L2/09-324

Comments on Kent Karlsson's document L2/09-277

Shriramana Sharma – jamadagni-at-gmail-dot-com

2009-Sep-26

This document comprises my comments on Kent Karlsson's document L2/09-277: "Comments on L2/09-141, Proposal to encode the Grantha Script". This is *not* a defense of L2/09-141 by Naga Ganesan since I myself have submitted L2/09-316 which is also intended to comment on and strongly object to various aspects of Ganesan's L2/09-141. Rather, it should be understood as merely a reply to various viewpoints presented by Kent Karlsson in L2/09-277 from the viewpoint of myself as a native Grantha user.

I would here like to thank Kent Karlsson for his promptness in giving me a copy of L2/09-277. He has caused me to examine aspects of Grantha I had not considered before.

What constitutes a different script

In my opinion, the perception of which two scripts are the "same script" or "different scripts" is at least to an extent subjective. A native user may easily perceive differences between two scripts whereas an non-native might tend to generalize and categorize both as superficial variants of a single script. This may be said to be akin to the determining of which two languages are indeed two languages and not merely two dialects of the same language, but I will not insist on that comparison. I will only consider scripts.

Say we have two samples A and B of similar but different written samples. A practical question which might help us determine whether the belong to the same script or not is "Can an intelligent person who can understand A and has no previous specific exposure to B easily overcome the small variations between A and B and understand B also? And is the converse also true?" In my opinion, only if both (the "proverse" and the converse) are true, can both A and B be considered variants of the same script.

In the case of Tamil and Grantha, one can confidently say that Tamilians (those whose mother tongue is Tamil) who can read Grantha can also read the Tamil script. However, the converse is not true. Only a very small fraction of those who can read Tamil can also read Grantha. There are also some people who are not Tamilians but are comfortable with Grantha owing to have studied the Veda-s using that script. These people cannot read Tamil well. (I speak from personal experience with such people.)

Therefore, going by the criteria I suggested above for the disunification of scripts, Tamil and Grantha are indeed different scripts.

Importance of native users' perception

To a native user of both the Tamil and Grantha scripts like me, saying that Tamil and Grantha are the same script reminds me of what they say about people from one part of the world thinking that people from another part of the world all look alike. It is possible that people from one continent may be unable to find distinguishing characteristics among people from another continent. However, everyone is able to distinguish among themselves. An outsider would evidently see the similarities first. A native would immediately see the distinguishing characteristics.

I think it is only fair to expect that native users' perceptions are given importance in determining identity/distinctness of scripts. I see many precedents to this.

Kaithi and Gujarati scripts are very similar and share a lot of characters with the same orthographic behaviour. The Wikipedia articles on the Balinese and Javanese scripts (as of 2009-Sep-26) even alleges that they are typographical variants. To me, Ranjana looks a lot like Fraktur Devanagari! I suspect many more examples like this can be found. Yet, all these are considered different scripts by the natives that use them, and hence they are correctly disunified by Unicode.

Behavioural distinction between Tamil and Grantha characters

Returning to the present case under discussion, one must note that there are a lot of behavioural differences between the characters of the Tamil and Grantha scripts.

Tamil exhibits ligatures of consonants with the vowel signs U/UU whereas modern Grantha does not use these ligatures. (Archaic ligatures are attested, though.) Tamil does not stack consonants, does not have a "repha", "ra-vattu" or "ya-phalaa" and uses only the single ligature K·SSA. Grantha, however, regularly stacks consonants, uses the "reph", "ra-vattu" and "ya-phalaa" consistently and has very many ligatures apart from K·SSA.

If we go beyond mere orthography, Tamil is phonemic and the same character represents different sounds, whereas Grantha is for the most part phonetic and uses each character for only one sound.

All these would suggest that these should be treated as two different scripts.

The suggested decompositions

Karlsson has also proposed a number of decompositions to minimize the number of new characters that need to be, in his opinion, added to the Tamil block to enable the proper representation of Grantha. I now consider them one by one.

1) கூ = கு + ா

I do not understand how கூ would be considered a ligature of கு and ா. Perhaps Karlsson is over-extending the analogy of Devanagari where glyphically आ = अ + ा ? But this would mean that he suggest a similar decomposition for Tamil and Malayalam too: ஆ = அ + ா and ആ = അ + ാ.

It is true that in Tamil, Malayalam and Grantha the glyphs for AA are formed by 'extending' the glyphs for A. But that does not necessary constitute a ligature. In Kannada we have DA = ದ and DHA = ಧ. One may say that the extra line at the bottom 'aspirates' the consonant and hence ಧ must be considered to be a ligature of DA ದ and HA ಹ. However, such is not the native users' perception, because there is no consistency in this.

