ृ	ூ	ூ	ᱠ'

ऋ	ஃ	ஃ	ரு'	ஶ்ரீ	ஶ்ரீ	ஶ்ரீ	ஶ்ரீ'
ॠ	ள	ள	லு'	ஶ்ரீ	ள	ள	ஶ்ரீ'
ॡ	ள	ள	லு'	ஶ்ரீ	ள	ள	ஶ்ரீ'
ए	ஊ	ஏ	ஏ	ஶ்ரீ	ஶ்ரீ	ஶ்ரீ	ஶ்ரீ
ऐ	ஊ	ஊ	ஊ	ஶ்ரீ	ஶ்ரீ	ஶ்ரீ	ஶ்ரீ
ओ	ஊ	ஊ	ஊ	ஶ்ரீ	ஶ்ரீ	ஶ்ரீ	ஶ்ரீ
औ	ஊ	ஊ	ஊ	ஶ்ரீ	ஶ்ரீ	ஶ்ரீ	ஶ்ரீ
क	க	க	க	த	த	த	த
ख	வ	வ	க ²	த	த	த	த ²
ग	ம	ம	க ³	த	த	த	த ³
घ	வ	வ	க ⁴	த	த	த	த ⁴
ङ	ஊ	ஊ	ஊ	ந	ந	ந	ந
च	ச	ச	ச	ப	ப	ப	ப
छ	ச	ச	ச ²	ப	ப	ப	ப ²
ज	ஊ	ஊ	ஊ	ப	ப	ப	ப ³
झ	ஊ	ஊ	ஊ ²	ப	ப	ப	ப ⁴
ञ	ச	ச	ச	ம	ம	ம	ம
ट	ட	ட	ட	ய	ய	ய	ய

ர	஠	஠	஠ ²	ர	ர	ர	ர
ஓ	ஔ	ஔ	ஔ ³	ல	ல	ல	ல
஖	஗	஗	஗ ⁴	ங	ங	ங	ங
ஞ	ஞ	ஞ	ஞ	ஞ	ஞ	ஞ	ஞ
ச	ச	ச	ச/ச'	ச	ச	ச	ச
ஞ	ஞ	ஞ	ஞ	ஞ	ஞ	ஞ	ஞ
஠	஠	஠	஠	஠	஠	஠	஠'
஠:	஠:	஠:	஠:	஠:	஠:	஠:	஠:
஠	஠	஠	(஠)	஠	஠	஠	(஠)

There are two points to be noted about the above chart. One is that in ET-C (Extended Tamil conservative version), a character which has already been encoded, namely 0BB6 TAMIL LETTER SHA, may need to take a different glyph from the standard one shown in the Unicode code chart depending on the orthographic style of the user. The other is that in the same ET-C, the double avagraha (just a sequence of two avagraha-s) will need to be rendered as a single glyph (by a ligature mechanism). I mention this here just to note that separate characters are not being encoded for these glyphic variations in Extended Tamil.

