Response to the ICTA's doc L2/14-048 on Tamil fractions and symbols

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

2014-Sep-10

§1. Background

N4430 L2/13-047 is the standing proposal for the encoding of 55 Tamil characters denoting fractions and symbols. These characters are currently in Stage 5. As can be seen in the 4th edition's 2nd amendment draft N____ L2/14-142 pp 5 and 40, 6 of these characters are proposed for the Tamil BMP block and the remaining 49 for a new Tamil Supplement block in the SMP.

The rationale provided in the original proposal (L2/12-231 p 60 §9.3.2) for thus separating the characters is as follows: the BMP already contains symbols for clerical use such as 0BF7 TAMIL CREDIT SIGN and 0BFA TAMIL NUMBER SIGN and so on. Thus it would be meaningful to place characters for clerical use along with them. Other symbols and all of the fractions are proposed for the Tamil Supplement block in the SMP.

In the meanwhile, in L2/14-048, the ICTA of Sri Lanka "strongly discourage"s the encoding of any (further) Tamil symbol characters in the BMP. They also "discourage" encoding fractions.

§2. Tamil fractions

§2.1. Concerns about never-ending set

The ICTA "discourage"s encoding fractions on the ground that it would lead to a "never-ending set of additions". Why the possibility of more written forms for fractions existing should prevent encoding those attested already is not made clear, however.

My proposal for these fractions and symbols is based on attestations from quite a number of reference texts, a significant number of which were in fact provided by G Balachandran of the ICTA as noted with thanks at the very outset of my proposal. The set of fractions is quite consistent throughout these texts. It does *not* showing signs of a runaway imagination producing arbitrary symbols for more and more random fractions. On the contrary, it is well defined with two separate sets of fractions for denoting area/volume (based on $^1/_5$) and currency (based on $^1/_4$). The former only goes down as far as $^1/_{320}$, and the latter as $^1/_{64}$, the ratio of an old paisa to a rupee. Further lower fractions are depicted as scaled-down versions of these. I have explained all this in L2/12-231 §4.4 and §4.5. For convenience of the WG2, a summary is provided in the Appendix.

Thus the concern of a "never-ending" set of additions is quite misplaced.

*§2.2. Possibility of notation using font tables**

The ICTA suggests that the fractions be handled by substitution rules in fonts as follows:

க/௳௨ → ௸

i.e.

0BE7 TAMIL DIGIT ONE + 2044 FRACTION SLASH** + 0BE9 TAMIL DIGIT THREE + 0BE8 TAMIL DIGIT TWO

→ TAMIL FRACTION ONE THIRTY-SECOND

However it is my understanding that Unicode encodes characters based on their written form, not based on their meaning. Therefore written forms which are *not* presentation variants of or otherwise orthographically derived from the nominal glyphs of one or more characters are *not* handled by font substitution rules in Unicode.

Fractions in international format like $^{1}/_{32}$ are indeed just a presentation form of the components 1, 3 and 2. However this is not the case in Tamil fractions. For instance, ௸ (with value 1 ÷ 32) is neither a presentation variant of nor orthographically derived from the shapes of the nominal components க (1) and ௳௨ (32).

As such, it is not appropriate to represent the uniquely shaped Tamil fractions such as ௸ using a sequence of the regular Tamil digits by font substitution.

The sequence suggested above would in fact be used for something like: $^{க}/_{௳௨}$ (which uses the Tamil digits for 1, 3 and 2 in the modern format for fractions) and not for the distinct ௸. Thus distinct written forms for fractions such as ௸ should be encoded separately.

This is in accordance with the ICTA's own earlier recommendation in L2/09-416:

> *We recommend that all the known Tamil fractions be encoded together (with spaces left for any others which may be identified later)*

§3. Tamil symbols

The ICTA "strongly discourages" encoding new symbols in the BMP, permitting that symbols "may be encoded in the SMP" only. The only reason provided is that these are "not used". However, the supposed distinction that the BMP is for "living" scripts and the SMP is for "dead" scripts no longer exists. Many historical characters are encoded in the BMP and many contemporary (albeit minority) scripts are encoded in the SMP. Given this, I feel "not in current use" is a non-argument and other factors should be considered for where to encode these rarely used characters.

