L2/13-222

Comments on revised PRI 250 document on Malayalam conjoining forms

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

2013-Nov-02

This is w.r.t. Cibu Johny’s document L2/13-219 requesting distinct encoded representations of C2-conjoining forms in Malayalam Unicode. Cibu’s document is a revised version of his earlier doc L2/13-087 which is the background doc for PRI 250. My feedback to that PRI doc was L2/13-048. Now I wish to record my observations on Cibu’s current doc L2/13-219.

§1. Summary of Cibu’s document

Cibu’s (revised) doc is based on the presumption that in the Malayalam script, consonants occurring in non-initial position in clusters can take “two different conjoining forms” as per whether the old or new orthography of the script is followed.

Cibu asserts :— There is a need for distinct encoded representations (which nevertheless should be equivalent after ignoring joiner characters) for the forms from the old vs new orthographies. Since a font by default conforms to either the old or new orthography, using such encoded representations one could unambiguously request the written forms belonging to one orthography from a font which would normally render the other orthography.

Cibu substantiates this claim by pointing out :— In the “Akṣarapallī” system of denoting numbers, old ligatures have numerical values. For instance the old ligature:

ഗ്ര

which represents /gra/ has the numerical value 7. The unligated modern form of /gra/:

ഗ്ര

is not known to have the value 7. Hence a user wishing to use the ligated /gra/ as a representation of 7 in Malayalam text while otherwise using the modern orthography for the rest of the text would find a sequence to unambiguously request the ligature useful.

By extension, it would be useful in general to be able obtain the old orthography even from a font that normally renders the new orthography and likewise vice versa.

Note that Cibu analyses the old ligature /gra/ seen above as the basic GA with a "cursively connected" C2 conjoining form of RA. Thus in his view (some) consonants in Malayalam have two different C2-conjoining forms, one "cursively connected" and the other "disconnected", these belonging respectively to the old and new orthographies.

Cibu goes on to propose that the sequence ZWJ + VIRAMA + C2 should unambiguously represent the "cursively connected C2-conjoining form" of the old orthography and ZWNJ + VIRAMA + C2 should represent the "disconnected C2-conjoining form" of the new orthography.

§2. Problems in Cibu's document

As I see it, there are two major problems in Cibu's doc.

§2.1. Ligatures, not "cursively joined conjoining forms"

The first is that Cibu analyses ligatures such as:

ഗ്ര

… for /gra/ as involving conjoining forms. It is curious that while Cibu asserts that his model follows the existing model outlined by L2/04-279 ("the PRI 37 doc"), he has ignored the distinction made by that earlier doc between ligatures and conjoining forms.

As I understand it, in the context of consonant clusters exemplified by the character sequence CA + VIRAMA + CB, a ligature is a single glyph that represents the whole sequence, whereas if one consonant is represented by its nominal glyph and the other is represented by a special glyph, then that special glyph is called the conjoining form of that consonant.

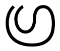
I agree that in the above ligated form of /gra/ the lower leftward-curving stroke is indeed a "conjoining" form of RA *as far as the orthographic nature of the script in the real world is concerned*. However, the existing definition of conjoining forms is *in terms of glyph substitution in text shaping software*. Indeed, the PRI 37 doc prescribed the model it did specifically to ensure that the glyph substitution rules such as:

RA + VIRAMA + ZWJ → REPH

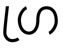
ZWJ + VIRAMA + CB → C2-conjoining form of CB

… would work uniformly throughout even in combination with 00A0 NBSP, 25CC DOTTED CIRCLE, in modes displaying control glyphs and so on.

Therefore, from the POV of glyph substitution, the /gra/ glyph seen above can not be analysed as involving a conjoining form, but it only has to be analysed as a ligature. It hence automatically follows that Cibu's proposed model, whereby:

GA + ZWJ + VIRAMA + RA → 

... cannot be valid because the glyph shown is a ligature and does not involve a C2-conjoining form. As per the existing model ZWJ + VIRAMA + RA has to produce the C2-conjoining form of RA and hence the following is what should result:

GA + ZWJ + VIRAMA + RA → 

The same argument applies to the various ligatures (like /tsa/ etc) which Cibu analyses as involving "cursively connected C2-conjoining forms".

§2.2. No need for any sequence to request ligatures

One should remember that the existing Indic encoding model does *not* provide any sequence to request ligatures. Thus, short of encoding a new ZERO WIDTH LIGATOR (which apparently has already been rejected) there can be no sequence to specifically request the ligated forms from the old orthography of Malayalam (or any other Indic script).

Nor is there a need to. Sanskrit scholars like me probably have the greatest use for ligatures (since the heaviest consonant clusters in Indian languages are indubitably to be found only in Sanskrit), and whatever ligatures we need we can get by using an appropriate font (such as Sanskrit 2003 for Devanagari or Rachana for Malayalam) which produces these ligatures as the default representations of the basic cluster sequence CA + VIRAMA + CB.

Cibu pointed out the use-case of Akṣarapallī where ligatures represent numerals and the non-ligated forms are not known to represent numerals. This is true, but it is not possible in all cases to distinguish letters used for numerals from letters used for language content in plaintext.

