L2/12-236

Update on Brahmi and other Indic unification issues

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

2012-Jul-18

I had submitted a proposal L2/12-226 for two characters – a separate virama and LLA – to be added to the Unicode Brahmi encoding to cater to the attested Tamil language writing in that script. I hereby withdraw that proposal and document my reasons for the same. Further, as this has implications on the glyph for the existing Brahmi LLA character, I provide an update on that issue. I also reply to Vinodh Rajan's document on "encoding 'duplicate' characters in Indic" as it also pertains to the Brahmi proposal.

§1. Phoneme-unification in Brahmi

I had proposed an additional virama and LLA for Tamil Brahmi because these two written forms have an independent palaeographic origin and written form than the corresponding existing characters. Further the Northern virama produces ligatures/joining behaviour whereas the Tamil virama does not. Other reasons are listed in p 5 of my proposal.

The written forms shown in the Brahmi code chart are modeled after the Northern Brahmi written forms. As Unicode encodes written forms and not the sounds thereof, I had proposed distinct characters to represent the special Tamil written forms.

The original Brahmi proposal had clearly stated (L2/07-342 p 8) that the special written forms of Tamil Brahmi should be unified with those of Brahmi of other regions. However the rationale behind was not extensively discussed therein.

I wrote to Stefan Baums and Andrew Glass, the authors of the above document, regarding my concerns about such unification. Stefan replied that he did not think different styles of writing the same character did not warrant distinct encoding. (This view is mirrored by that of Vinodh Rajan.) However this presupposes that the various written forms are a single character. The basis for this presumption only seems to be the unity in the underlying phonemic value and not any uniformity in the written form. However, in the absence of a common glyphic skeleton, I feel one cannot assert character identity.

Further detailed discussion with Andrew indicated that there was another concern – viz that of potentially confounding users as to which character to use where. As per my

understanding, Andrew's concern was that since Brahmi underwent very many regional and transitory changes, the written forms recognized as Brahmi can vary greatly. Sometimes written forms for different phonemes may look the same without being intended so, some one-off-experiments might be seen, and so on. It would not spell for a practical or useful encoding model to unify the accidental similarities in encoded text or disunify each and every experimental form.

Personally, I feel that there is a difference between one-off-experiments and a consistent subsystem (that Tamil Brahmi or Bhattiprolu Brahmi is to the broader Brahmi writing system) where a phoneme is always written using a particular distinct written form, but I do not think it is very important to press the point given the larger picture.

Thus while normally in Unicode one does not unify per phonemic value, it is found advisable to do so for Brahmi alone. Super-unification of characters as per phonemes is what has been done in Brahmi*, and I no longer have any reason for strongly objecting to it. I hence use the terms "phoneme-unification" and "phoneme-characters" for Brahmi.

As has been noted in the original proposal (L2/07-342 §7), a distinct font would be required for each "epigraphic ductus" and such fonts can present each phoneme-character as attested. This would both cater to the differences in written forms (whether glyphic variants or otherwise) and also make it so that distinct phonemes which may be written almost the same in a particular ductus** are recorded distinctly for search and other operations to be meaningful.

I hope to work with Andrew to produce a Unicode Technical Note on this.

One point I should note however is that Andrew initially suggested that it might be advisable to disunify the Tamil Brahmi virama alone while not disunifying the LLA. His reasoning was that while the difference in the case of LLA is merely graphic, that of the

---

* Note that while this seems to be excepted in the case of Bhattiprolu vowel sign AA vs the regular vowel sign AA, it is not really so, since within the Bhattiprolu system the two marks denote distinct phonemes!

