Response to L2/10-363 from INFITT

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

2010-Oct-02, Mahatma Gandhi Jayanti

The objection to the word TAMIL in the character names

I read with interest the document L2/10-363 submitted by INFITT in response to my Extended Tamil proposal L2/10-256R. While the INFITT people are still 'deliberating' on my proposal, they have raised objections to associating the word 'Tamil' with these characters:

> The proposed characters ... should not be called as "Tamil" or encoded in Tamil block or the names be recommended with Tamil in the character names. ... it is best if ... no connection to any Tamil characters.

I wonder if they feel that encoding these characters which are required for Sanskrit and other languages would 'taint' the Tamil script somehow.

In the original version of my proposal L2/10-256, I had asked that these characters be named with the words TAMIL EXTENDED in the character names to avoid precisely this kind of objection from Tamil purists. I quote myself from L2/10-256 p 11:

> Regarding the character names, I feel that it is better, for the same reason of avoiding clashes with Tamil purist parties as above, to name these characters beginning with the words TAMIL EXTENDED and not just TAMIL.

It was by instruction of the UTC via Rick McGowan as per Action Item 124-A129:

> Respond to Sharma re document L2/10-256 with UTC feedback that "EXTENDED" should be taken out of the names; and re-submit.

that I re-submitted the proposal as L2/10-256R asking for the characters to be named TAMIL LETTER KHA etc omitting the word EXTENDED. EXTENDED or no EXTENDED, apparently the very idea of associating the word TAMIL to characters that are needed for writing Sanskrit properly is undesirable to the INFITT people. Tamilians interested in representing Sanskrit properly in the Tamil script ask me on this: "If so, then why did INFITT get TAMIL LETTER SHA encoded when it is required only for Sanskrit?" What reply does INFITT have for that?

The claim of "technical problems"

The objections of the INFITT people mentioned in L2/10-363 are baseless. Vague wordings like "potentially causing confusion with technical implications" only recall to our mind the age-old and equally vague "difficulties in Natural Language Processing" argument that the

Tamil purists were posing complaining about how Unicode supports Tamil. I challenge the INFITT people to present any *real* and logically constructed arguments describing any *real* technical difficulties in the encoding of these Extended Tamil characters.

If one does not want these characters, one is always free to ignore them – whether one is a font-maker, rendering engine programmer or end user. How would the very existence of these characters with the word TAMIL in their name make life difficult for anyone? If the INFITT people bring forth any properly constructed logical arguments in this matter, their view can be considered. If not, mere objections will not hold water.

Importance of encoding Extended Tamil to represent the Tamil script properly

Anybody interested in Unicode representing the true breadth of the Tamil script must remember that Tamil is not used for the writing of Tamil alone. Even the Unicode chapter on Tamil has noted that Tamil with numbers serving as diacritics is used for the writing of Sanskrit and Saurashtra. I have also submitted a separate followup document to the Extended Tamil proposal showing that Extended Tamil is also used for writing Hindi, Marathi, Telugu and Kannada. To properly represent these *attested* and *real* use-cases, the encoding of separate characters for Extended Tamil is required.

I have clearly presented the arguments in favour of encoding Extended Tamil characters on p 8 of my proposal. If the INFITT people want to prevent the encoding of these characters, they should logically refute those arguments and provide alternative ways of addressing the problems that the proposed encoding intends to solve.

Problem with using existing encoded superscript digits

I have clearly pointed out in L2/10-085 p 11 the problem with using superscript digits as currently prescribed by TUS. It is that in printings of Extended Tamil one observes that the superscript digits are placed immediately after the consonant and *before* any vowel signs to be placed to the right of the consonant, but such a rendering is not achievable in the prescribed model. Sequences like KA + SUPER-2 + VOWEL SIGN AA would not be appropriate since applying vowel matra-s to GC=No characters would certainly upset Indic rendering.

Faulty rendering: கா² கோ²

Desired rendering: க²ா கே²ா

For proof that the desired rendering is indeed as shown, see the proposal L2/10-256R p 7.

