

Comments on Encoding “Duplicate” Indic Characters

Vinodh Rajan

vrs3@st-andrews.ac.uk

This document is in response to the following recent documents submitted to the L2 registry authored by Shriramana Sharma - *Proposal to encode ODF5 Malayalam Letter Archaic II* (L2/12-225) and *Proposal to add two characters for Brahmi* (L2/12-226).

Encoding Written Forms as opposed to “phonemes”

In L2/12-225 Shriramana Sharma recommends for encoding an archaic form of Malayalam Independent vowel letter “ഓ”. Shriramana writes “*Unicode encodes written forms and not the sounds thereof*”. The same maxim is repeated at L2/12-226 as well to dis-unify Tamil-specific Brahmi characters LLA & VIRAMA SIGN. While the statement is true, it must also be taken into account, that Unicode doesn’t actually recommend encoding of all the glyphic variants of a character as well.

The Indic scripts usually have alternate independent forms for several letters, which cannot be just considered as simple glyphic variants of their base forms. It is not possible to encode all those “archaic” or “alternate” glyphs as independent characters in the UCS.

Consider the below examples:

Siddham

i െ െ െ

ī െ െ

u െ െ

Devanagari

a अ अ

jha झ झ

ṇa ण ण

Brahmi



ī :: ·|·

Tamil

ī  𑌦

Most probably this latter alternate form of Brahmi ī (which was more prevalent in Southern India & Sri Lanka) was the source of the corresponding alternate letter form in Siddham & Malayalam.

Quite surprisingly, L2/12-226 doesn't seem to request an independent code point for this source variant.

The interesting case is that of Tamil – 𑌦 is the oft used form, while  derived “independently” from the corresponding short vowel  is the alternate form.

The maxim “Unicode encodes written forms and not the sounds thereof” cannot be over applied to every one of the examples above. Any such application will probably result in the encoding of all the above sample glyphic variants as separate characters in the UCS.

Such a move is quite unwarranted. This would set a serious precedent to encode a dozen or more glyphic variants as independent characters in the UCS. Paleographic origins of the letters are immaterial for the process of encoding.

On “dual” encoding of /ra/ in Bengali Block & Devanagari Prishtamatra E

U+09F0 ळ & U+09B0 ळ were needed to represent “two” different modern languages in plain text. As for Prishtamatra ‘E’ since there were re-orderings involved, this subsequently posed problems for the rendering engines. Both the cases are not precise precedents to encode further duplicate characters.

For “Archaic” Malayalam II - both the alternate glyphs belong to the same language.

For the Brahmi additions, Brahmi block is itself highly unified in the UCS which requires extensive tailoring for its implementation. In all probability, the Brahmi font has to be customized for a single variant and it cannot incorporate all the myriad variants found in the inscriptions.

Brahmi Specific Dis-Unifications

Just a few Brahmi characters cannot be dis-unified citing independent paleographic origins. It must be noted that independent “Bhattiprolu” variants are already unified with the existing characters. As noted earlier, while L2/12-256 requests the dis-unification of LLA and VIRAMA, but it doesn't seem to recommend the dis-unification of the “independent”

alternative $\bar{\iota}$ · $\bar{\iota}$ · . Therefore, any dis-unification in the Brahmi block needs to be taken only after considering the overall unification/dis-unification paradigm of the entire block.

Graphemic Segmentation

Graphemic Segmentation is quite complex for Indic scripts again requiring several customizations and is not a very strong argument to dis-unify the VIRAMA for Tamil-Brahmi.

Even for Devanagari (or any other Indic script), there cannot be a single uniform graphemic segmentation. For the same word, /<MA><NGA><VIRAMA><GA><LA><MA><VIRAMA>/, मङ्गलम् is of five graphemes, while मङ्गलम् is of four graphemes, dependent on the conjunct behavior. Therefore, even for the same word, behavior of the segmentation rules are dictated by the font, and therefore must be taken care only at the application level.

Font Level Handling

As with all glyphic variations, the characters proposed in L2/12-226 & L2/12-225 must be handled at the font level.

Already, the “Malayalam Classical” font available as a part of Indolipi package (<http://www.aai.uni-hamburg.de/indtib/INDOLIPI/Indolipi.htm>) has the alternate “ \circ ” at U+0D08, and supports classical orthography of Malayalam script such as extensive conjunct behavior, ligated vowel signs for U, UU, Vocalic R and Vocalic RR etc.

In case of Brahmi, a single font cannot possibly support both Tamil variant of Brahmi and other Brahmi variants at the same time. As discussed earlier, the font needs to be specifically tailored for the particular Brahmi variant. In the absence of such a use case, there are no possible issues in unifying the proposed new Brahmi characters with the existing characters.

OpenType supports language tags, and Graphite has features which can be enabled. Hence, in case there any specific cases where both the variants must be used in the same text, the above can be harnessed.

Conclusion

Based on all the afore mentioned arguments, the UTC should not recommend the encoding of the glyphic variants proposed in L2/12-225 & L2/12-226, and instead advice the unification of the characters with the already existing characters, to be handled at the font level.