§4. Variations in Extended Tamil

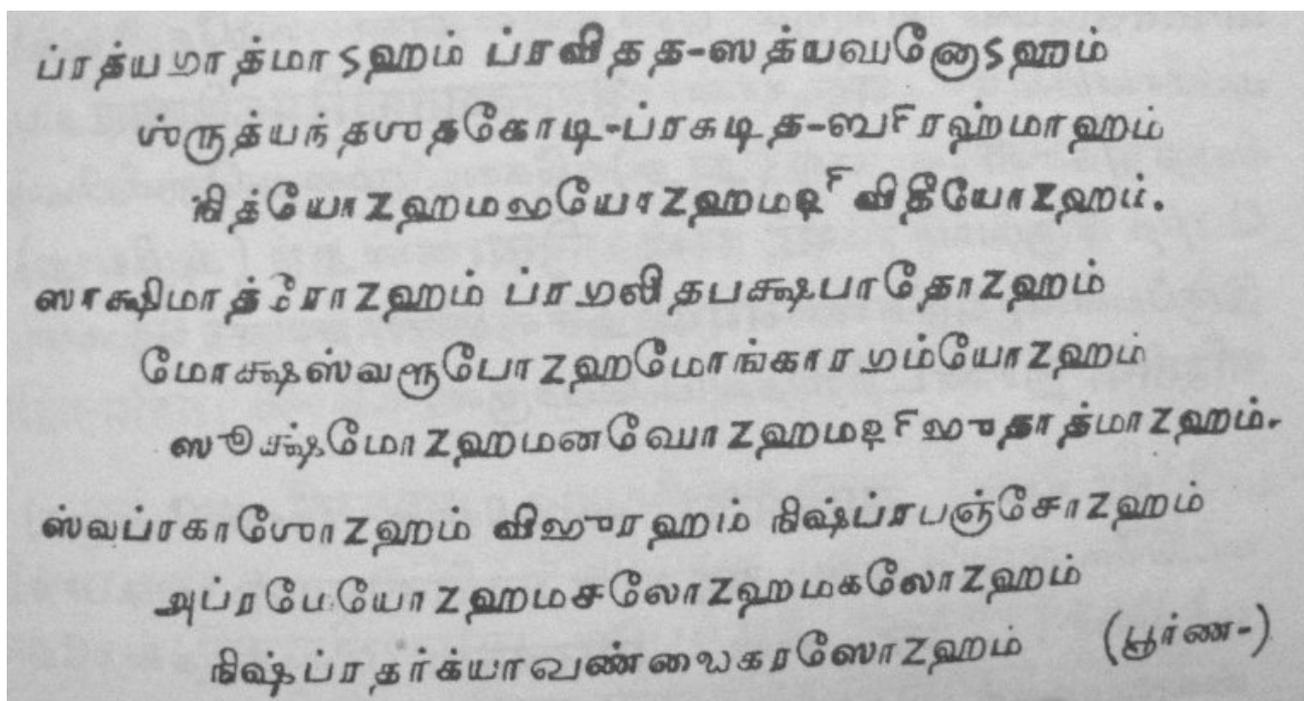
We mentioned the two versions of Extended Tamil – liberal and conservative. (I hasten to remark that there are no political overtones here!) These terms merely refer to the extent to which the new Extended Tamil characters borrow glyphs from Grantha.

The liberal version of Extended Tamil (“ET-L”) imports glyphs from Grantha for all the new characters that need to be encoded for Extended Tamil. The glyphs were shown in the table above. It also uses Grantha-style glyphs for consonants that carry Grantha-style vowel signs, even if those consonants are already part of the Tamil script.

The conservative version, on the other hand, chooses to employ existing Tamil glyphs with diacritic-like marks or other indication. The latter may sometimes be so conservative as to use for the already-encoded 0BB6 TAMIL LETTER SHA, instead of its default representative glyph, the glyph of 0B9A TAMIL LETTER CA with an apostrophe or other modification. (We have mentioned this above.)

In the liberal version (ET-L), we have observed two variants. One variant uses only Grantha-style consonant glyphs even in the presence of Tamil equivalents when the Grantha-style vowel signs for Vocalic R etc need to be attached and likewise uses the Grantha-style virama glyphs for Grantha-style consonants. Another uses only Tamil-style glyphs in these cases even with the Grantha-style vowel signs and uses the Tamil-style virama even with Grantha-style consonants.

The following are samples for the two variants:



p 65, Śiva Mānasa Pūjā, Kīrtana-s and Ātma Vidyā Vilāsa of Śrī Sadāśiva Brahmendra, 1951, Kamakoti Koshasthanam, Chennai

ஸுணுய்வம் மஹேதோ மே ஸ்ய
 ஜந்ம முஞ்ஜஸ்ய பெளர்ஷிகம்।
 ஸ புரா லில்ல ஏவாஸமூத்
 ஶந்யமௌௌ யதூர்யரஃ॥

p 26 of PDF, Bhoja Caritram by T S Narayana Sastri, 1916, <http://www.archive.org/stream/bhojacharitrama00sastgoog>