** I recently (Mar 20, 2012 at a workshop in New Delhi) learnt that in certain varieties of Brahmi, TA and NA are almost non-distinct and are subtle mirror images of each other. And in the daughter-script of (Tamil) Brahmi named Vatteluttu, some variants present KA and CA virtually identically (Ref: Burnell's South Indian Palaeography, plates 17 and 32.) [Increase in such ambiguous representations is one of the reasons attributed to the eventual demise of Vatteluttu. Ref: South Indian Temple Inscriptions, T N Subramanian.]

virama is functional. That is, the Tamil Brahmi virama is productive in the sense of producing ligatures, which would have an effect on the placement of vowel signs etc.

Further, in Tamil Brahmi the virama is also applied to the vowels E/O (both independent and as vowel signs) which would mean that sequences such as C1 + VS-E/O + VIRAMA + C2 would have a valid rendering in Tamil Brahmi whereas they are invalid for Northern Brahmi and should not produce a subjoined form of C2 if rendered using a fall-back font which might be based on Northern Brahmi.

Andrew felt a rendering engine might need to be aware of this at the encoding level.

However I feel that this is largely an OpenType-specific problem. Using the alternate rendering technology Graphite (specifically, its support for user-defined features) I can already mark-up text within the same document and using the same font as either Northern Brahmi or Tamil Brahmi etc, whereby the sequences would be rendered correctly in the respective orthography.

In my opinion, OpenType is ill-suited to handling rare orthographies of palaeographic scripts (at least those of the Brahmi family) and such variations are best catered to by Graphite. As already remarked, Graphite can handle the Tamil/Northern Brahmi situation very well. It is only OpenType that faces difficulties due to the hard-coding of presumptions regarding the structure of orthographic syllables.

Therefore, this is an implementation-level problem. As Vinodh has remarked, the Brahmi encoding "requires extensive tailoring for its implementation". As such, the encoding need not be fixed to cater to an implementation-level problem.

Therefore the phoneme-unification principle for Brahmi need/should not be excepted even in the case of the virama, and additional characters representing the same phonemes (or having the same phonemic function, such as vowel-reduction) as existing characters need/should not be added for Brahmi.

§2. Representative glyph for 11034 BRAHMI LETTER LLA

Given the phoneme-unification policy for Brahmi outlined above, the question I previously raised in L2/12-106 §5 is to be reconsidered. This is with regard to the representative glyph for 11034 BRAHMI LETTER LLA.

The earliest indubitable attestation for the LLA phoneme-character in any subsystem of Brahmi is in Tamil Brahmi. The written form is:

... derived from Brahmi LA 𑀮. The Northern LLA derives from DDA, but its indubitable attestations are of a later date (as shown on L2/12-106 p 7). However, (as noted ibid p 8,) a few scholars have analysed some written forms even from Ashoka's edicts (which are earlier than Tamil Brahmi) as LLA. These are like DDA but with a bulge at the bottom:

However, not all scholars agree that these are LLA and some merely read it as DDA. Andrew has said he will consult experts on this. If they feel the reading of the above Ashokan glyph as LLA is sufficiently satisfactory, then this would be the oldest and appropriate glyph for 11034. Otherwise the Tamil Brahmi LLA would be the oldest glyph for the chart.

§3. On glyphic variants and Indic unification

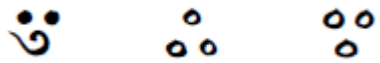
This final section replies to the document recently submitted by Vinodh Rajan objecting to my proposal L2/12-225 for an archaic II for Malayalam and L2/12-226 for the two additional characters for Brahmi mentioned above. While I have withdrawn my Brahmi proposal, I have not withdrawn the Malayalam one. This is because while phoneme-unification is appropriate for Brahmi, which is a palaeographic script with too many variations and inconsistencies to unify, phoneme-unification has not been and should not be done for modern Indic scripts, which are furthermore required to be used in IDNs where unpredictable glyphic changes would cause security issues.

Vinodh writes:

> *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.

