L2/07-393

Correction to L2/07-090, L2/07-129 on how Malayalam Chillus used in boundary cases

Cibu C. Johny
cibu@google.com
2007-10-15
version 1.0

The need for correction

The recommendations in documents L2/07-090 and L2/07-129 call for using Virama immediately following a Chillu. However, in the Indic model, Virama acts as the vowel remover for a consonant with default vowel /a/. The Chillus does not have an inherent vowel. So <chillu, virama> sequence could be violating the Indic model.

This document retracts the above usage and lists how various visuals can still be achieved while being consistent to Indic model. Following sections describes the changes in Unicode 5.1 and clarifications after making the correction.

Changes incorporated in Unicode 5.1

Effective of Unicode 5.1, representation of Malayalam Chillus would change and the encoding of /nta/ conjunct and dot style Reph would be clarified. The new representation could be summarized as below:

- All the Chillus are atomically encoded
- The /nta/ conjunct is represented by <NA, Virama, RRA>
- Dot reph is represented by <RA, Virama, Consonant> (no change from Unicode 5.0)

The tables 1, 2 and 3 detail these changes with examples. Left side of the tables describes what is being changed in Unicode 5.0 and right side describes what the final encodings in Unicode 5.1 would be.

What happens to existing sequence for Chillu?

Rendering systems should render <Consonant, Virama, ZWJ> as chillus, where consonants is NNA(0D23) NA(0D28) RA(0D30) LA(0D32) or LLA(0D33).

In the above codepoint context, ZWJ should not be stripped off as far as possible.

However, <Consonant, Virama, ZWJ> is highly discouraged from being used for chillu.

Changes related to representation of Chillus:

Table 1	Visual	Codepoint sequence in 5.0	Codepoint sequence in 5.1.0 (atomic encoding)	Sample Words
C1	ൺ	ണ, ്, ZWJ NNA, VIRAMA, ZWJ (0D23, 0D4D, 200D)	ൺ (0D7A)	ആൺ /aann/
C2	ൻ	ന, ്, ZWJ NA, VIRAMA, ZWJ (0D28, 0D4D, 200D)	ൻ (0D7B)	അവൻ /avan/
C3	ർ	ര, ്, ZWJ RA, VIRAMA, ZWJ (0D30, 0D4D, 200D)	ർ (0D7C)	ഞായർ /njaayarr/
C4	ൽ	ല, ്, ZWJ LA, VIRAMA, ZWJ (0D32, 0D4D, 200D)	ൽ (0D7D)	ഉൽപ്പത്തി /ulpatti/
C5	ൾ	ള, ്, ZWJ LLA, VIRAMA, ZWJ (0D33, 0D4D, 200D)	ൾ (0D7E)	അവൾ /avall/
C6	ൿ	Not defined	ൿ (0D7F)	ദൃക്സാക്ഷി /drksaakssi/

Clarification on Conjunct rendering fallbacks:

Table 2	Levels of rendering fallbacks for the code sequence <C1, VIRAMA, C2> In the examples, C1 = ന (0D28) and C2 = മ (0D2E)					
	Unicode 5.0				Unicode 5.1	
	Microsoft Kartika Interpretation		Debian Rachana Interpretation			
	Visual	Example	Visual	Example	Visual	Example
F1	Conjunct of C1 and C2	ന്മ	Conjunct of C1 and C2	ന്മ	Conjunct of C1 and C2	ന്മ
F2	C1, Virama, C2	ന്മ	Chillu-C1, C2	ൻമ	C1, Virama, C2	ന്മ
F3			C1, Virama, C2	ന്മ		

Changes specific to Chillu NA(ൻ) and RRA(ർ) combinations:

Table 3	Visuals	Shortest Codepoint sequence in 5.0		Sequence in 5.1.0	When joiner is lost	Fallback	Sample Words
		Microsoft Kartika interpretation	Rachana, Varamozhi, etc.				
N1	ന്‍ര	ന, ്, ര (0D28, 0D4D, 0D31)	ന, ്, ZNWJ,ര (0D28, 0D4D, 200C, 0D31)	ന, ്, ര (0D28, 0D4D, 0D31)			ഉപ്പുമാവിന്‍രവ /uppumaavin‍-rava/
N2	ൻര	ന, ്, ZWJ, ZWNJ, ര (0D28, 0D4D, 200D, 200C, 0D31)	ന, ്, ZWJ, ര (0D28, 0D4D, 200D, 0D31)	ൻ, ര (0D7B, 0D4D, 0D31)			ഹെൻറി /heṇri/
N3	ൻര	Not defined		ന, ്, ZWJ, ര (0D28, 0D4D, 200D, 0D31)	N4	N4	എൻര /ente/
N4	ന്റ	ന, ്, ZWJ, ര (0D28, 0D4D, 200D, 0D31)	ന, ്, ര (0D28, 0D4D, 0D31)	ന, ്, ര (0D28, 0D4D, 0D31)		N3	എന്റ /ente/

Position for the left part of the vowel sign, if used.

Following table indicates the changes related to dot-reph representation:

Table 3	#	Visual	Codepoint sequence in 5.0	Codepoint sequence in 5.1.0	Fallback	Sample Words
ര(RA - 0D30) with യ(YA - 0D2F)	D1	ര്യ	Not defined	ര, ്, ZWJ, യ (0D30, 0D4D, 200D, 0D2F)	D4	No examples found yet.
	D2	ർയ	ര, ്, ZWJ, യ (0D30, 0D4D, 200D, 0D2F)	ർ, യ (0D7C, 0D2F)		ചതുർയുഗം /chaturyugam/
	D3	ര്യ	ര, ്, യ (0D30, 0D4D, 0D2F)	ര, ്, യ (0D30, 0D4D, 0D2F)	D4	ഭാര്യ /bha:rya/
	D4	ര്‍യ	ര, ്, ZWNJ, യ (0D30, 0D4D, 200C, 0D2F)	ര, ്, ZWNJ, യ (0D30, 0D4D, 200C, 0D2F)		No examples found yet.
	D5	ര്‍യ്യ	ര, ്, യ, ്, യ (0D30, 0D4D, 0D2F, 0D4D, 0D2F)	ര, ്, യ, ്, യ (0D30, 0D4D, 0D2F, 0D4D, 0D2F)	ര്യ്യ (D4)	ഭാര്‍യ്യ /bha:rya/
	D6	ർയ്യ	ര, ്, ZWJ, യ, ്, യ (0D30, 0D4D, 200D, 0D2F, 0D4D, 0D2F)	ർ, യ, ്, യ (0D7C, 0D2F, 0D4D, 0D2F)		No examples found yet.
ര(RA - 0D30) with വ(VA – 0D35)	D7	ര്‍വ	ര, ്, വ (0D30, 0D4D, 0D35)	ര, ്, ZWJ, വ (0D30, 0D4D, 200D, 0D35)	D10	നിര്‍വാണം
	D8	ർവ	ര, ്, ZWJ, വ (0D30, 0D4D, 200D, 0D35)	ർ, വ (0D7C, 0D35)		പാർവതി /bha:rya/
	D9	രാ	Not defined	ര, ്, വ (0D30, 0D4D, 0D35)	D10	വരോ /varvo/ (colloquial)
	D10	ര്‍വ	ര, ്, ZWNJ, വ (0D30, 0D4D, 200C, 0D35)	ര, ്, ZWNJ, വ (0D30, 0D4D, 200C, 0D35)		No examples found yet.
	D11	ര്‍വ്വ	ര, ്, വ, ്, വ (0D30, 0D4D, 0D35, 0D4D, 0D35)	ര, ്, വ, ്, വ (0D30, 0D4D, 0D35, 0D4D, 0D35)	ര്വ്വ (D10)	പാര്‍വ്വതി /bha:rya/
	D12	ർവ്വ	ര, ്, ZWJ, വ, ്, വ (0D30, 0D4D, 200D, 0D35, 0D4D, 0D35)	ർ, വ, ്, വ (0D7C, 0D35, 0D4D, 0D35)		സർവ്വം /sarvvam/
RA with other consonants	D13	ര്‍ഷ	ര, ്, ഷ (0D30, 0D4D, 0D37)	ര, ്, ഷ (0D30, 0D4D, 0D37)	ർഷ	വര്‍ഷം /varsham/