In the conservative version also, there are many variants. The selection of glyphs shown for ET-C in the table in §3 is my personal choice (with SHA taking the Grantha-style glyph) used by myself in publications edited and translated by myself, such as *Jagadguru Ratna Māla Stava of Sadāśiva Brahmendra and other related works* and *Saparyā Paryāya Stava of Sadāśiva Brahmendra and other works*, both to be published by Śrī Sadāśiva Brahmendra Bhakta Jana Samiti, Chennai. Another variant is seen at the Indic transliteration website <http://tamilcc.org/thoorihai/thoorihai.php> (retrieved 2010-Mar). The document <http://tamilcc.org/thoorihai/Manual.pdf> from that site discusses some more variants. The following samples from pages 28-31 of *Śiva Kavaca and Indrākṣī Stotra*, published in 1996 by Giri Trading Agency, Chennai, shows yet another particular variant:

த்யான-வதாதாகில-கர்மபீந்திச்-
 சிரம் சிதானந்த-நிமக்ந-சேதா: |
 ஷட்குரந்யாஸ-ஸமாஹிதாத்மா
 ஷைவேன சூர்யாத் கவசேன ரக்ஷாம் ||
 மாம் பாது தேவோ(அ)கில தேவதாத்மா
 ஸம்ஸாரகூபே பதிதம் கீபீரே |
 தந்நாம தீவ்யம் பரமந்தர்-மூலம்
 துஜேது மே ஸர்வமக்ம் ஹ்ருதீஸ்தீம் ||
 ஸர்வத்ர மாம் ரக்ஷது விச்வமூர்திர்-
 ஜீயோ திர்மயானந்த-கீனச் சிதாத்மா |

அனூரணீயா-லுருசுக்தி-ரேக:
 ஸ ஈச் வர: பாது ப்யாத்ஷேஷாத் ||
 யோ பூஸ்வரூபேண பீபீர்தி விச்வம்
 பாயாத்ஸ பூமேர்கீரிசே(அ)ஷ்டமூர்தி: |
 யோ(அ)பாம் ஸ்வரூபேண ந்ருணம் சுரோதி
 ஸஞ்ஜீவனம் ஸோ(அ)வது மாம் ஜீவேபீய: ||
 கல்பாவஸானே பீவ்ரணி தீக் தீவா
 ஸர்வாணி யோ ந்ருத்யதி பூரீலீல: |
 ஸ காவருத்ரோ(அ)வது மாம் த்வாகீனேர்
 வாத்யாதீ பீதே ரகீவாச்ச தாபாத் ||

ப்ரதீபஸ்தியுத்-கனகாவபீரஸோ
 வித்யா-வராபீதி-குடார-பாணி: |
 சதுர்முகீஸ்-தத்புருஷஸ் த்ரிநேத்ர:
 ப்ராச்யாம் ஸதீதோ ரக்ஷது மாமஜீஸ்ரம்
 குடார-கேடாங்குச்-குல்-டீக்கா-
 கபால-பாசர்-கூ-குண் தீதான: |
 சதுர்முகோ நீலநிஸ த்ரிநேத்ர:
 பாயாதீகோரோ திசி தீக்ஷிணஸ்யாம் ||
 குந்-தேந்நுச்-ங்கீ-ஸ்பீடி-காவபீரஸோ
 வேதீசுமாலா-வரதீபயாங்க: |

தூர்யக்ஷச்-சதுர்வக்தர்-உருப்ரபீரவ:
 ஸதீயோ(அ)தீஜீரதோ(அ)வது மாம் பாதீஸ
 வராக்ஷமாலாபீய-டங்கஹஸ்த:
 ஸரோஜீ-கிஞ்ஜீல்க-ஸமானவர்ண: |
 த்ரிவோசனச் சாரு-சதுர்முகோமாம்
 பாயா-தீதீச்யாம் தீசி வாமதேவ: ||
 வேதீபீயேஷ்டாங்குச்-டங்க-பாசர்-
 கபாலடீக்காக்ஷ-குல்பாணி: |
 ஸிததீயுதி: பாஞ்சமுகோ(அ)வதான்மா-
 மீசான ஊர்தீவம் பரமப்ரகாச: ||

Finally on discussing variants I should remark that there are many imperfections in real-world books, such as not differentiating the consonant /m/ and the anusvara, using the glyph of Tamil NNNA for NA etc (as seen in the samples above). Some books even go to the atrocious (in my reaction as a Sanskrit expert) extent of using bold formatting to merely differentiate between the voiced and voiceless class consonants (with bold denoting voiced) and not differentiating between the aspirated and unaspirated forms thereof. These are imperfections, and cannot be considered legitimate variants in their own right.