---

* This section is the same as that in L2/14-160 but I'm repeating it here for the WG2 record.

** The ICTA doesn't specifically mention U+2044 FRACTION SLASH against U+002F SOLIDUS but it seems it would be the logical choice for such a sequence.

AFAICS extant CTL/smart font technology does not distinguish between BMP and SMP codepoints. And in this age of terabyte drives, storage requirements are not really an issue either way, especially for rare use characters. Thus no forcible case can be made for recommending either the BMP or SMP for these characters based on these factors alone.

§3.1. The case for using the free space in the BMP Tamil block

In the 128-cell BMP Tamil block, 72 characters are currently encoded. 56 cells remain empty. Of these, the 31 in the first five columns are best left alone as they mostly come between existing Tamil letters parallel to other Indic codepoints in the ISCII layout. 25 cells remain in the last three columns. Of these, 0BE4 and 0BE5 are parallel to the danda-s at 0964 and 0965 and marked as reserved. It is highly unlikely that Tamil dandas will be encoded and these can probably be reused if really necessary. But even discounting these, there are 23 empty cells in the last three columns.

The question now arises whether it is wise to waste all this free space. If this space is used better, the SMP Tamil Supplement block, currently at four columns from 11FC0-11FFF, can be trimmed down even to two columns and relocated to make better use of SMP space, say between the already published Modi and Takri blocks from 11660-1167F.

Space in the SMP is now at a premium as it is fast being filled up by other (historical) scripts. It would be meaningful to free it up as far as possible for such use. OTOH, the empty space in the Tamil BMP block cannot be used for anything other than Tamil characters.

Possible future uses of Tamil BMP space I have heard suggested include “diacritic marks for pronunciation dictionaries” and “characters produced by script reform”, but one wonders if such hypothetical needs from an unknown future time warrant *current* disuse of empty codepoints which are otherwise immediately usable for the proposed characters. If at all such needs arise and cannot be met by the remaining BMP cells, a Tamil Supplement B block may be allocated later on.

§3.2. One possible option for better utilization of the BMP Tamil free space

The 23 empty cells in the last three columns of the BMP Tamil block (excluding the “danda spaces”) can be filled up as follows:

Since the symbols already encoded in the BMP Tamil block are for clerical use, the standing proposal recommends encoding newly identified clerical use characters in the BMP block to fill the 5-cell hole at 0BFB-0BFF with one extra character at 0BDF. The following may be done in addition:

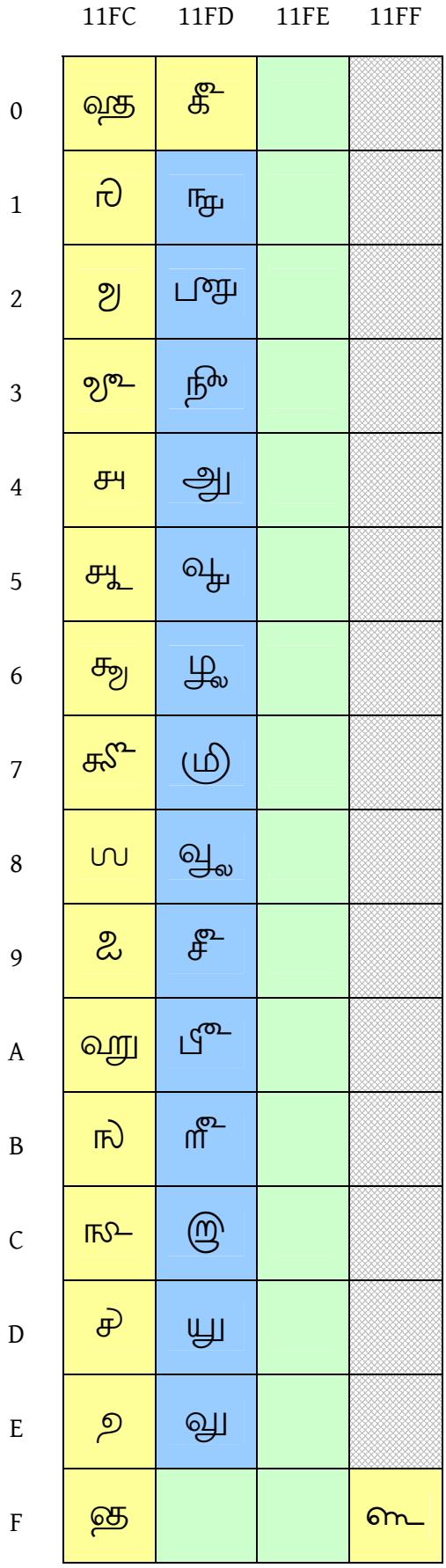
a) Currency symbols from 11FD8-11FDD move to fill 6-cell hole at 0BD1-0BD6. **b)** Measures from 11FD1-11FD7 move to fill 7-cell hole at 0BD8-0BDE. **c)** Symbols for weight etc from 11FDE-11FE1 move to fill 4-cell space at 0BE0-0BE3. **d)** 32 remaining SMP characters can be defragmented.

