Proposal for a Mechanism to Select from Multiple Malayalam C2-conjoining Forms

Cibu Johny, Google Inc.
November 2, 2013

This document is an updated version of PRI 250[1].

Orthography reform[2] in 1971 divided the Malayalam script into traditional and reformed orthographies with differing typographic conventions. This created multiple C2-conjoining forms for some Malayalam consonants. This proposal provides a mechanism to select from multiple C2-conjoining forms, by extending the usage of ZWJ and ZWNJ with VIRAMA, described in PRI 37.

This document incorporates the feedbacks received for PRI 250 - specifically, the cases 3 & 4 are withdrawn, relationship to PRI 37 is clarified, and the motivational section is elaborated. An FAQ is provided at the end to address the common questions.

Introduction

In Malayalam there are two prevailing orthographies - traditional and reformed - both are written as digital text using same Malayalam encoding. Today the difference between them is manifested by both spelling and typographic conventions (i.e., renderings). Traditional orthography rendering accommodates a lot more C2-conjoining ligatures, while reformed orthography would instead use nominal consonants separated by visible virama (*chandrakkala*). Along with that, for the vowel signs of U, UU, and Vocalic R, RR and for the C2-conjoining form of RA, a reformed orthography font would use visually disconnected forms instead of the cursively connected forms.

Examples of multiple C2-conjoining forms

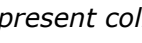
Following are some representative examples for the multiple C2-conjoining forms that occur due to various reasons. Please observe that, these forms differ in the level of cursiveness or in the closeness to the base consonant.

K-RA

Differ in traditional and reformed typographic conventions. This represents most common multiple C2-conjoining forms.

LLL-VA

[1] http://www.unicode.org/review/pri250/
[2] Details of the orthography reform of 1971:
http://en.wikipedia.org/wiki/Malayalam_alphabet#Orthography_reform

First form is used in traditional renderings of the words like  *(V-AA LLL-V VIRAMA) and the corresponding reformed rendering uses visible virama. Second form is used to represent colloquial tongue as in*  *(V-II LLL-V-OO).*

Y-YA

First form is used by both traditional and reformed orthographies. Second form is used for renderings of circa 1900 CE. Please observe that, the separation of the conjoining form from the base does not always mean reformed orthography rendering.

Spelling and rendering congruence

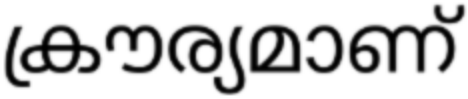
The difference between the two Malayalam orthographies, comes with spelling and rendering changes. There are traditional and reformed spellings, but also traditional and reformed orthography renderings, and mixing traditional spelling with reformed rendering, or vice versa, is going to look bad. For example, consider the word /krauryamāṇŭ/ in its traditional spelling:

*Traditional spelling <**K-RA** AU-sign Dot-Reph Y-YA MA AA-sign **NNA U-sign Virama**> is rendered with the traditional orthography font Meera*[3]

This traditional orthography text is rendered in reformed orthography rendering as below. Orthography rendering mismatch is as indicated by bold in the character sequence.

Traditional spelling; mismatched rendering with the reformed orthography font Noto Malayalam[4]

In the reformed spelling below, the difference in spelling is indicated in colors:



Modern spelling <K-RA AU-length-mark R-YA MA AA-sign NNA Virama> rendered with the reformed orthography font Noto Malayalam

In order to avoid the mismatch between traditional spelling and reformed orthography rendering, it should be possible to create a font that supports both orthographies.

When the spelling used is unambiguously traditional, that font might be able to detect that and provide a traditional rendering, even if the font were primarily intended for reformed typography. For example, <Consonant, U-SIGN, VIRAMA> sequence can always be rendered in traditional orthography. However, for differing C2-conjoining forms like that for the K-RA ligature, there is no spelling difference and hence the font is unable to choose right rendering for that ligature. Since the rendering implementations are unable to determine which presentation is intended, today a reformed orthography font cannot avoid the above described mismatch.

[3] http://download-mirror.savannah.gnu.org/releases/smc/fonts
[4] https://code.google.com/p/noto/

The reverse of the above scenario also can be imagined; that is, to request disconnected C2-conjoining form when cursively connected form is the default in the font, for example, as seen in the Y-YA ligature.