The feature of the extra line being used to 'aspirate' the consonant or 'lengthen' the vowel is not consistently observed in Tamil, Malayalam or Grantha, *unlike* the case of Kannada ೀ which is used in the vowel signs II ಿೀ, EE ೇೀ and OO ೊೀ consistently to distinguish them from the short vowel signs I ಿ, E ೇ and O ೊ.

Therefore, while Unicode provides decompositions for the above-mentioned long vowel signs in Kannada, Unicode does not consider the Tamil and Malayalam independent vowels AA as ligatures or provide decompositions for them. This is in accordance with native users' perceptions.

Therefore Grantha கூ is not a ligature and should not be treated as such.

2) ஊ = உ + ள

Similar arguments as above are to be read for this too. Saying that this is a ligature is not in accordance with native users' perception. The precedent with existing Tamil and Malayalam blocks not having decompositions for independent vowel UU also goes against Karlsson's saying that this is a ligature of the independent vowel U with the 'length mark'.

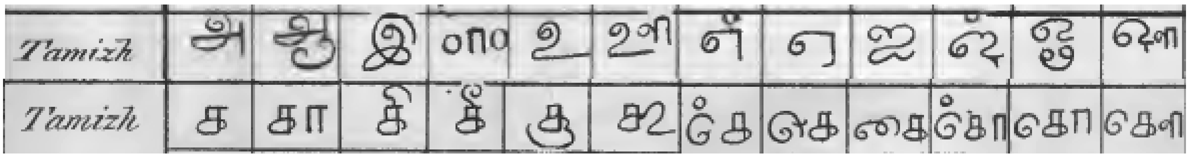
Here I wish to mention that the name 'Length Mark' is not really appropriate for Grantha ள, nor is it appropriate for all the other Indic characters named this way. The only character in all the Indic scripts to which this name really fits is Kannada ೀ as shown above. (Telugu also has ీ which is also consistently used as a length mark, but for some reasons, the Telugu vowel signs are not provided decompositions using this.) All the other characters unfortunately named as 'Length Mark' are merely the left-out part of two-part vowel signs where one part is already encoded separately for another purpose. Oriya even has two such badly named 'Length Marks' – one for AI and one for AU!

Further, in Malayalam and Grantha, the so-called ‘Length Mark’ should more appropriately be named ‘Vowel Sign AU New’ because they are used as the vowel sign AU in modern orthography. (It is too late now for Malayalam, though not so for Grantha.)

3) ஔ̊ = ஔ + ◌̊, ஒ̊ = ஒ + ◌̊

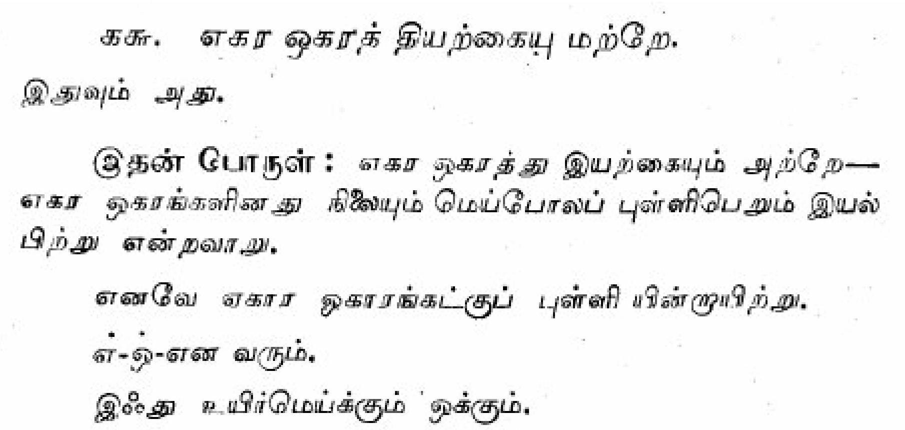
This would be a valid decomposition, but the ‘pulli’ (the modifier mark on top) can look like either a dot or a ring. Mr Ganesan has shown a dot, but a ring is also attested (and I prefer it personally). We must remember that this ‘pulli’ on top is ‘borrowed’ from archaic Tamil usage. Referring to Tamil usage, we find that both forms of the ‘pulli’ are valid.