§5. The need to encode Extended Tamil

While as shown above there do exist many variants, Extended Tamil is essentially one. It is an extension of the Tamil script to support Sanskrit, as said at the outset. The particular glyphs used for the additional characters being used with the native Tamil characters may differ. But whether Grantha-style glyphs are used for the additional characters or diacritic-like marks are used with native Tamil glyphs, the underlying characters are the same. I may liken this to the Old Italic and Brahmi situation where (as I am informed) there are many (seriously different) glyphic variants but the Old Italic and Brahmi scripts are encoded as one script each. Therefore Extended Tamil may be encoded as one with the variants being taken care of at the font level.

It may be asked what is the need for encoding Extended Tamil, when the existing Unicode recommendation is to use the characters 00B2, 00B3 and 2074. I however believe that that portion of Unicode does not do justice to the real complexity behind Extended Tamil. Some reasons are given below:

- 1) That recommendation disables (or at least makes difficult or less-than-elementary) one-to-one transliteration by computer of Sanskrit texts from Devanagari or other Indic scripts to Tamil.
- 2) It does not consider the anusvara, visarga, avagraha etc at all (or the vowels vocalic R etc) but only talks about “consonants”. These characters are not analysed by natives (or to my knowledge by others) as consonants.
- 3) It does not consider the problem of rendering pointed out by me in page 11 of my document L2/10-085 (Feedback to Dr Anderson's Grantha Summary). A good look at the samples for ET-C provided hereinbefore will show that the problem is genuine and cannot be resolved by existing means. (Note that in those samples it is not only the superscript digits 2, 3 and 4 but also the apostrophe which gets placed between consonants and their vowel signs.)
- 4) It does not even consider the existence of the variant of Extended Tamil using Grantha-style glyphs for the additional characters.

To maintain the recommendation, and yet address the problem of point 4 above, it may be suggested that after the encoding of the Grantha script, codepoints from the Grantha block may be used to achieve Grantha-style glyphs, but such a suggestion should be pronounced dead on arrival because it goes against the essence of Unicode. In Unicode one does not use different characters to handle glyphic variants but different fonts. I have also discussed

other problems with this option in page 6 of L2/10-085. (It is also highly probable that that the word-boundary problem I have mentioned there also exists with the current Unicode recommendation of using 00B2 because those characters are all GC=No.)

Therefore the cleanest solution is to encode separate characters, with an appropriate selection of representative glyphs (as done for Old Italic and Brahmi). It would solve all the problems outlined above and more. All the variants (in both the liberal and conservative versions) can be handled by appropriate fonts and smart font technologies. More is discussed in the rendering section below. I have also mentioned some other advantages to the separate encoding of Extended Tamil characters in page 8 of L2/10-085.

Finally I should say that I must also not forget the Saurashtra language, which also is written with Extended Tamil. As I neither know the language nor any of its experts, I can only go by the Saurashtra block code-chart for A880-A8DF and the corresponding description in TUS 5.2 pp 329-330. The only point that the above discussion centered on Sanskrit misses out from the Saurashtra situation seems to be the Saurashtra Haaru. However, it seems that it is an analog of the Tamil āytam, and hence perhaps may be transliterated by the existing Tamil āytam character 0B83 placed appropriately. Further, the Saurashtra language also apparently uses the short E and O sounds (which do not exist in Sanskrit) but these can easily be represented (in either ET-L or ET-C) by the existing encoded Tamil characters for those sounds.

Therefore, an encoding of Extended Tamil as described above should also be able to support the writing of the Saurashtra language using Tamil characters.

§6. Rendering

In general, all rendering rules are as in Tamil, since, as mentioned before the underlying ‘grammar’ (orthography) of Extended Tamil is still that of Tamil. The letters should function like the normal Tamil letters, and the spacing combining marks are all displayed to the right of their base. The avagraha is as in other scripts that have it.