Rendering order in some of the edge case conjuncts

Following tables clarify the storage representation of specific Malayalam visuals for which the representation is not apparent from the Indic model. These visuals are based on the assumption that, the font has all the possible conjuncts. If the font lacks a specific conjunct, the fallback visual is also specified.

U-sign(0D41) variants and Samvruthokaram

Table 4	Visuals related to Samvrutho-karam	Storage Representation	Fallback	Sample Words	Notes
U1	നു OR നൗ	ന, ു (0D28, 0D41)	U5	മനു /manu/	Rendering engine decides the visual
U2	ന്നു്	ന, ZWNJ, ു, ് (0D28, 200C, 0D41, 0D4D)	U2	അവന്നു് /avanṉ/	Requesting one specific visual
U3	നൗ്	ന, ZWJ, ു, ് (0D28, 200C, 0D41, 0D4D)	U1	അവനൗ് /avanṉ/	Same reading as U2; Requesting one specific visual
U4	ന്	ന, ് (0D28, 0D4D)	None	അവന് /avanṉ/	
U5	നു	ന, ZWNJ, ു (0D28, 200C, 0D41)	U6	മനു /manu/	
U6	നൗ	ന, ZWJ, ു (0D28, 200D, 0D41)	U5	മനൗ /manu/	Same reading as U3

VA(0D35) and RA(0D30)-sign forms

Table 5	Visuals related conjoining വ	Storage Representation	Fallback	Sample Words	Notes
V1	ന്വ	ന, ◌്, വ (0D28, 0D4D, 0D35)	ന്വ	തന്വി /tanvi/	
V2	ത്സ്വ	ത്സ, ◌്, വ (0D34, 0D4D, 0D35)	V3	താത്സ്വാരം /ta:r'va:ram/	Same reading as V3. Fallback cannot be ത്സ്.
V3	ത്സ്‌വ	ത്സ, ◌്, ZWNJ, വ (0D34, 0D4D, 200C, 0D35)	None	താത്സ്‌വാരം /ta:r'va:ram/	
V3	ത്സ്വ	ത്സ, ZWJ, ◌്, വ (0D34, 200D, 0D4D, 0D35)	V3	താത്സ്വോ /ta:rvo:/	
V4	ഉ്വ	വ, ◌്, വ (0D35, 0D4D, 0D35)	വ്‌വ	ഉ്വ /uwwa/	
V5	ഉ്വോ	വ, ZWJ, ◌്, വ (0D35, 200D, 0D4D, 0D35)	വ്‌വ	ഉ്വോ /uw'wo:/	

Conjuncts, as in V2, is formed only in following regular expression context: /[ത്സ]◌്(വ|ര)/. That is, in perl like syntax:
`/[\x{0D34}\x{0D2F}]\x{0D4D}(\x{0D35}|\x{0D30})/`.

Fallbacks for these forms will not be with VA or RA subjoining form. Instead, it will be with explicit virama as in V3.

Common conjunct that are often misrepresented in currently available fonts

Table 6	Common Conjuncts	Storage Representation	Sample Words
L1	ങ്ക	ങ, ്, ക (0D19, 0D4D, 0D15)	പങ്ക /panka/
L2	ഞ്ച	ഞ, ്, ച (0D1E, 0D4D, 0D1A)	ഇഞ്ച /incha/
L3	ണ്ട	ണ, ്, ട (0D23, 0D4D, 0D1F)	മണ്ട /manda/
L4	ന്ത	ന, ്, ത (0D28, 0D4D, 0D24)	മൊന്ത /monta/
L5	മ്പ	മ, ്, പ (0D2E, 0D4D, 0D2A)	പമ്പ /pampa/