See for example the following samples from pages 1 and 2 of “Ancient and Modern Alphabets of the Popular Hindu Languages of the Southern Peninsula of India”, Captain Harkness, 1837 (http://www.archive.org/details/ancientmodernalp00harkrich):



which both show the ‘pulli’ as a ring. (I removed the empty spaces in the samples.)

Other samples such as the following from a 1937 edition of the earliest known work on Tamil grammar, the Tolkāppiyam, shows a dot for the same. (See page 40 of the PDF from http://noolaham.org/wiki/index.php/தொல்காப்பியம்_எழுத்ததிகாரம்.)



கசு. எகர ஒகரத் தியற்கையு மற்றே.

இதுவும் அது.

இதன் பொருள்: எகர ஒகரத்து இயற்கையும் அற்றே— எகர ஒகரங்களினது நிலையும் மெய்போலப் புள்ளிபெறும் இயல் பிற்று என்றவாறு.

எனவே ஏகார ஓகாரங்கட்குப் புள்ளி யின்றாயிற்று.

எ்-ஒ்-என வரும்.

இஃது உயிர்மெய்க்கும் ஒக்கும்.

The summary translation is: “The nature of short E and O is the same as that of the vowelless consonants which is to take a ‘pulli’ on top.” and the last line specifically says “this applies to the syllables (i.e. the vowel signs used in syllables) as well”.

Thus it is obvious that the pulli is rendered interchangeably as both a dot and a ring. Though today the ‘pulli’ is not used in Tamil for these short vowels, it is still used as the virama sign and natives write that virama sign as a dot or ring as they wish.

Hence to provide a decomposition using either a combining ring or dot (both at the same time is of course not possible) would be to discriminate against the other valid usage and should not be done. Without the decomposition, designers of fonts will be under no pressure to conform to the dot or the ring and can present the 'pulli' as either.

4) 𑌐 = 𑌏 + ◌𑍇, 𑌔 = 𑌓 + ◌𑍗

These two are hard to speak against as they (especially the latter) in fact have a precedent in the decomposition of Tamil 0B94 ஔ as ஒ + ◌ௗ. However, the Malayalam equivalents ഐ and ഔ and Sinhala equivalents ඓ and ඖ are not decomposed. In fact, none of the Indic independent vowels (from 0900 to 0DFF) are decomposed despite many opportunities to do so. (Tamil 0B94 ஔ is the only aberration. I do not know the reason for this aberration.) Unicode in fact recommends against the use of अ + ◌ा to represent आ etc. Therefore if decompositions are permitted here for Grantha, Sanskrit texts composed using such decompositions may suffer when 'roundtripping' from Grantha to Devanagari and back. Therefore I would prefer to go by the majority and not decompose these vowels.

5) ◌𑍇̊ = ◌𑍇 + ◌̊

I think Karlsson has overlooked something here. When two combining marks occur in sequence, they both modify only the base character. I do not know of any case in Unicode where a combining mark modifies another combining mark. Of course, successive combining marks in European scripts are placed one above another, but that occurs only when both are of the same combining class. For one, Indic vowel signs are given a combining class of zero, and for another, even if ◌𑍇 had a non-zero combining class, it would not match that of ◌̊. Further, the argument given above against decomposition of 𑌣̊ and 𑌕̣ regarding impropriety of discriminating between dot and ring applies here too.

6) ◌𑍈 = ◌𑍇 + ◌𑍇

This might be valid but Malayalam ◌ൈ and Sinhala ◌ෛ are not decomposed this way and so Grantha ◌𑍈 should also not be decomposed. Same argument about 'roundtripping'.

7) ◌𑍇̊𑌾 = ◌𑍇̊ + ◌𑌾, ◌𑍋 = ◌𑍇 + ◌𑌾, ◌𑍌 = ◌𑍇 + ◌𑍗

These decompositions are provided. There is nothing new here.

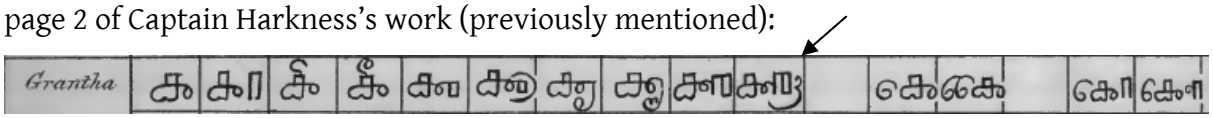
7) 𑍠 = 𑌋 + ◌𑍄, 𑍡 = 𑌌 + ◌𑍄

As I said before, none of the Indic independent vowels (except one in Tamil) are provided decompositions. I remain yet to be convinced of the need for Grantha to differ.