6.1. General Category property

Of the characters to be newly encoded for Extended Tamil (see §2) 18 are independent letters (both vowels and consonants, GC=Lo), 4 are Indic vowel signs (GC=Mc), 2 other spacing combining marks (GC=Mc) and 1 avagraha (GC=Lo).

As for single remaining character, the anunasika sign, it is to be noted that in a variant of ET-C (as described by the Thoorihai PDF mentioned before), the anunasika sign is

transliterated by (the glyphic equivalent of) TAMIL MA + TAMIL VIRAMA + SUPERSCRIPT THREE. If the character TAMIL SIGN ANUNASIKA is given GC=Mn, then one must consider how this variant is to be implemented because the character should then properly get GC=Mc. My suggestion is that this character take GC=Mc.

For the variants where the character is non-spacing, the situation is to be handled like the TAMIL VOWEL SIGNS U/UU which both have GC=Mc but for all native Tamil consonants are effectively non-spacing as they ligate with their base. Going by this argument I have chosen to give this character GC=Mc.

6.2. Substitution rules

Within ET-L, there exist two major variants as described above. One system uses Grantha-style glyphs for even native Tamil consonant characters when they take Grantha-style Extended Tamil vowel signs. It also uses only the Grantha-style virama glyph with Grantha-style Extended Tamil consonants even though the virama character to be used in the Unicode representation of Extended Tamil is the Tamil virama. Another system does not use Grantha-style glyphs for native Tamil consonant characters and uses the Tamil-style virama glyphs for even Grantha-style consonants. These two versions may be handled by smart-font rendering by the turning on or off of the following two rules:

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

I have also already mentioned that in (at least one variant of) ET-C, the double avagraha would have to be handled as follows:

TAMIL EXTENDED SIGN AVAGRAHA →
 LEFT PARANTHESIS + TAMIL LETTER A + RIGHT PARANTHESIS
 TAMIL EXTENDED SIGN AVAGRAHA + TAMIL EXTENDED SIGN AVAGRAHA →
 LEFT PARANTHESIS + TAMIL LETTER AA + RIGHT PARANTHESIS

6.3. Consonant clusters

For consonant clusters in which Tamil Extended consonants are involved, there is no ligature formation except for K·SSA which is already present in Tamil. This ligature may be rendered Tamil-style or Grantha-style, the minute difference being in the bottom left quadrant of the glyph. There are no conjoining forms. While the question itself does not arise in ET-C, it does arise in ET-L where Grantha-style consonants are present. However,

except for the single ligature K:SSA, consonants are written with visible virama as appropriate (in one variant of ET-L with Grantha-style virama glyph for Grantha-style consonants not existing in Tamil and Tamil-style for native Tamil consonants).

§7. Collation and linebreaking

As the language being represented is Sanskrit, the Sanskrit collation order (described in detail in my Grantha proposal L2/09-372 §10) is to be followed. It is to be remembered that native Tamil characters and newly Tamil Extended characters which be naturally mixed up in the sorting order. The rules for line breaking are as in Tamil.

§8. Unicode character properties

8.1. Discussion

I have already discussed in L2/10-085 whether to encode these characters in the Tamil block or elsewhere. Since the positions in the Tamil block corresponding to the ‘missing’ characters from other blocks are yet empty, it would be very easy to simply fill in those codepoints. However, I strongly suspect that it would not be welcomed by some parties that are already asking for the removal of Grantha-style characters from the Tamil block (which is of course absurd). Therefore, to avoid such a problem, these characters may be encoded in a separate “Tamil Extended” block (just like “Devanagari Extended”), another name for the Tamil Supplementary block I requested in L2/09-317. They may be placed sequentially.

Regarding the character names, I felt that it is better to name these characters beginning with the words TAMIL EXTENDED and not just TAMIL. However, as per instructions from the UTC I have used just TAMIL. For the anusvara and visarga, however, I had to introduce the adjectives SPACING and GRANTHA-STYLE to differentiate them from the existing “anusvara” and “visarga” characters in the Tamil block 0B82 and 0B83.