Case study: attempt to harmonize traditional spellings in reformed orthography fonts

Above described issue of mismatch between spelling and typographic conventions must have been a pain point for the font maintainers of Lohit Malayalam and Raghu Malayalam; both are reformed orthography fonts. They invented presentation forms to accommodate traditional spellings they might have to render, but that would never have been encountered in reformed spellings. For the common traditional sequence of <U-SIGN, VIRAMA>, the presentation form they invented is as below.[5]

<NNA, U-SIGN, VIRAMA> rendering by Raghu Malayalam[6]

This approach to introduce totally new presentation forms into the script might be too naive. However, it demonstrates the definite need in the community for an on-demand traditional rendering for the traditionally spelled text.

Glyph variant or orthography difference

The fact that, Malayalam has both traditional and reformed orthographies that needed separate typographic treatment, is well established in the expert and user communities. Since the reform has happened in the near past, both orthographies have significant following.

The OpenType specifications defines two separate Language System tags, MAL and MLR, for traditional and reformed scripts respectively[7]. By far Malayalam is the only script in Indic to have such clear distinction made between orthographies.

Along with the spelling differences that happen in traditional versus reformed orthographies, it is hard to see this as just a few glyph variations. It is much more consistent to view the phenomena as orthography/spelling distinction with differing systems of typographic conventions.

Historic fractions and numbers

Traditional fuller conjuncts have other usages as well. Historically conjuncts or letters are used to represent fractions and numerals. For example, traditional orthography conjunct P-TA, represents 1/320 along with recently encoded Malayalam fractions.

---

[5] The reasoning behind this addition might have been like this: Today there is no way to indicate the suitable C2-conjoining form in text. Since the font was intended for the reformed orthography, it had to use reformed rendering for conjuncts like K-RA. So only change that could be imagined was, to introduce a new presentation form, that is reformed orthography friendly, for the unambiguously traditional spellings; i.e., <U-SIGN, VIRAMA>.

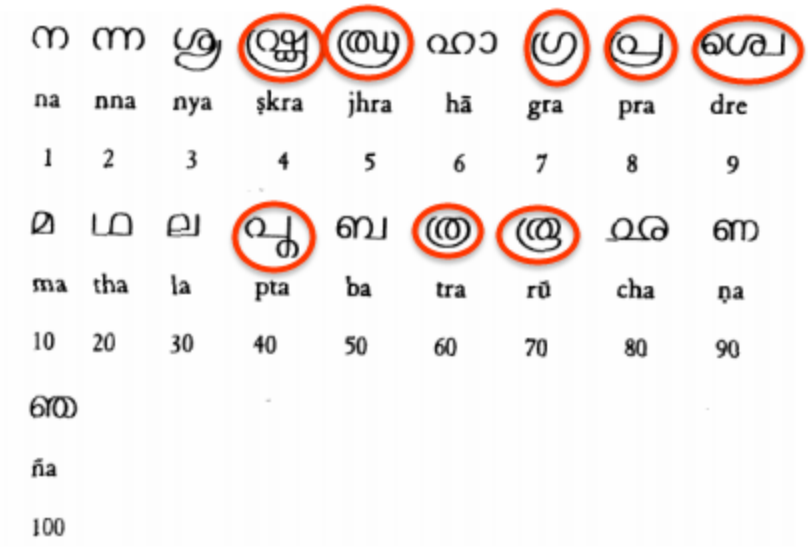
[6] http://download-mirror.savannah.gnu.org/releases/smc/fonts This presentation form was in the font for years and it is just rescinded in the latest version 6.0, released on Oct 19, 2013.

[7] http://www.microsoft.com/typography/otspec/languagetags.htm

	Glyph	x/320	Value	Words	Malayalam name
1.	ൟ	1/320	1/320	one three-hundred-and-twentieth	muntiri
2.	ൡ	2/320	1/160	one one-hundred-and-sixtieth	arakkāṇi
3.	ൠ	4/320	1/80	one eightieth	kāṇi

Historic fractions[8]. Conjunct that was was not atomically encoded is circled in red.

Once popular, alternative number coding scheme called Akṣarappaḷḷi system, uses many traditional orthography conjuncts.



Akṣarappaḷḷi system[9]. Traditional orthography conjuncts are circled in red.

These conjuncts need not be atomically encoded since they are proper Malayalam letters or conjuncts. However, they need to be in traditional orthography rendering to effectively convey the meaning. Just like any numeral system, this system also should be possible in plain text.

Current status

Indic conjoining model favors the full conjunct for a given set of characters. Half forms are produced by ZWJ, which acts like an invisible consonant that would always try to form a ligature with the consonant on the other side of the Virama.

