L2/10-425

Comments on L2/10-256R & L2/10-363

Vinodh Rajan, vinodh@virtualvinodh.com

Tamil Script lacks voiced and aspirated consonants as well as signs of vocalic vowels. So, people who wished to perfectly represent other Indic languages in Tamil script resorted to the usage of superscript/subscript numerals with the Tamil consonants to represent the missing consonants. Some, who were more traditional, imported the consonants from the Grantha script to fill in the gaps in the Tamil Script. Similar measures were taken for the vocalic vowels as well. Since, enough has been discussed about these issues in L2/10-256R, I stop here and move to other issues.

Encoding *extended* Tamil in BMP

The document L2/10-256R submitted by Shriramana proposes a set of characters to be encoded in the SMP, to support the representation of other Indic languages in Tamil. Encoding such characters is also very necessary to facilitate a Pan-Indic transliteration to Tamil Script.

As noted in the above revised document, the term “extended” was itself deemed unnecessary by the Unicode committee. Since the represented script is Tamil, and there was absolutely no necessity to add an extra qualifier.

Considering the above fact, I am not sure why the characters are proposed in the SMP. It is more convenient to have them in the BMP. The code-points corresponding to the voiced/aspirated consonants and vocalic vowels in the other Indic scripts are currently marked as reserved in the Tamil code chart. When there are already reserved code points for the missing characters in the Tamil code block, it seems more logical to make use of those code points and encode the proposed characters in the BMP itself.

Arguments for puritanism [The Tamil code chart already has other supposedly ‘non-Tamil’ consonants SSA, SHA, SA, HA, JA] and other political reasons are not good arguments for encoding the characters in the SMP. Unless any valid technical reasons are proposed, it would be a very bad decision to encode the characters in SMP citing political arguments. Unicode Consortium should give more weightage to the Technical arguments rather than the Political arguments presented by some parties.

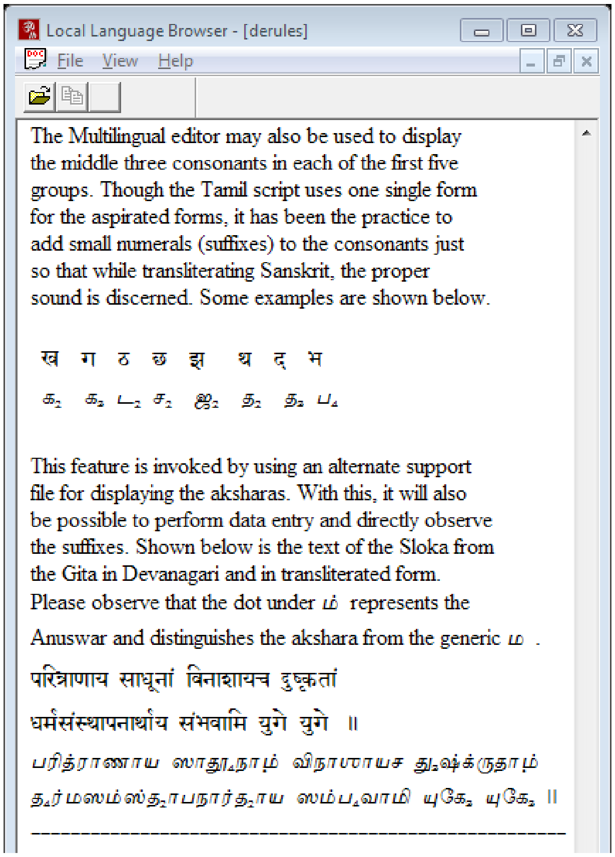
The Unicode code charts of the Indic scripts have a similar layout. Given an Indic Character, it is possible to get the corresponding character in other scripts [if they do exist] based on the layout. A simple bit-level manipulation on the character is sufficient to do a transliteration.

Many Indic transliteration programs like *Eemata* [http://eemaata.com/indic2indic/index.php], *Girgit* [http://girgit.chitthajagat.in/] etc. make use of this similarity of the Indic script encoding to do a one-to-one transliteration between the Indic Scripts. It will be easy for the programs to adapt the new characters, if they are encoded in the already reserved code points. Because, the characters would be in the typical pattern, similar to the other Indic code point blocks. Encoding the characters in SMP would unnecessarily complicate the technical implementation.

Hence, the Unicode consortium must strongly consider encoding the proposed characters in the BMP, rather than the SMP.

Further comments on L2/10-256R

It must be further noted that Indian Institute of Technology - Madras's [IIT-M] multilingual editor [http://acharya.iitm.ac.in/software/iitmed.php] supports the additional "Tamil" characters since late 90's. The software allows entering the Tamil characters and transliteration between other Indic scripts to Tamil script employing the additional characters. It shows that Tamil with digits have been recognized and supported as a part of Multi Lingual computing in India since the past decade.