There is only one point about naming I would like to mention, however. Everywhere else (Devanagari, Gujarati etc), the anunasika sign has been named <SCRIPT> SIGN CANDRABINDU and not <SCRIPT> SIGN ANUNASIKA. Here, however, I request for the anunasika character to be named anunasika and not candrabindu. The reason is that, as I have mentioned above, there are many variants to Extended Tamil, and not all of them use the candrabindu for the anunasika, as I have remarked above in §6.1. Therefore, it would not be appropriate this character CANDRABINDU, and therefore I ask for it to be named TAMIL EXTENDED SIGN ANUNASIKA.

The script property of these characters should be script=tamil to enable their painless use among Tamil characters, since, as is being several times repeated, it is still the Tamil script which is but being extended.

8.2. Properties listing

```

xx00;TAMIL SIGN ANUNASIKA;Mc;0;L;;;;N;;;;;
xx01;TAMIL SIGN SPACING ANUSVARA;Mc;0;L;;;;N;;;;;
xx02;TAMIL SIGN GRANTHA-STYLE VISARGA;Mc;0;L;;;;N;;;;;
xx03;TAMIL LETTER VOCALIC R;Lo;0;L;;;;N;;;;;
xx04;TAMIL LETTER VOCALIC RR;Lo;0;L;;;;N;;;;;
xx05;TAMIL LETTER VOCALIC L;Lo;0;L;;;;N;;;;;
xx06;TAMIL LETTER VOCALIC LL;Lo;0;L;;;;N;;;;;
xx07;TAMIL LETTER KHA;Lo;0;L;;;;N;;;;;
xx08;TAMIL LETTER GA;Lo;0;L;;;;N;;;;;
xx09;TAMIL LETTER GHA;Lo;0;L;;;;N;;;;;
xx0A;TAMIL LETTER CHA;Lo;0;L;;;;N;;;;;
xx0B;TAMIL LETTER JHA;Lo;0;L;;;;N;;;;;
xx0C;TAMIL LETTER TTHA;Lo;0;L;;;;N;;;;;
xx0D;TAMIL LETTER DDA;Lo;0;L;;;;N;;;;;
xx0E;TAMIL LETTER DDHA;Lo;0;L;;;;N;;;;;
xx0F;TAMIL LETTER THA;Lo;0;L;;;;N;;;;;
xx10;TAMIL LETTER DA;Lo;0;L;;;;N;;;;;
xx11;TAMIL LETTER DHA;Lo;0;L;;;;N;;;;;
xx12;TAMIL LETTER PHA;Lo;0;L;;;;N;;;;;
xx13;TAMIL LETTER BA;Lo;0;L;;;;N;;;;;
xx14;TAMIL LETTER BHA;Lo;0;L;;;;N;;;;;
xx15;TAMIL SIGN AVAGRAHA;Lo;0;L;;;;N;;;;;
xx16;TAMIL VOWEL SIGN VOCALIC R;Mc;0;L;;;;N;;;;;
xx17;TAMIL VOWEL SIGN VOCALIC RR;Mc;0;L;;;;N;;;;;
xx18;TAMIL VOWEL SIGN VOCALIC L;Mc;0;L;;;;N;;;;;
xx19;TAMIL VOWEL SIGN VOCALIC LL;Mc;0;L;;;;N;;;;;

```

§9. References

1. *Śiva Mānasa Pūjā, Kīrtana-s and Ātma Vidyā Vilāsa of Śrī Sadāśiva Brahmendra*, 1951, Kamakoti Koshasthanam, Chennai
2. *Bhoja Caritram* by T S Narayana Sastri, 1916,
<http://www.archive.org/stream/bhojacharitrama00sastgoog>
3. <http://tamilcc.org/thoorihai/Manual.pdf>, retrieved 2010-Mar
4. *Śiva Kavaca and Indrākṣī Stotra*, 1996, Giri Trading Agency, Chennai

§10. Official Proposal Summary Form

A. Administrative

1. Title

Proposal to encode characters for Extended Tamil

2. Requester's name

Shriramana Sharma

3. Requester type (Member body/Liaison/Individual contribution)

Individual contribution

4. Submission date

2010-Jul-10

5. Requester's reference (if applicable)

6. Choose one of the following:

6a. This is a complete proposal

Yes, except for the actual code points which should be allotted based on the answer to L2/09-317.