BMP Tamil block layout per suggestion in §3.2 to make better use of free space

	0B8	0B9	0BA	0BB	0BC	0BD	0BE	0BF
0		ஐ		ர	◌ை	ஒ	றை	ய
1				ற	◌ை	டை	கை	ர
2	◌ஃ	ஒ		ல	◌ை	ஆை	ஙை	ற
3	ஃ	ஒ	ண	ள		கை	ஙை	ல
4		ஒள	த	ழ		ப		ள
5	அ	க		வ		ை		ழ
6	ஆ			ஶ	ெ	வை	ஂ	ப
7	இ			ஷ	ே	◌ள	க	ஈ
8	ஈ		ந	ஸ	ை	ஐ	உ	ஷ
9	உ	ங	ன	ஹ		ஸை	ங	ஸை
A	ஊ	ச	ப		ொ	ஸ	ச	ஸ
B					ோ	ஸி	ஞ	ஹ
C		ஜ			ௌ	ஹ	ஜ	ஹை
D					◌்	ஹ	ட	ஜை
E	எ	ஞ	ம	◌ா		ஹ	அ	ஜ
F	ஏ	ட	ய	◌ி		ஹ	க	ஆை

legend: dark grey = has cognate Indic characters; yellow = already proposed; blue = newly moved

SMP Tamil Supplement block layout per suggestion in §3.2 to make better use of free space



legend: yellow = already proposed; blue = newly moved ; green = new empty space

Of course, other possibilities for reallocation (somewhat) different from the above do exist. In any case, it would be best to not break up the groups of related characters explained in the original proposal and seen in the draft amendment doc L2/14-142 pp 41 and 42.

§3.3. The case for leaving some breathing space in both blocks

It is clear that not all the newly proposed characters can be meaningfully fitted in within the BMP block itself. Thus an SMP block is unavoidable. Given this, a pertinent question here is as to what happens to the SMP block after some of its contents are moved to the BMP and the remaining characters are defragmented.

In the above scheme, 32 remaining SMP characters fit perfectly within two columns, and so the block can be shrunk to two columns and moved to fill two-column holes such as 11660-1167F between Modi and Takri as mentioned above. However, this does not permit of future additions. While hypothetical requirements as mentioned in p 3 may be ignored for the present, the probability of future proposals for other historic characters is greater, since we already have identified some requiring further attestation for encoding (see L2/12-231 p 21 §4.9).

Of course, the possibility of a future Supplement 'B' block always exists, but it would be useful to avoid block fragmentation in the future if there is at least one (partially) free column in the already proposed Supplement block. To justify such a column, at least one group of characters previously suggested to move to the BMP can be retained in the SMP block itself. This would leave sufficient breathing space in both the BMP and SMP blocks for any possible future characters which would warrant placement along with the currently proposed BMP/SMP characters.

§3.4. The case for leaving the BMP Tamil block alone

Discounting the "these are not used" argument which is not really valid as discussed before, there still exists one, albeit somewhat non-technical, argument for leaving the BMP Tamil block alone.