The Multilingual editor may also be used to display the middle three consonants in each of the first five groups. Though the Tamil script uses one single form for the aspirated forms, it has been the practice to add small numerals (suffixes) to the consonants just so that while transliterating Sanskrit, the proper sound is discerned. Some examples are shown below.

ख ग ठ छ झ थ द भ

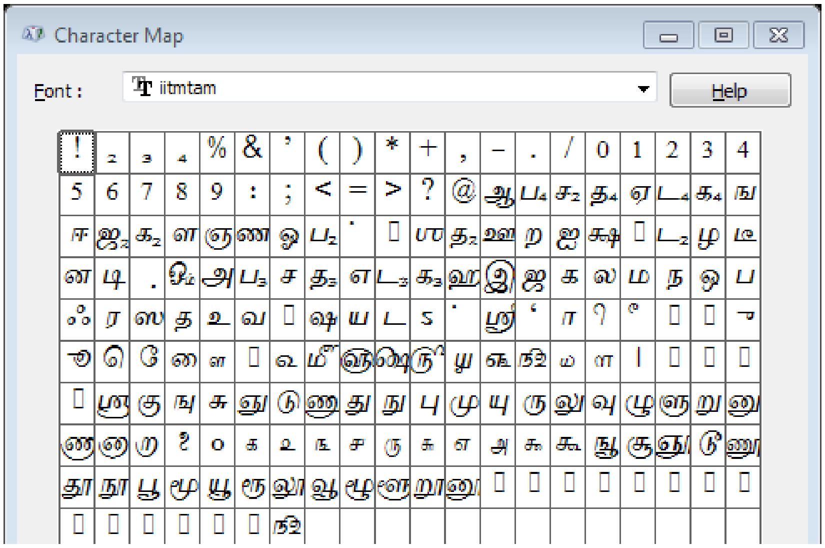
க$_{2}$ க$_{3}$ ட$_{2}$ ச$_{2}$ ஜ$_{2}$ த$_{2}$ த$_{3}$ ப$_{4}$

This feature is invoked by using an alternate support file for displaying the aksharas. With this, it will also be possible to perform data entry and directly observe the suffixes. Shown below is the text of the Sloka from the Gita in Devanagari and in transliterated form. Please observe that the dot under ம்̣ represents the Anuswar and distinguishes the akshara from the generic ம .

परित्राणाय साधूनां बिनाशायच दुष्कृतां
धर्मसंस्थापनार्थाय संभवामि युगे युगे ॥
பரித்ராணாய ஸாது$_{4}$நாம்̣ விநாஶாயச து$_{3}$ஷ்க்ருதாம்̣
த$_{4}$ர்மஸம்ஸ்தா$_{2}$பநார்தா$_{2}$ய ஸம்ப$_{4}$வாமி யுகே$_{3}$ யுகே$_{3}$ ||

Do note the usage of த$_2$ா, where the subscript numeral is appearing before the vowel sign. This is the correct rendering, where the numerals appear before the vowels signs.

The IIT-M editor being a legacy application uses 8-bit fonts to represent Tamil. It must be highlighted that, even in the 8-bit legacy encoding, the additional characters are encoded as distinct code-points, as shown in the below screenshot.



Font : iitmtam

Hence, these additional characters are no way alien to Tamil or Tamil language computing. Unicode Consortium must also take this into account, while encoding the characters.

The facts that it is used by many publishers and it has also been used in Indic multilingual processing are further arguments in favor of encoding the characters in BMP.

The word “Tamil” in the character names

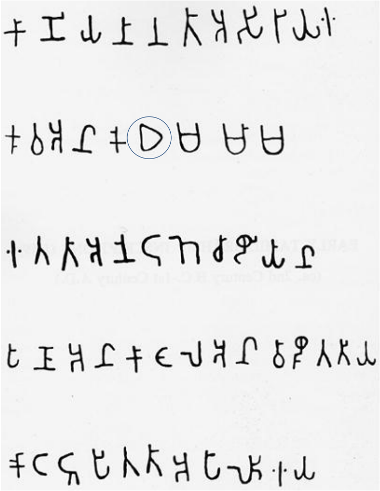
Document L2/10-363 submitted by INFITT in response to L2/10-256R had put forth the argument that, the characters shouldn’t be called Tamil, but instead “Linear Grantha”.

I am not sure why such an argument is put forth. As noted in L2/10-256R and its follow up documents, many usage samples in “Tamil” from various publishers have been shown. Using Grantha forms is one of the ways (in fact a minority way) to represent the voiced/aspirated consonants & vocalic vowels in Tamil. Most publishers tend to use superscript/subscript numerals to denote the missing characters. It makes absolutely no sense to call them as “Linear Grantha”.

Given the fact, they are Tamil character albeit in modified version, they should named as “Tamil” characters and placed in the reserved positions in the Tamil code chart as discussed in the above section.