He proceeds to give various shapes representing the same sound in various Indic scripts. However, except in the case of Tamil, these are to me clearly glyphic variants sharing a common glyphic skeleton. For example, in the case of Siddham I:



... the general skeleton is that of three circular strokes. The difference between the three is only that of closing the stroke vs ending it with an open swish, or in the orientation. Given this, I personally do see these as glyphic variants. The Devanagari "Northern" (or to be precise "Calcutta") style glyphs for A, JHA, NNA etc also have a common evolutionary origin and skeleton with the regular glyphs even though this may not superficially be apparent. As for Brahmi, I have always analysed the ∶ ∶ vs ·।· shapes as variants where the latter form is just a re-orientation of the former with two points joined by a straight line. (In any case, Brahmi has phoneme-unification as discussed before.)

Of course, some may not see the uniform skeletons as I do, but certainly these glyphs Vinodh provides as representing the same character have more in common between them than do ൟ and ഈ where there is no postulatable common skeleton at all! As noted in my Malayalam proposal L2/12-225 p 2, the latter form was an intentional new invention by a Malayali scholar for a simplification of the script by making the I II pair ഇ ഈ similar to U UU ഉ ഊ. This new form ഈ does not share palaeography (implying a common glyphic skeleton) with the old form ൟ unlike in the Devanagari examples.

As such, the Malayalam case is parallel to the one proper pair that are *not* glyphic variants provided by Vinodh, viz ഀ vs ഋ for Tamil II. However unlike ഈ, ഀ does not seem to have come to consistent usage. If Vinodh has any attestations for such usage of ഀ, he is free to submit a proposal for it.

Apart from the above, which discusses the main thrust of Vinodh's document, there are a few other points which I should mention.

Vinodh notes that the two forms of Devanagari E i.e. the pṛṣṭhamātrā ◌ॅ vs the regular ◌े were disunified as they need to be treated differently by rendering software. This is correct and appropriate.

He further states that Bengali RA র vs Assamese RA ৰ were disunified (in a unified encoding of the two writing systems) to different two different languages in plain-text. However, I understand that Han is used both for Chinese and Japanese, and while there are distinct written styles for the two orthographies, the (core common) encoding is the same. The disunification in the case of Bengali-Assamese hence cannot be brushed aside as for representing different languages in the same text. The disunification is because the two forms are not glyphic variants unlike in the case of Chinese vs Japanese usage of Han.

One point Vinodh correctly notes is that graphemic segmentation for Indic is complicated, and the same word within the same script when rendered with conjuncts or otherwise would properly take distinct segmentation; for example, maṅgalam written in Devanagari as म | ङ् | ग | ल | म् vs म | ङ्ग | ल | म्. (However many implementations do not test whether the font provides the conjunct or not, and blindly treat encoded sequences following a certain regex pattern as a single syllable. This is certainly not ideal.) Ken Whistler also expressed (in private mail) similar sentiments to my grapheme segmentation argument. As such, I agree that that argument is not a strong one for disunifying the Tamil Brahmi virama from the Northern one. But I have withdrawn my Brahmi proposal anyway!

Finally, Vinodh has suggested the usage of environmental language tags to distinguish Malayalam ൎ from ഩ (as are probably used for Han to handle Chinese vs Japanese). However, he has ignored the confusability issue I have noted on p 6 of L2/12-225. The currently encoded ഩ is confusable with the sequence 0D07 ഇ + 0D57 ◌ൗ, whereas the newly proposed ൎ is confusable with 0D02 ◌ം + 0D30 ര + 0D02 ◌ം. Unicode security protocols cannot meaningfully map a single character to both ഇ + ◌ൗ and ◌ം + ര + ◌ം. Therefore it is advisable to disunify from ൎ from ഩ. I have already noted (op. cit.) that the archaic ൎ may be prohibited from use in IDNs if the user community wishes so.

(In the case of palaeographic scripts like Brahmi the security problem as noted above is not taken into consideration since they are not intended to be used in IDNs!)

In sum, ൎ should be disunified from ഩ whereas a different policy is adopted for Brahmi as described above.

-o-o-o-