Background Document for PRI 263: Proposal for Sequences to Select From Multiple C2-Conjoining Forms in Malayalam

This PRI is an updated version of the earlier 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 consonant letters. This document proposes a mechanism to select from multiple C2-conjoining forms, by extending the usage of ZWJ and ZWNJ with VIRAMA. Existing best practice is described in PRI 37[3].

This document addresses the feedbacks received for PRI 250 - specifically, sub-proposals 3 & 4 are withdrawn, relationship to PRI 37 is clarified, and the motivational section is elaborated.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*). Reformed orthography rendering uses visually disconnected forms for the vowel signs of U, UU, and Vocalic R, RR and for the C2-conjoining form of RA, while traditional rendering uses the cursively connected forms.

Examples of multiple C2-conjoining forms

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

Y-VA

ക്യ   യ്വ

P-RA

പ്ര   പ്ര

---

[1] http://www.unicode.org/review/pri250/
[2] Details of the orthography reform of 1971:
http://en.wikipedia.org/wiki/Malayalam_alphabet#Orthography_reform
[3] http://www.unicode.org/review/pr-37.pdf

Y-YA

 യ്യ   യ്യ

First form is used by both traditional and reformed orthographies. Second form is used for renderings of circa 1900 CE. This example indicates that, the disconnected conjoining form does not always mean reformed orthography rendering.

The need to select C2-conjoining form in plain text

Following sub-sections describe specific scenarios to prefer a particular C2-conjoining form through spelling difference.

C2-conjoining forms with semantics difference

The C2-conjoining forms of some sequences present semantic difference. Following forms of Y-VA sequence is an example of that:

യ്വ   യ്വ

First C2-conjoining form is used in traditional renderings of the words like:

പോയ്വരൂ <P-OO **Y-VA** R-UU> a word meaning *good-bye*.

Attestation:

Malayalam Fifth Reader (1918), page 179. The word മനസ്താപമുണ്ടായ്വരും (manastāpamuṇṭā**yva**ruṃ) shows the example usage of the first C2-conjoining form.

Following alternate shaping with second C2-conjoining form is never attested in similar words and will be considered as a spelling mistake:
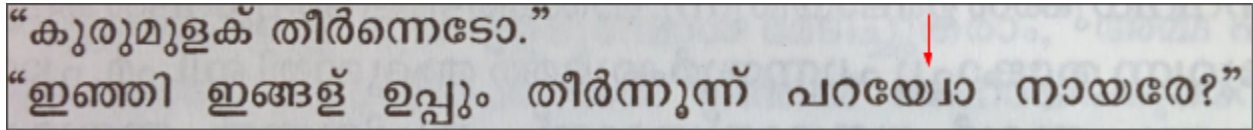
പോയ്വരൂ

The correct reformed orthography rendering is with the visible virama:

പോയ്വരൂ

The corresponding second C2-conjoining form has a special usage today. It is used to represent the vernacular as below:

പറയ്യോാ <PA RRA **Y-V**-OO> meaning 'would [you] say?'.

കുരുമുളക് തീർന്നെടോ."
"ഇഞ്ഞി ഇങ്ങള് ഉപ്പും തീർന്നുന്ന് പറയോാ നായരേ?"

Nīrmātaḷaṃ Pūttakālaṃ (1997), by Mādhavikkuṭṭi, page 56. Gives attestation to the sequence above.

Similar to the previous case, following alternate shaping with first C2-conjoining form is never attested and will be considered as a spelling mistake:

പറയ്യോാ

So there is a need to specifically indicate which form is requested in a word so that shaping engine can choose the correct C2-conjoining form.

