Submitted by N. Ganesan, 2006-04-17

Dravidian script “markers” code-points in Unicode

This note is about the allocation of "markers" for the Dravidian scripts such as Tamil, Telugu, Kannada, Malayalam in Unicode code-charts. The "markers" will be very useful for working with historic letters, and esp. the cillakSaram consonants in Malayalam since the special cillu-markers will free up ZWxJ functionality, and no special properties in ZWJ/ZWNJ just for the case of Malayalam in the family of Indian scripts will be ineeded in TUS and implementation. Here are 5 examples of 'marker' codepoints

(1) Telugu alveolar marker

There are two historic Telugu letters not used in current print books, but found in grammar books. Usually called TCA and TJA due to alveolar modification upon Telugu letters, CA and Ja respectively.

TCA and TJA can be generated by "Telugu alveolar marker" sign with an annotation something like "This Telugu sign works only on CA and JA". This combining sign, with properties like anusvara, will have a dotted circle.

(2) Telugu abbreviation marker:

In Telugu script, words are abbreviated and shown as the first letter (abugida or vowel) of the word followed immediately by two closely spaced vertical lines. This combining sign, with properties like anusvara, with a dotted circle followed by two vertical lines II will be "Telugu contraction (or abbreviation) marker". There are many example words with Telugu contraction marker listed in books.

(3) Malayalam cillu marker

The Indic list has gone through several inputs on this problem for implementation. I did some research, and I do not recall a glyph for cillu m, that is in line with cillu n, cillu nn, etc., Like samvruthokaram example where Virama properties for just Malayalam alone needs to be changed if we don't have "short u" marker code-point and a corrsponding combining sign, it is better if we do not use special properties for ZWxJ, Virama in the case of Malayalam cillus. Antoine Leca mentioned a cillu-y in Indic list, possibly there are some more cillus (that will be brought to attention). So, the question is: does UTC want to encode, say, 10,12 or 14 code-points for cillus (which will divorce them from their root consonants which is not good linguistically)?

In Unicode, cillu letters of Malayalam can be called as "Malayalam prepausal consonant marker" or "Malayalam cillu marker". This combining sign, with unicode properties like

anusvara, will have a dotted circle. Cillus can come at the end of words and word-medially for ka, na, nna, la, lla, llla, ra, ta and ma Please note that one Unicode cillu marker will provide cillus for all these 9 consonants (otherwise, 9 separate code-points! which will remove them from their root genetic consonants unaccepatble linguistically,) And, cillu marker will do away with any special rules for ZWJ/ZWNJ just for Malayalam among Indian scripts also.

In usual transliteration, cillu_consonant = consonant (roman) followed by a colon : So, eg., cillu_nna (in Malayalam) = n:na in roman, otherwise nna only. In the Rachana document on cillu letters, cillu_nma set (pg. 2 of L2/05-210), can be transliterated as "n:ma" with : representing the cillu-marker codepoint. In word-final position, cillu_r is spoken out as Malayalam letter RR. Transliteration of cillu is done with a : sign.
See Note 12 in
http://homepage.ntlworld.com/stone-catend/trinotes.htm

Please note the distinct cillu-m (which not in the shape of Malayalam anuswaram). The distinct shapes of Malayalam cillus (Ref. : R. Gruenendahl) are given in this document at the end. Cillus shown are for 9 consonants: ka, na, nna, ma, ra, ta, la, lla, llla.
Note the llla-specific cillu in the pdf. Also, the cillu_l and cillu_t can be differentiated with the glyphs given. Take for example, the third glyph for cillu_t and the second one for cillu_l. This is also adhered to in the Library of Congress ALA-LC romanization table:
http://www.loc.gov/catdir/cpso/romanization/malayala.pdf
Of course, the code-points for cillu_l and cillu_t are different.

(4) Malayalam short u marker

Sometimes, especially in North Kerala, words ending with u has a short u indicated in orthography. The word-terminal [consonant] + short u is shown visibly using a virama. This is called saMvRttokaram in Sanskrit/Malayalam, and is used in Malayalam.

Unicode has a Virama based model where the Virama normally deletes/"kills" inherent "a" in "consonants"/akSarams like [ka]. So, in order to make abugidas with short u, no need to stack a Virama after [consonant] + [vowel modifier u] abugidas. That will break the normal Unicode meaning of Virama in Indic scripts, and create an unusual function for Virama only in Malayalam. Similarly, no need to use [ku], ZWxJ followed by Virama for saMvRthokaram u in Malayalam. Typically, samvRuthokaram u is transliterated as [consonant] + u with breve (U+016D).
http://homepage.ntlworld.com/stone-catend/trinotes.htm

It can be encoded as "Malayalam short u marker" which works with words ending as [consonant] + u This "Malayalam short u marker" is the last code-point/sign in a word. There is a separate document on the importance of Samvruttokaram in Malayalam script authored by Drs. Chtrajakumar and Gangadharan. This combining "short u marker" in Malayalam Unicode, with properties like anusvara, will have a dotted circle. It has to work with [consonant]+ [u-vowel modifier].

(5) Malayalam gemination marker:
It has a ramp/saw_tooth shape which ligates at the bottom in conjuncts like cca and rvva, etc., In transliteration, the geminate marker can be represented for cca as c.ca .

These 5 examples are given to illustrate the use of "marker" codepoints in Unicode among Dravidian scripts. There are some more markers that can be added in Unicode over time.

Naga Ganesan
Houston, TX

PS:
From Indic list post

A. Leca wrote:
>Until now, it is not known if cillu-l (and,
>as far as I can see, your putative cillu-t as well)
>should be encoded as <0D31, 0D4D, 200D>
>or U+0D7B. But nothing more.

Please note that there is *no* separate cillu_rr,
so code point for a Malayalam cillu with 0D31
does *not* arise. Refer ALA-LC romanization
or ISO 15919 etc., In word-final position, cillu_r
is spoken out as Malayalam letter RR. So,
in word-final position, cillu_r is transliterated
as _r (r with an underline) in Roman script.
But it is still a cillu_r like the rest of cillu_r's.
ISO 15919, ALA-LC tables, and other books
do not give any cillu_rr.

Cillu marker code-point is highly recommended
(1) for not imposing new properties on ZWxJ
just for Malayalam among Indic scripts
(2) cillus are too many to be given separate code
points (Future may throw up more cillus)
which will move them away from root consonants
(Chitrajakumar/Gangadharan document).

N. Ganesan

Reinhold Grünendahl

South Indian Scripts in Sanskrit Manuscripts and Prints

Grantha Tamil – Malayalam – Telugu – Kannada – Nandinagari

Prepausal Consonants

k	ക്ക & ക്ക & ഹ്ക	r	ർ & ർ
ṇ	ൺ & ൺ	l	ൽ & ൽ
t	ത്ത & ത്ത & ന്ത	ḷ	ൿ & ൿ
n	ൻ & ൻ	ḻ	ഴ്
m	മ്		

pg. 92,
(2001 : Harrassowitz Verlag, Wiesbaden, Germany)

Note the chillu-m glyph.

All these [Consonant] can be encoded by a Malayalam followed by "chillu marker" (one codepoint).

Will free up special use of ZWJ etc. in Malayalam.

No chillu for RR. So, just word-finally chillu-r is ർ (=RR)