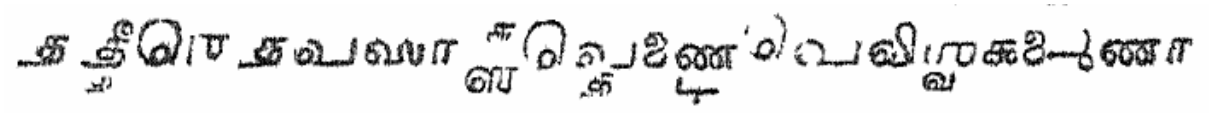
Independent vowels and vowel signs for vocalic L/LL

Karlsson has suggested that the independent vowel and vowel sign for vocalic L be unified. The same may be said of vocalic LL too since though it is preferably presented as ◌ൢ, ◌ൣ is also attested. Thus though there is glyphic identity, and though both glyphically follow the preceding character as in logical order, they should *not* be unified. The reason is that the vowel signs have different behavioural patterns than the independent vowels:

They are attested to ligate with the consonant KA, as in the following sample from page 2 of Captain Harkness's work (previously mentioned):



The independent vowels obviously would not ligate. Further, sometimes (though very rarely) the vowel signs are placed below the consonant (as in most other Indic scripts) as in this sample from http://ambaa.org/pdf/halasya_mahatmyam.pdf (70 MB!) page 4:



And as for vocalic LL being a ligature, the same argument presented before for ക്ഝ not being a ligature applies. Neither do native users perceive it that way nor is there reason to.

Further, Karlsson's repeatedly saying that 'ligatures are common for Grantha and Tamil vowel signs' is not at all appropriate. In fact, Tamil and Grantha keep most of their vowel signs glyphically separate from the base consonant, unlike Kannada and Telugu which glyphically attach most vowel signs to the base consonant. Thus it is entirely incorrect to say that this is "common" for Tamil and Grantha vowel signs.

Grantha Virama

Karlsson's suggestion that the visual and functional aspects of the Grantha virama be disunified as for Khmer does not fit within the Indic model, but will only make roundtrip conversions between Indic scripts unnecessarily cumbersome. Roundtrip conversions are especially important for Sanskrit to which Grantha is dedicated, since Sanskrit can be written in most Indic scripts and in fact a very large number of religious texts in Sanskrit are published in all the major Indic scripts. Thus there should be only one Grantha Virama.

Characters common to Tamil and Grantha

Karlsson has pointed out that various characters are common to Tamil and Grantha. I do not deny this. It is a well known fact. In fact, this makes it easier for those who can read the Tamil script to learn Grantha. However Karlsson mixes up those characters which are

identical between the scripts and those which are only similar. I shall provide a clarified list of the characters common to Tamil and Grantha.

Karlsson also says that LLA is common between the scripts too, but this is incorrect. There is perhaps only a not-very-close similarity between Tamil ൽ and Grantha 𑌲.

Glyphically and semantically identical

The independent vowels U, UU, the consonants NNA, TA, NA, YA and VA, the confessedly 'Grantha' consonants JA, SHA, SSA and HA, the vowel signs AA, I and II, the digits 0-9 and the numbers 10, 100 and 1000 are glyphically and semantically identical in both scripts. The archaic Grantha vowel sign for AU is likewise identical to the vowel sign AU in Tamil.

Note that I do not accept the disunification of Tamil NNNA, RRA and LLLA myself since they have absolutely no attested usage in Grantha. They are *not* considered by native Grantha users to be a part of the Grantha character repertoire. If they are to be used for transliteration, they may be 'borrowed' from the Tamil block. So I do not list them here.

Glyphically identical but semantically different

The Grantha vowel signs for long EE and OO are identical to those of short E and O in Tamil. It is to be noted however, as we hinted at before, that Tamil also formerly used the same vowel signs for long EE and OO as Grantha and differentiated the short vowels by the 'pulli'.

Glyphically similar, semantically identical

The independent vowels A, AA, II, OO and AU, the consonants KA, TTA and RA, the 'Grantha' consonant SA may be said to be glyphically very similar between the two scripts. The ligature K·SSA is also very similar to that used in Tamil.

The similarities may not be resolved by change of font

Karlsson suggests that the characters which are glyphically very similar may be unified. In my opinion, this would be mere overzealousness at unification. When a character has a consistent shape within a particular usage context, I do not think it appropriate to unify it with a glyphically similar character of a different context. If a character has glyphic variations within the same context, however, they may be unified, but not across contexts.

