

Unicode Public Review Issue #66: Encoding of Chillu Forms in Malayalam

L2/05-085

Author: Cibu C Johny
Email: cibu (at) yahoo.com
Date: May 2, 2005

Abstract

This document proposes two solutions to the public review issue #66 and suggests introducing code points for Chillu letters as the preferred solution. Also describes various issues with the current representation of Chillu letters.



Conventions used in this document

- Pronunciations are transliterated to Latin and indicated by single quotes (``).
- Unicode code points for characters used in this document:
 - A - U+0D05
 - U - U+0D09
 - TA - U+0D24
 - NA - U+0D28
 - MA - U+0D2E
 - RA - U+0D30
 - LA - U+0D32
 - VA - U+0D35
 - Virama - U+0D4D

Proposed solution to the Issue #66

Since Chillu-NA and NA + visible Virama can give different meaning to a word, we cannot let the rendering system choose the output of NA + Virama. Here are my preferences in the decreasing order:


- 1) Explicitly encode Chillu characters. Various issues are discussed in detail below.
- 2) <NA, Virama> (without any joiner) should be mapped to NA with visible Virama since it enforces uniformity. That is, Consonant + Virama will always produce visible Virama symbol, irrespective of whether the consonant is capable of forming a Chillu or not. If we follow this, both of following sample combinations without any joiner will have visible Virama symbol.


VA + Virama = 
 NA + Virama = 

Issues in current representation of Chillu letter as Consonant + Virama + ZWJ

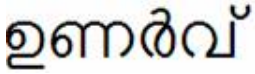
- 1) ZWJ and ZWNJ are supposed to be font directives, directing a font to select from two or more semantically same renderings. In case of Malayalam, this is no longer true. ZWJ becomes an alien language construct introduced to Malayalam by Unicode to produce Chillu letters. Thus, it is possible to produce two semantically different words, which differ only by ZWJ in their Unicode representation. In the following examples, words differ only by ZWJ.


Example 1.1:

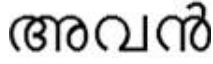
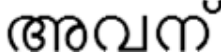
 This word is with visible Virama after NA and pronounced as 'avanu'. This word means "for him".

 This word is with Chillu NA and pronounced as 'avan'. This word means "he".

Example 1.2:

 This word is with Chillu RA. This is a valid word in Malayalam.

 This word is with RA in full form and VA in C2-conjoining form. This is NOT a valid word in Malayalam.

- 2) When a word is searched in Unicode text, the search algorithm should ignore ZWJ and ZWNJ because it should not care about the rendering of the word. From the argument 1, Malayalam can have words differ by a joiner alone. So the search for, say,  will return  also. That is plain wrong.

As a work around, the search algorithm could match joiners, only in the case of Malayalam. Then the algorithm will not match those words that are semantically same but rendered differently by using or

omitting a joiner (ZWJ or ZWNJ). For example, search for 

will not match ഉത്സവം, if later is written using ZWNJ.

This issue has repercussions beyond the search algorithm. Future development of language tools (for example grammar checker) for Malayalam will be impeded by this inconsistency.

3) Confusion on whether ൽ (Chillu LA/TA) belongs to LA or TA.

For Sanskrit words used Malayalam, ത (TA) is pronounced as it is, only when a vowel or semi-vowel comes after it. For all other occasions, it is pronounced as ല (LA).

An example would be ഉത്സവം ('ulsavam'). Even though, it's Sanskrit originated form is ഉത്സവം ('uthsavam'), it is pronounced in Malayalam as ഉല്സവം ('ulsavam').


This means, Chillu form of ത (TA) should be pronounced as if it is Chillu form of ല (LA). Thus, ൽ (Chillu LA/TA) is in a very curious situation:

Grapheme level: Graphically it is Chillu of ത (TA).

Character level: It can represent the characters - either ത (TA) or ല (LA).

Phoneme level: Its pronunciation is the Chillu of ല (LA).

Since Unicode is standardizing characters, this Chillu has to be considered the Chillu of both LA and TA. However, this will lead to two representations of a word with same rendering.

- 4) Chillu of a consonant is phonetically different from its C1-conjoining form without inherent  (A). This is in direct contrast with that Unicode assumption and this inconsistency produces issues described in arguments 1 and 2.


Consider the combination: Vow + CC + Con

Vow - a vowel

CC - a consonant capable of forming Chillu


Con - a consonant

When CC takes its Chillu form, it joins more with Vow. This effect produces a noticeable small stop between CC and Con.

When CC without inherent  (A) forms a conjunct ligature with Con, it is pronounced together with Con without any pronunciation stop in-between.

Two sample letter combinations to show the pronunciation difference:

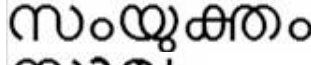
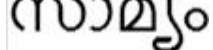
 - RA in Chillu form

 - Full form of RA with C2-conjoining form of VA

- 5) Chillu of a consonant can be treated like Anusvara

R. Raja Raja Varma states in his [Keralapanineeyam](#) (which is the foremost grammar book of Malayalam) "Anusvara is the Chillu form of MA". This is essentially same as saying Malayalam Anusvara and other Chillu characters share same properties.


As a demonstration of that fact, we can see that, the half-stop phonetic property described in argument 4 is same for Anusvara and other Chillu characters. Following two sample letter combinations show the pronunciation similarity with the example in argument 4:

Background

- A) Overloading of visible Virama in Malayalam

Following are the functions of Visible Virama:

A.1) At end of a word, it acts as quarter vowel  (U). Example:

അവൻ ('avanu')

A.2) In the middle of a word, it means the consonant before is forming a conjunct with consonant after. For example, consider

Sabdam').

In this context, it does not produce any sound what so ever.

Functionality-(A.2) has been overloaded with this grapheme when typesetting friendly new orthography has been introduced. Unicode recognizes functionality-(A.2) alone with visible Virama of Malayalam. This contributes to the problem that Unicode representation of the

words അവൻ ('avan') and അവൻ ('avanu') differ only by a joiner (ZWJ or ZWNJ). However, they have two different meanings.

Reference: [kEraLapaaNineeyam, peeThika - A. R. Raja Raja Varma](#)