6b. More information will be provided later

No.

B. Technical – General

1. Choose one of the following:

1a. This proposal is for a new script (set of characters)

No. This is a proposal for extending the existing Tamil script.

1b. Proposed name of script

1c. The proposal is for addition of character(s) to an existing block

Yes.

1d. Name of the existing block

Tamil Extended (to be allocated based on the request in L2/09-317)

2. Number of characters in proposal

26 (twenty-six)

3. Proposed category (A-Contemporary)

Category A.

4a. Is a repertoire including character names provided?

Yes.

4b. If YES, are the names in accordance with the “character naming guidelines” in Annex L of P&P document?

Yes.

4c. Are the character shapes attached in a legible form suitable for review?

Yes.

5a. Who will provide the appropriate computerized font (ordered preference: True Type, or PostScript format) for publishing the standard?

Shriramana Sharma.

5b. If available now, identify source(s) for the font (include address, e-mail, ftp-site, etc.) and indicate the tools used:

6a. Are references (to other character sets, dictionaries, descriptive texts etc.) provided?

Yes.

6b. Are published examples of use (such as samples from newspapers, magazines, or other sources) of proposed characters attached?

Yes.

7. 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.

8. 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.

See detailed proposal.

C. Technical – Justification

1. Has this proposal for addition of character(s) been submitted before? If YES, explain.

No.

2a. 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. The proposer himself is a member of the user community.

2b. If YES, with whom?

Dr Mani Dravid, lecturer at Madras Sanskrit College, Chennai. Dr Venugopala Sharma, lecturer at Shri Jayendra Saraswathi Ayurveda College, Nazarathpet, Kanchipuram. Vinodh Rajan, Chennai.

2c. If YES, available relevant documents

None specifically. Mode of contact was personal conversation.

3. Information on the user community for the proposed characters (for example: size, demographics, information technology use, or publishing use) is included?

Tamilians in their lakhs residing in Tamil Nadu and elsewhere who read Sanskrit (religious) texts.

4a. The context of use for the proposed characters (type of use; common or rare)

Common in the context of Sanskrit religious books printed in Tamil Nadu.

4b. Reference

5a. Are the proposed characters in current use by the user community?

Yes, often.

5b. If YES, where?

In publications in Tamil Nadu of Sanskrit religious texts.

6a. After giving due considerations to the principles in the P&P document must the proposed characters be entirely in the BMP?

No.

6b. If YES, is a rationale provided?

6c. If YES, reference

7. Should the proposed characters be kept together in a contiguous range (rather than being scattered)?

Yes, since it is only logical to keep mutually related characters together.

8a. Can any of the proposed characters be considered a presentation form of an existing character or character sequence?

No.

8b. If YES, is a rationale for its inclusion provided?

8c. If YES, reference

9a. Can any of the proposed characters be encoded using a composed character sequence of either existing characters or other proposed characters?

No.

9b. If YES, is a rationale for its inclusion provided?

9c. If YES, reference

10a. Can any of the proposed character(s) be considered to be similar (in appearance or function) to an existing character?

In some glyphic variants, the characters resemble Grantha characters and in others they resemble Tamil characters in combination with superscript digits

10b. If YES, is a rationale for its inclusion provided?

Yes.

10c. If YES, reference

The rationale is that these characters are used as part of the (Extended) Tamil script and are different from those glyphically similar characters in function and behaviour. The existence of many glyphic variants (as in the case of Old Italic and Brahmi) is also a justification for separate encoding.

11a. Does the proposal include use of combining characters and/or use of composite sequences (see clauses 4.12 and 4.14 in ISO/IEC 10646-1: 2000)?

No.

11b. If YES, is a rationale for such use provided?

11c. If YES, reference

11d. Is a list of composite sequences and their corresponding glyph images (graphic symbols) provided?

No.

12a. Does the proposal contain characters with any special properties such as control function or similar semantics?