It is my belief that the Kaithi danda-s accepted for Unicode 5.2.0 at 110C0 and 110C1 were disunified from the generic Indic danda-s 0964 and 0965 *against* the general principle of not encoding script-specific danda-s for Indic scripts *precisely* because they are consistently represented in Kaithi by those distinct glyphs and using the generic Indic danda-s would *not* satisfactorily represent Kaithi in plaintext. Grantha also consistently

represents its KA in a particular way that is glyphically different from the KA in Tamil. Thus Grantha's KA should also be permitted to be distinctly represented in plaintext.

As I see it, each character in a script has a 'right' to have a proper representative glyph. The glyph for Tamil KA is *not* representative of Grantha KA.

Further, unifying Grantha KA with Tamil KA would make it impossible for Grantha texts to visibly distinguish between the digit 1 and KA. In Grantha, the glyphic distinction between 𑌕 and 𑍗 is important for readers to distinguish the two. This distinction, which is important especially in Sama Vedic texts where it is often impossible to predict from context whether KA or 1 (the latter marking a Sama Vedic svara) is to be understood, would be lost if Grantha and Tamil KA are unified. While similar cases cannot be presented for TTA and RA, the 'consistent usage of distinct glyph' argument still applies to them.

If these characters are forcefully unified citing that their distinct glyphs may be presented by change of font, it would be absolutely unacceptable to the Grantha user community as they would be unable to clearly present these Tamil and Grantha characters distinctly in plaintext. This is all the more important seeing as the Tamil and Grantha scripts are mixed to a great extent in writing Tamil Manipravalam in which native users diligently follow the Grantha style for the Grantha characters and the Tamil style for the Tamil characters. Forcing them to use the same glyph for these characters in both scripts would only be seen by the native users as unwarranted interference by extraneous parties! Therefore, native users' sentiments regarding the proper representative glyphs of characters must definitely be respected and such over-unification must not be done.

The excellent parallel of Kaithi and Gujarati

I believe that the UTC also respects the native users' perception this way, seeing as they have disunified the Kaithi script from Gujarati for Unicode 5.2.0. As Anshuman Pandey, author of the Kaithi proposal, states in defense of disunifying Kaithi from Gujarati:

> The standardization and official recognition of a script and the subsequent adaption of the script in print technology suggests that the script is an independent writing system with a distinct typology and scribal tradition.

Any suggestion of unifying Tamil and Grantha is hence akin to unifying Kaithi and Gujarati. If it is said that Tamil and Grantha share many identical characters, especially for the vowel signs and the numerals, the same is the case with Kaithi and Gujarati. If it is said that there are many similar characters between Tamil and Grantha, the same is true for Kaithi and

Gujarati. I also point out that the parallel goes so far as some characters being glyphically similar but semantically different between the scripts!

So when Kaithi and Gujarati are encoded as different scripts, and the minimalistic approach of merely adding whatever extra characters are needed for Kaithi to the Gujarati block is not considered sensible, Grantha and Tamil should certainly not be unified.

Extending the same argument, while the vowels and consonants of Kaithi are disunified from those of Gujarati, the numerals of Kaithi were judged to be identical or mere stylistic variants of those of Devanagari and hence not disunified. By the same argument, the numerals encoded in the Tamil block should not be disunified for Grantha. I have previously presented in L2/09-316 samples that prove beyond doubt that the exact same numerals as exist encoded in the Tamil block are used for Grantha also.

The security concern

Finally, as for the argument of double encoding posing security issues, Latin and Cyrillic Letter A looking the same is a much bigger problem than any pair of Indic characters looking alike, obviously because there are more webpages and domain names with Latin-script content than Indic-script content. Whatever mechanisms exist or are devised in the future to ensure security in Latin-script environments will apply to Indic-script environments too. It would be hence injustice to shortchange an Indic script on this basis.

Conclusion

The Grantha and Tamil scripts, while sharing a lot of character repertoire, are considered by the native users as distinct scripts. Their linguistic nature is also distinct, with Tamil being phonemic and Grantha being phonetic. Most Grantha and Tamil characters that are glyphically similar are behaviourally distinct. Forcefully unifying Grantha and Tamil characters based on glyphic identity or similarity would lead to improper or unsatisfactory representation of Grantha in plaintext. Unicode precedent is for the disunification of scripts which share a lot of glyph/character repertoire, if there are demonstrably different characteristics that are unique to each script.

Hence Tamil and Grantha should be treated as different scripts and disunified. The numerals encoded in the Tamil block should however not be disunified for Grantha.

-o-o-o-