For instance, the glyphs of the Tamil letters PA ப, VA வ, NGA ங, TA த and LLA ள are also used as numerals or symbols but they are not disunified from the existing letters (see L2/13-047 §1) and their identification as numerals are to be left to higher-order protocols (presumably, such as some markup). Even in Malayalam, MA മ represents the fraction 1/80 (see L2/13-051 §3).

Thus if one requires ligatures for any particular purpose, the established and correct way of doing so is to either use an appropriate font throughout or to selectively enable smart-font features for the appropriate sections of text. I do not see any point in trying to invent an encoded representation of ligatures for plain-text Indic.

§2.3. ZWNJ breaks clusters, is not included in them

Apart from the incorrect analysis of ligatures as involving C2-conjoining forms, the other major problem in Cibu's doc is the suggestion to use ZWNJ to produce C2-conjoining forms. I have already drawn attention to this in my previous response to PRI 250. Vide L2/13-048 p 3 §2. I will summarize it below.

Cibu suggests that the sequence ZWNJ + Virama + CB be adopted to request the "disconnected C2-conjoining form" of CB such as in:

GA + ZWNJ + Virama + RA → ഗ്ര

The problem with this is first of all that this is the *only* C2-conjoining form of RA in the Malayalam script as explained previously. It hence should be selected by the sequence *ZWJ* + Virama + RA (and not *ZWNJ* + Virama + RA) as per the existing model, as in:

GA + ZWJ + Virama + RA → ഗ്ര

Further, the behaviour of ZWNJ has always been to merely act as an invisible obstacle to the rules which would cause conjoining or ligating behaviour and not to be itself included in glyph substitution rules of fonts. Vide TUS 6.2 p 551 (p 581 of PDF):

> *For modern font technologies, font vendors should add ZWJ to their ligature mapping tables as appropriate. … In contrast, ZWNJ will normally have the desired effect naturally for most fonts without any change,* ***as it simply obstructs the normal ligature/cursive connection behaviour****.*

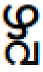
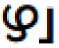
To request that ZWNJ + Virama should produce C2-conjoining forms would hence be a deviation from existing practice of fonts and rendering engines of handling ZWNJ. IIUC, layout engines would break grapheme clusters at ZWNJ. Indeed, only then would the name *non*-joiner be appropriate! I am not sure how appropriate it would be to user a *non*-joiner to producing a *conjoining* form.

Hence Cibu's suggestion of using ZWNJ + Virama to produce "disconnected" C2-conjoining forms is inappropriate on these two counts.

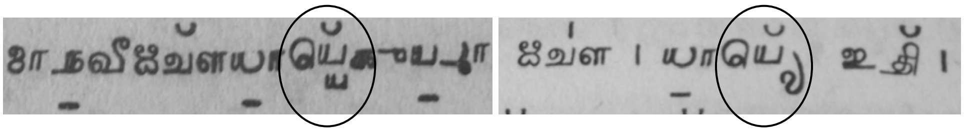
§3. Rare real cases of more than one C2-conjoining form

As seen above, Cibu's examples of "cursively connected C2-conjoining forms" such as for /gra/ are in fact ligatures and hence there is no case of two C2-conjoining forms there. However there is at least one consonant in Malayalam and three in Grantha which have attested variation in their C2-conjoining forms:

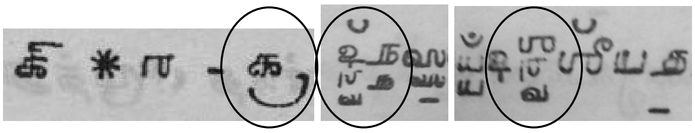
1) Malayalam VA, which may be written either as sa sub-base VA or more commonly as a post-base reduced glyph as in /l̲va/:

 and 

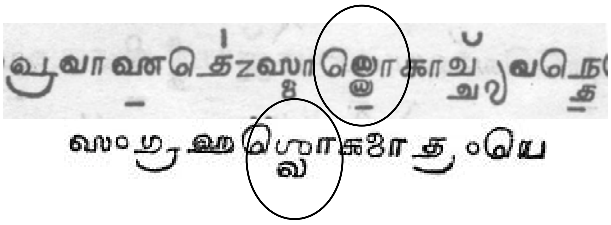
2) Grantha YA, which behaves likewise:



3) Grantha RA, alternating between a RA-vattu and a sub-base nominal glyph:



4) Grantha LA, alternating between a sub-base nominal glyph and a sub-base Tamil-style LA (note that Malayalam always uses this sub-base Tamil-style LA):



I had already recorded the Grantha forms in L2/10-259, my second follow-up to my Grantha proposal. It is not surprising that such variation should be seen in Malayalam as well.

The point is, there is no mechanism to distinguish between such variations in C2-conjoining forms. For Grantha I had proposed the use of VSs in my earlier doc but it was rejected. My proposal was as a theoretical exercise, and there is really no urgent need to be able to distinguish between variant C2-conjoining forms of Grantha in plaintext. I believe the same is true for Malayalam as well. Font features are quite sufficient in this regard.

-o-o-o-