No.

12b. If YES, describe in detail (include attachment if necessary)

13a. Does the proposal contain any Ideographic compatibility character(s)?

No.

§11. Code chart

As mentioned in §5, due to the wide glyphic variation of these characters (such as in Old Italic and Brahmi) a particular set of representative glyphs should be chosen for Extended Tamil. I suggest that the glyphs corresponding to (a variant of) ET-C be chosen, because the current TUS description of the Tamil script refers to ET-C only.

xx0 xx1

0	ம் ³	த ³
1	ம் ²	த ⁴
2	ஃ	ப ²
3	ரு ²	ப ³
4	ரு ²	ப ⁴
5	லு ²	(அ)
6	லா ²	்ரு ²
7	க ²	்ரு ²
8	க ³	்லு ²
9	க ⁴	்லா ²
A	சு ²	
B	ஐ ²	
C	ழ ²	
D	ழ ³	
E	ழ ⁴	
F	த ²	

This choice of ET-C-style glyphs would also avoid any problems with those disliking Grantha-style glyphs being used in Tamil. A note should however be added in the code chart to the effect that there are many other (and quite dissimilar) glyphic variants and that the given glyphs are only indicative. It should also be noted that we do not provide any decompositions of these characters to other similar-looking characters exactly because of the presence of glyphic variants. A decomposition based on (one variant of) ET-C would not work in ET-L and even would not work in another variant of ET-C.

The (mandatory) chart description now follows:

Various Characters:

xx00	◌̣̣̣ ³	TAMIL SIGN ANUNASIKA
xx01	◌̣̣̣ ²	TAMIL SIGN SPACING ANUSVARA
xx02	◌̣̣̣̣	TAMIL SIGN GRANTHA-STYLE VISARGA

For ardhavisarga, use 1CF2 ◌̣̣̣̣ VEDIC SIGN ARDHAVISARGA or 1CF3 ◌̣̣̣̣ VEDIC SIGN ROTATED ARDHAVISARGA.

Independent Vowels:

For independent vowels not present here, use from the Tamil block 0B85-0B94.

xx03	◌̣̣̣ ²	TAMIL LETTER VOCALIC R
xx04	◌̣̣̣̣ ²	TAMIL LETTER VOCALIC RR
xx05	◌̣̣̣̣ ²	TAMIL LETTER VOCALIC L
xx06	◌̣̣̣̣̣ ²	TAMIL LETTER VOCALIC LL

Consonants:

For consonants not present here, use from the Tamil block 0B95-0BB9.

xx07	◌̣̣̣ ²	TAMIL LETTER KHA
xx08	◌̣̣̣ ³	TAMIL LETTER GA
xx09	◌̣̣̣ ⁴	TAMIL LETTER GHA
xx0A	◌̣̣̣̣ ²	TAMIL LETTER CHA
xx0B	◌̣̣̣̣̣ ²	TAMIL LETTER JHA
xx0C	◌̣̣̣̣ ²	TAMIL LETTER TTHA
xx0D	◌̣̣̣̣̣ ³	TAMIL LETTER DDA

xx0E	ட ⁴	TAMIL LETTER DDHA
xx0F	த ²	TAMIL LETTER THA
xx10	த ³	TAMIL LETTER DA
xx11	த ⁴	TAMIL LETTER DHA
xx12	ப ²	TAMIL LETTER PHA
xx13	ப ³	TAMIL LETTER BA
xx14	ப ⁴	TAMIL LETTER BHA

Various Signs:

xx15	(அ)	TAMIL SIGN AVAGRAHA
------	-----	---------------------

Dependent Vowel Signs:

For dependent vowel signs not present here, use from the Tamil block 0BBE-0BCC.

xx16	்ரு ²	TAMIL VOWEL SIGN VOCALIC R
xx17	்ரு ²	TAMIL VOWEL SIGN VOCALIC RR
xx18	்லு ²	TAMIL VOWEL SIGN VOCALIC L
xx19	்லு ²	TAMIL VOWEL SIGN VOCALIC LL

Various Signs:

For the virama, use from the Tamil block 0BCD.