Apparently when font foundries are commissioned to make fonts for a script, especially by Governments for public usage, or by makers of OS/wordprocessing/publishing software, the clients would like glyphs for the whole block to be provided. However, font foundries would really like to *not* be obliged to spend human-hours on so many glyphs (presumably, in multiple styles and weights for each family) for such characters of highly limited usage, but would find it difficult to convince their clients that these glyphs aren't really needed for normal contemporary use of the script (or to increase the payment overmuch for the significant additional work these glyphs entail). It has been suggested that relegating the characters to an SMP block will help take the pressure off these foundries to support these characters.

OTOH, if the contract is by script and not by block, then making a separate block, even if in the SMP, will not help (assuming that the client knows about the extra block of course).

In any case, the terms of business contracts do not really form a technical argument, but a practically meaningful solution to this problem would be clearly labeling these characters in the names list as "historical and not in current usage". This should help alleviate the font-makers' problem while not wasting valuable BMP space in overreaction to a non-technical argument.

§4. Conclusion

Not encoding any of these characters, especially the fractions, is *not* an option. As to *where* to encode them, I have merely discussed in this document the pros and cons of the various options and their ramifications. I have no personal preference that I would insist upon except that related characters should be placed together per the groups shown in the original proposal and as currently reflected in the draft amendment. That is but logical, I hope.

I would also *like* to see TAMIL PUNCTUATION END OF TEXT currently proposed for 11FFF retained at the end of the block, however the block may be resized or the characters relocated/consolidated within it, but that's just my poetic (not technical) sense talking! :-)

After considering all this and any other input, if the UTC/WG2 decides that any relocation should be done, I'll be happy to submit a document formally requesting it for the record.

-o-

Appendix – The systematic nature of Tamil fractions

(A summary of L2/12-231 §4.4 and §4.5)

Basic fractions

¾ ஞ, ½ ஏ, ¼ வ, ⅕ சி

Fractions for area/volume, based on ⅕ சி

⅕ சி × ¼ வ, ½ ஏ, ¾ ஞ → 1/20 ப, 1/10 இ, 3/20 ஊ

The least of these is: 1/20 ப.

1/20 ப × ¼ வ, ½ ஏ, ¾ ஞ → 1/80 ஆ, 1/40 சு, 3/80 கூ

The least of these is: 1/80 ஆ.

1/80 ஆ × ¼ வ, ½ ஏ, ¾ ஞ → 1/320 �ஹ, 1/160 ஶ, 3/320 (*)

Here we do not have attestation for a distinct form for 3/320 presumably due to rare use.

Fractions for currency, based on ¼ வ

¼ வ × ¼ வ, ½ ஏ, ¾ ஞ → 1/16 �full, 1/8 ஷ, 3/16 ஸ

The least of these is: 1/16 �full.

1/16 �full × ¼ வ, ½ ஏ, ¾ ஞ → 1/64 ஃபு, 1/32 சூ, 3/64 ஸு

Glyphic comparison of area/volume vs currency fractions with ratio 4:5

We can see that many currency fractions were glyphically derived from area/volume fractions:

3/20 ஊ, 3/80 கூ, 1/40 சு, 1/80 ஆ

3/16 ஸ, 3/64 ஸு, 1/32 சூ, 1/64 ஃபு

Lesser fractions

Prefixing கீ scales the above fractions down to 1/320 of their nominal value. These are *kīḻ* i.e. "low" fractions. The lowest value in this scale is கீஹ i.e. 1/320 × 1/320. Below this we even have *immi* ("tiny"), *nuṇmai* ("fine") and *ciṉṉa* ("small") scales but without attested prefixes.

Conclusion

From the above evidence it should be clear that the set of fractions the Tamils chose to represent was very systematic and certainly not haphazard or "never-ending".

Note that Malayalam also has a very similar set of fractions since the Malayalam/Tamil regions have been socio-culturally very close. These are accepted for encoding as per N4429 L2/13-051R. Telugu fractions are encoded at 0C78-0C7E, Oriya ones at 0B72-0B77 and generic North Indic ones at A830-A835. Likewise, the attested Tamil fractions should also indeed be encoded.

-o-o-o-