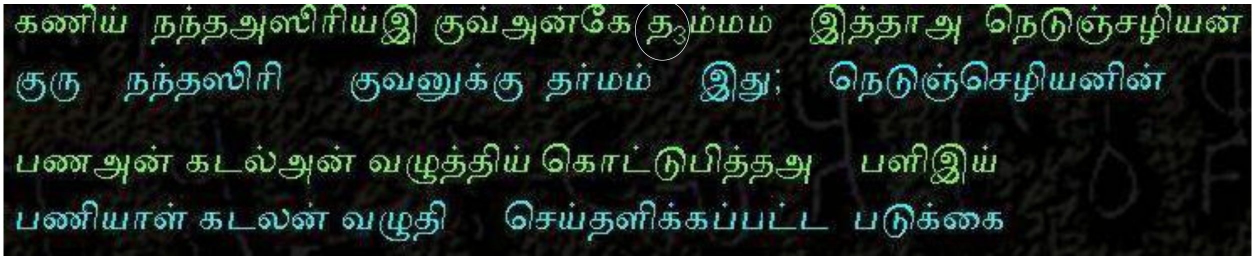
It seems INFITT considers the characters as alien to the Tamil Script. It must be noted that the extra varga consonants have been used in Tamil as early as 150 BCE [i.e in the Tamil Brahmi variant, deriving from the Asokan Brahmi characters].

Such as the inscription below [From*: Indoskript* - http://userpage.fu-berlin.de/~falk/]:



Even if this Tamil Inscription is to be transliterated into Tamil, the proposed extra consonants are needed.

Ref:[http://www.tamilheritage.org/kidangku/DrSwaminathan/scripts/08a_Scripts_of_Tamilnadu_1_Tamil-Brahmi.pdf] exactly does this by using த$_3$ [sic] (Ideally த$_4$ should have been used) to transliterate the consonant.



கணிய் நந்தஅஸிரிய்இ குவ்அன்கே த$_3$ம்மம் இத்தாஅ நெடுஞ்சழியன்
குரு நந்தஸிரி குவனுக்கு தர்மம் இது; நெடுஞ்செழியனின்

பணஅன் கடல்அன் வழுத்திய் கொட்டுபித்தஅ பளிஇய்
பணியாள் கடலன் வழுதி செய்தளிக்கப்பட்ட படுக்கை

Conclusion

The usage of additional consonants in Tamil date back as far as 150 BCE, and they are in no way alien to the language. As discussed in the previous sections, the additional consonants have been supported in Tamil language computing, within a Multi lingual environment since the past decade. The Extended Tamil proposal L2/10-256R and its follow up documents such as L2/10-379 has also shown various publication using the proposed characters.

Considering all the arguments presented here, the Unicode Consortium must consider encoding the proposed additional consonants in the reserved BMP code points in the Tamil Code chart.

Personally, I am more interested in the Transliteration between the various Brahmic scripts. I have already implemented scripts [obviously with lots of non-standard workarounds] to supports transliteration of other Indic scripts into Tamil using superscript numerals [http://www.virtualvinodh.com/tamil] & Grantha-style consonants [http://www.virtualvinodh.com/tamil-grantha]

I eagerly look forward towards the encoding of these additional characters [ideally in the BMP] as it would be very much helpful specifically in the representation of Languages like Saurashtra, Sanskrit and in general transliteration of the other Indic scripts into Tamil script.

As an appendix, the proposed code chart for Tamil with the additional characters in BMP is present in the next page.

Proposed Tamil Code Chart in BMP

	0B8	0B9	0BA	0BC	0BD	0BD	0BE	0BF
0	ம்^{2}	ஐ	ட^2	ர	◌ீ	ௐ	ரூ^{2}	௰
1	◌ँ		ட^3	ற	◌ு		லூ^{2}	௱
2	◌ஂ	ஒ	ட^4	ல	◌ூ		◌்லு^{2}	௲
3	ஃ	ஓ	ண	ள	◌்ரு^{2}		◌்லூ^{2}	௳
4	◌𑌃	ஔ	த	ழ	◌்ரூ^{2}			௴
5	அ	க	த^2	வ				௵
6	ஆ	க^2	த^3	ஶ	ெ◌		௦	௶
7	இ	க^3	த^4	ஷ	ே◌	◌ள	௧	௷
8	ஈ	க^4	ந	ஸ	ை◌		௨	௸
9	உ	ங	ன	ஹ			௩	௹
A	ஊ	ச	ப		ெ◌ா		௪	௺
B	ரு^{2}	ச^2	ப^2		ே◌ா		௫	
C	லு^{2}	ஜ	ப^3		ெ◌ள		௬	
D		ஜ^2	ப^4	(அ)	◌்		௭	
E	எ	ஞ	ம	◌ா			௮	
F	ஏ	ட	ய	◌ி			௯	

Except for the following characters, the others are placed in their equivalent positions in accordance with the other Indic blocks. An *annotation* must be provided for the additional characters indicating they are used while *representing other Indic languages* in Tamil Script.

0B80 ம்^{2} - Tamil Sign Spacing Anusvara

0B84 ◌𑌃 — Tamil Sign Grantha-Style Visarga