The existing Unicode model does not support the proper desired rendering as it would force one to place the superscript digits after any vowel signs, and that is not right, as it is neither the place nor intention of Unicode to force orthographic reform. Unicode is to support to the best of its abilities the existing usage as evidenced by proper attestations. Such existing usage was shown in my proposal L2/10-256R as also in the follow-up document. To enable Unicode to support such existing usage and solve the rendering problem (among other matters), these characters were proposed and should be encoded.

Having not even hinted that they are interested in or concerned about addressing the rendering problem described in my previous documents as above, the INFITT people have brushed the matter off (in their email on the unicore list dated 2010-Sep-30) saying:

> *Tamil letters with superscripts ... for almost 10 years ... are encoded as sequences in Unicode.*

Without properly addressing the rendering problem stated above, any party claiming to be interested in the proper representation of the Tamil script in Unicode should not merely decry the proposed encoding with passive statements like the above.

'Linear Grantha'

The INFITT people suggest that these written forms are to be analysed as 'Linear Grantha' with the word 'Linear' possibly intending to mean (somehow) that there are neither stacks nor conjoining forms nor ligatures. If this were true, then the written forms would be represented by the proposed Grantha encoding itself, with an appropriately 'dumbed-down' font not providing for any of these 'sophistications' like stacks.

However, the observed orthography follows that of Tamil and not of Grantha, and the intention of those either importing Grantha written forms or adding superscript digits as diacritics is obvious – to extend the Tamil script to enable the representation of sounds that are not native to Tamil. Thus this is indeed 'Extended Tamil' and not 'Linear Grantha'.

The very presence and usage of the so-called 'Grantha consonants' JA, SHA, SSA, SA and HA in the Tamil script and their consequent presence in the Tamil block is already a minimal form of Extended Tamil. My proposal only seeks to complete the set and thus is nothing radically new or original in conception. (In short, I did not invent Extended Tamil.) If the characters JA, SSA, SA and HA can exist in the Tamil block with the word TAMIL in their names, and if INFITT itself can propose and get encoded a TAMIL LETTER SHA, why can not other Tamil natives like me ask for the encoding of the remaining Brahmic characters? Especially when sufficient attestation for such usage within the context of Tamil exists?

Thus the writing system intended to be supported should not be labeled 'Linear Grantha', especially when it consistently uses the Tamil written forms for those vowels and consonants needed for Sanskrit but already present in the Tamil script, i.e.: A, AA etc, KA, CA etc. Non-Tamilians who learnt Grantha for studying Vedic texts but cannot read Tamil would not immediately be capable of reading the writing system in question, while they will be able to read 'dumbed-down Grantha' (Grantha without stacks etc). This also shows that the writing system is not a development on Grantha but on the Tamil script.

(In consideration of the usage of Tamil written forms, would the INFITT now suggest the script name to be 'LINEAR MIXED TAMIL-GRANTHA'? That would however require the usage of the word 'TAMIL' in a way that might not be desired by INFITT.)

In short, the writing system for which attestations were presented and which is sought to be represented in Unicode by the encoding of the requested characters can *not* be considered 'Linear Grantha' or any other form of Grantha, especially in consideration of the existence of a variant of this writing system in which no (new) Grantha glyphs are used and only superscript digits are employed (and placed between consonants and vowel signs).

This writing system is an extension of the Tamil script to support the representing of non-Tamil linguistic content. As it is, it is indeed Extended Tamil, and therefore it is fully justified to encode these characters with script=tamil and with the word TAMIL in the character names. I should however note that I still think it would be good if the UTC also included the word EXTENDED in the character names, especially in view of such objections coming from parties that seek to (artificially) ensure the 'purity' of the Tamil script.

Conclusion

It is granted and in fact emphatically stated that the characters proposed by the name Extended Tamil are not used for Tamil. They are chiefly used for Sanskrit and Saurashtra and to an extent for other languages like Hindi, Marathi, Telugu and Kannada. They should be encoded on the strength of the attestations and arguments previously provided. Since the objections of INFITT have no real logic behind them, they are to be ignored.

Those that write only Tamil using the Tamil script are not the only 'true' Tamil community. Sanskrit scholars like me, Saurashtra speakers and those publishing religious texts of other languages like Sanskrit, Hindi, Marathi etc in the Tamil script are also native Tamilians and have a right to use the Tamil script extended in meaningful ways. INFITT has no right to object to the encoding of these characters which we require for such purposes.

–o-o-o–