So the <Consonant, VIRAMA, ZWJ> sequence provides the half form of the initial consonant. Similarly the C2-conjoining form is specified as <ZWJ, VIRAMA, Consonant>. The <Consonant, Virama, ZWNJ, Consonant> sequence is used to create an visible virama and the sequence <Consonant, ZWNJ, Virama, Consonant> is left undefined. See the Figure 9-7 in the standard version 6.2 for examples:

[8] http://www.unicode.org/L2/L2013/13051r-malayalam-fractions.pdf (PDF page 1)
[9] http://www.unicode.org/L2/L2013/13051r-malayalam-fractions.pdf (PDF Page 10)

The rendering fallback sequence employed is:

1. Full conjunct
2. Conjunct with half-forms
3. Consonants in nominal form with visible Virama.

Implication of PRI 37 resolution

The resolution of PRI 37[10] established the overarching Indic conjoining behavior model with Virama and joiners. However, with respect to Malayalam, some cases are left undefined. For example, it does not account for multiple C2-conjoining forms that could occur.

Moreover, PRI 37 does not distinguish between cursively connected and disconnected conjoining forms. See the example of cursively connected Oriya K-RA cited in the PRI. This example could imply that, if there are multiple C2-conjoining forms, the PRI prefers cursively connected form with <ZWJ, VIRAMA> sequence.

The < ZWJ, VIRAMA > solution also avoids problems with display modes that show control-picture glyphs for control characters. Consider again an example using Oriya:

Sequence	Normal display	Possible display with control-picture glyphs shown
< KA, ZWJ, VIRAMA, RA >	କ୍ର	କ୍× ◌
< SPACE, ZWJ, VIRAMA, RA >	◌୍ର	◌୍× ◌

Table 13. Rendering using control-picture glyph for ZWJ

Accepted resolution of PRI 37; page 13

PRI 37 leaves the sequence <ZWNJ, VIRAMA> undefined. Probably, Malayalam can use this sequence as well. That is, <ZWNJ, VIRAMA> can produce discrete C2-conjoining forms, while <ZWJ, VIRAMA> producing traditional, cursively connected C2-conjoining forms. This goes well with the general principle of ZWNJ; that is to obstruct the fully ligatured or cursively

[10] http://www.unicode.org/review/pr-37.pdf

connected behavior.

Harbuzz and Uniscribe with traditional fonts

The behavior of the two popular traditional orthography fonts Rachana and Meera with Harfbuzz[11], is to cursively connect C2-subjoining forms, irrespective of whether the joiner is ZWJ or ZWNJ.

Meera[12] traditional font rendering <KA, ZWJ/ZWNJ, VIRAMA, RA> with Harfbuzz

Uniscribe[13] also does not distinguish between ZWJ or ZWNJ. It produces disconnected C2-conjoining form that is not correctly reordered.

Meera traditional font rendering <KA, ZWJ/ZWNJ, VIRAMA, RA> with Uniscribe

These tests indicate that there is no pre-existing, well established <ZWJ/ZWNJ, VIRAMA> usage in Malayalam.

Proposal

The proposal below is enhanced PRI 37 mechanism in the above indicated manner. Only minimal changes are introduced and it preserves backward compatibility.

Conjoining of consonants in Indic scripts follows a three-level precedence hierarchy; a dead consonant C_d followed by a consonant C2 can be displayed in three levels:

1. the combination of C_d and C2 can form a conjunct ligature
2. either C_d or C2 takes on an alternate conjoining form and is combined with the full form of the other consonant
3. C_d is displayed with a visible halant, followed by the full form of C2

If **no joiners are used, font or rendering system decides the level** to be used for the specific Virama involving sequence. It can fallback from level 1 to level 2 and then to level 3 when a conjoining form is not supported or does not exist.

The characters ZWJ and ZWNJ can direct the possible renderings as follows:

- For all Indic scripts, ZWNJ can be used in a sequence <C1, virama, ZWNJ, C2> to explicitly restrict the display to the level-3 alternative, the visible halant form.
- For a C1-conjoining consonant, ZWJ can be used in a sequence <C1, VIRAMA, ZWJ, C2> to restrict the display to level 2 or level 3. Specifically, this sequence requests the half form of C1, to be combined with the full form of C2. If C1 has no half form, then fallback to the level 3 display is used.

[11] version 0.9.23, released on Oct 28, 2013
[12] version 6.0, released on Oct 26, 2013. Tested with Rachana version 6.0 as well with the similar results.
[13] version 6.3.9431.0 (Windows 8.1)

- For a C2-conjoining consonant, ZWJ can be used in a sequence <C1, ZWJ, VIRAMA, C2> to restrict the display to level 2 or level 3. Specifically, this sequence requests the sub- or post-base form of C2, to be combined with the full form of C1. If C2 has no sub- or post-base form, then fallback to the level 3 display is used. **If the script allows more than one C2-conjoining forms, then the fuller or cursively connected form is selected.**
- **[*Addional rule*]** For a C2-conjoining consonant, ZWNJ can be used in a sequence **<C1, ZWNJ, VIRAMA, C2>** to restrict the display to level 2 or level 3. Specifically, this sequence requests the sub- or post-base form of C2 **that is not cursively connected**, to be combined with the full form of C1. If C2 has no **discrete** sub- or post-base form, then fallback to the level 3 display is used. **If the script allows more than one C2-conjoining forms, then the shallow or disconnected form is selected.**
- For a C1-conjoining consonant, the sequence <C, VIRAMA, ZWJ> can be used to display the half form of C in isolation.
- For a C2-conjoining consonant, the sequence <SPACE, ZWJ, VIRAMA, C> **or <SPACE, ZNWJ, VIRAMA, C>** can be used to display the **respective** sub- or post-base form of C in isolation.

Following two cases illustrate this additional semantics with examples. Please note that, the sequences of the form <Consonant, VIRAMA, ZWJ> which were formerly used for requesting Chillus are not used for any of the cases.

Usage examples with *<ZWJ, VIRAMA>*

Malayalam C2-conjoining ligatures can be either a subjoining or a post-base. See S-KA conjunct for the subjoining form:

സ്ക → സ്ക

The T-SA conjunct can produce the post-based C2-conjoining form[14]:

ത്സ → ത്സ

Examples showing conjoining behavior with the **reformed** orthography default rendering:

SA + VIRAMA + KA → സ്ക

SA + **ZWJ** + VIRAMA + KA → സ്ക
(KA has only one C2-conjoining form)

TA + VIRAMA + SA → ത്സ

[14] Examples of some common conjuncts that do not follow C2-conjoining form: NG-KA, NG-NGA, M-PA

TA + **ZWJ** + VIRAMA + SA → 

KA + VIRAMA + RA → 

KA + **ZWJ** + VIRAMA + RA → 

LLLA + VIRAMA + VA → 

LLLA + **ZWJ** + VIRAMA + VA → 

YA + VIRAMA + YA → 

YA + **ZWJ** + VIRAMA + YA → 

Examples showing the behavior for the same sequences with the **traditional** orthography default rendering:

SA + VIRAMA + KA →

SA + **ZWJ** + VIRAMA + KA → 
(KA has only one C2-conjoining form)

TA + VIRAMA + SA →

TA + **ZWJ** + VIRAMA + SA → 

KA + VIRAMA + RA → 

KA + **ZWJ** + VIRAMA + RA → 

LLLA + VIRAMA + VA → 

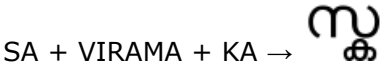
LLLA + **ZWJ** + VIRAMA + VA → 

YA + VIRAMA + YA → 

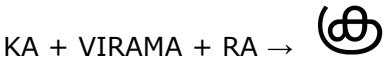
YA + **ZWJ** + VIRAMA + YA → 

Usage examples with *<ZWNJ, VIRAMA>*

Examples illustrating the conjoining behavior with <ZWNJ, VIRAMA> in the **traditional** orthography default rendering for the above sequences:

SA + VIRAMA + KA → 

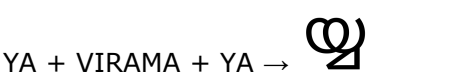
SA + **ZNWJ** + VIRAMA + KA → 
(KA has only one C2-conjoining form)

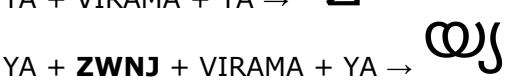
KA + VIRAMA + RA → 

KA + **ZWNJ** + VIRAMA + RA → 

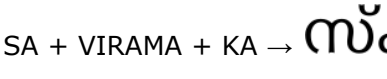
LLLA + VIRAMA + VA → 

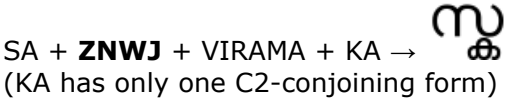
LLLA + **ZWNJ** + VIRAMA + VA → 

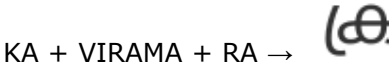
YA + VIRAMA + YA → 

YA + **ZWNJ** + VIRAMA + YA → 

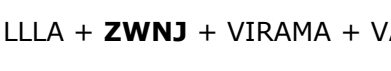
Examples showing the behavior with the **reformed** orthography default rendering:

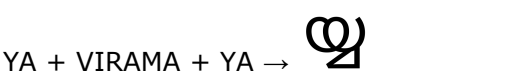
SA + VIRAMA + KA → 

SA + **ZNWJ** + VIRAMA + KA → 
(KA has only one C2-conjoining form)

KA + VIRAMA + RA → 

KA + **ZWNJ** + VIRAMA + RA → 

LLLA + VIRAMA + VA → 

LLLA + **ZWNJ** + VIRAMA + VA → 

YA + VIRAMA + YA →

YA + **ZWNJ** + VIRAMA + YA → 

FAQ

1. Aren't these just glyph variants those are better distinguished in rich text than in plain text?

Existence of dual orthography in Malayalam is a well established fact. Opentype spec defines two different Language System tags for them. Two orthographies are not just a scheme of glyph variations; it includes spelling changes as well. So, plain text or not, making sure traditionally spelled text gets displayed in traditional orthography rendering should be important to many users. More over, some traditional conjuncts are used for representing numerals in once popular alternate number systems. Numerals needs to be represented in plain text.

2. Can't we distinguish this using opentype language system tags MAL and MLR?

Historical fractions and numbers needs to be represented in plain text, just like any other numeral system. Without a mechanism to request cursively connected C2-conjoining form, the conjuncts used for a numeral can get split up in a reformed orthography rendering, obscuring its numeral sematics. Moreover MAL and MLR Language System tags are part of a specific rendering technology. Numeral system needs to be agnostic of that.

3. What about encoding a separate visible virama and get away from all joiner related complications?

It is true that joiner (ZWJ, ZWNJ) semantics may be complicated. At the sametime, separate visible virama is a much larger change and would introduce many confusables. This proposal aims at introducing minimal change to the existing joiner semantics. Also, straightforward visible virama encoding might leave out some of the issues addressed in this proposal. For example, it still may not be possible to distinguish between two possible C2-conjoining forms like the LLL-VA sequence:

4. Are we conflicting with PRI 37?

This proposal does not conflict with PRI 37. It clarifies the PRI 37 for the cases involving multiple distinct C2-conjoining forms. It also employs the otherwise unused <ZWNJ, VIRAMA> sequence to display disconnected C2-conjoining form.

5. What is the implication to rest of the Indic?

See the previous answer. Also, unless an Indic script has multiple C2-conjoining forms, this proposal does not have any implications for it.

6. Doesn't the layout engines need to deviate from the 'Indic model' change to accommodate this?

Since the proposal does not conflict with PRI 37, it does not deviate from the existing Indic model for the layout engines.

7. How huge is the impact to the existing user community?

The negative impact is minimal or even non-existent. Today only two type of joiner usages exists in Malayalam:

<Consonant, Virama, ZWJ> → for the chillus
<Consonant, Virama, ZWNJ, Consonant> → for the visible virama

This semantics are kept intact; hence, practically no negative impact to the user community. At the same time, there will be positive impact since the users gets the ability to select the right C2-conjoining form if they choose to, as the rendering mechanisms implement this.

8. What about the impact to the font developers?

Today only sequence involving joiners that Malayalam fonts explicitly encode is the <Consonant, Virama, ZWJ> for chillu. Since the semantics of this sequence is unchanged, there are no compatibility issues. If a reformed orthography font wants to support additional C2-conjoining forms as per this proposal, it could add additional glyphs for those conjuncts and would need to write substitution rules as per this proposal.

9. Isn't <ZWJ, VIRAMA> already assigned for the discrete subjoining form as in  ?

The PRI 37 does not distinguish between cursively connected or disconnected C2-subjoining forms. In fact, the Oriya example in PRI 37, table 13 is producing cursively connected C2-conjoining form using <ZWJ, VIRAMA>. The current behavior of traditional orthography fonts Meera and Rachana with Harfbuzz is to produce cursively connected subjoining irrespective of ZWJ or ZWNJ. Uniscribe also does not distinguish between ZWJ or ZWNJ. It produces cursively disconnected subjoining that is not correctly reordered. So the proposal does not disturb any well established rendering tradition.