Historic fractions and numbers

Historically, traditional conjuncts are used to represent fractions and numerals. For example, traditional orthography conjunct P-TA, represents 1/320 along with recently encoded Malayalam fractions:

Glyph	x/320
ഌ	1/320
ഝ	2/320
മ	4/320

Historic fractions[4]. Conjunct that was was not atomically encoded is marked

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

ന	ന്ന	ഗ്ള	ഷ്ക്ര	ഝ്ര	ഹാ	ഗ്ര	പ്ര	ദ്രെ
na	nna	nya	ṣkra	jhra	hā	gra	pra	dre
1	2	3	4	5	6	7	8	9

മ	ഥ	ല	പ്ത	ബ	ത്ര	രൂ	ച	ണ
ma	tha	la	pta	ba	tra	rū	cha	ṇa
10	20	30	40	50	60	70	80	90

Akṣarappaḷḷi system[5]. Traditional orthography conjuncts are marked

---

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

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

Consistency between spelling and rendering

The difference between the two Malayalam orthographies, comes with spelling and rendering changes. There are not only traditional and reformed spellings, but also traditional and reformed orthography renderings. Mixing traditional spelling with reformed rendering, or vice versa, is going to look bad. For example, consider the word /*saukaryapradamāṇŭ*/ in its traditional spelling:

സൌകയ്യപ്രദമാണ്

Traditional spelling <S-AU KA Dot-Reph Y-YA **P-RA** DA M-AA **NNA U-sign Virama**> is rendered with the traditional orthography font Meera[6]. Rendering mismatch with reformed orthography style is indicated by the bold clusters.

Above traditional orthography text is rendered in reformed orthography rendering below:

സൌകയ്യപ്രദമാണു്

Traditional spelling; mismatched rendering with the reformed orthography font Noto Malayalam[7]

The corresponding reformed orthography of the same above word:

സൗകര്യപ്രദമാണ്

Reformed spelling <S-**AU-length-mark** KA **R-YA** P-RA DA M-AA NNA **Virama**> rendered with the reformed orthography font Noto Malayalam. The difference in spelling is indicated by colors.

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 is unambiguously traditional, that font should 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 P-RA ligature, there is no spelling difference and hence the font is unable to choose right rendering for that ligature. Since the shaping implementations are unable to determine which presentation is intended, today a Malayalam font cannot avoid the above described mismatch.

Case study: an 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 encounter in reformed spellings. For the common traditional sequence of <U-SIGN, VIRAMA>, the presentation form they invented is as below.[8]

---

[6] http://download-mirror.savannah.gnu.org/releases/smc/fonts

[7] https://code.google.com/p/noto/

[8] 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 P-RA. So only change that could be imagined was, to introduce a new presentation form for the unambiguously traditional spellings; i.e., <U-SIGN, VIRAMA>. This would be reformed orthography friendly.
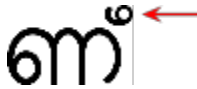
<NNA, U-SIGN, VIRAMA> rendering by Lohit-Malayalam[9]

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[10]. By far Malayalam is the only script in Indic to have such clear distinction made between orthographies.

With the spelling differences that happen between traditional and 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.

Current status

The resolution of PRI 37[11] established the overarching Indic conjoining behavior model with Virama and joiners. Indic conjoining model favors the full conjunct for <C1, VIRAMA, C2>. Half forms are produced by ZWJ, which acts like an invisible consonant letter that would always 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.

Corollary to PRI 37

PRI 37 has following properties which are relevant to the objective of this document:

1. PRI 37 does not specify which form to be used, if there are multiple C2-conjoining possibilities for some sequences.

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

---

[9] 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.
[10] http://www.microsoft.com/typography/otspec/languagetags.htm
[11] http://www.unicode.org/review/pr-37.pdf

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

Table 9. Rendering using control-picture glyph for ZWJ

Example from the accepted resolution of PRI 37; page 13

3. PRI 37 leaves the sequence <ZWNJ, VIRAMA> undefined. Probably, Malayalam can use this sequence to produce disconnected C2-conjoining forms, while <ZWJ, VIRAMA> producing connected C2-conjoining forms. This goes well with the general principle of ZWNJ; that is: "obstructs the normal ligature/cursive connection behavior"[12].

Proposal

The proposal below is enhanced PRI 37 mechanism in the above indicated manner. Only minimal changes are introduced while preserving 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 shaping 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.
- 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 form connected to C1 is selected.**
- [*Addional rule*] 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,

---

[12] http://www.unicode.org/versions/Unicode6.2.0/ch16.pdf Section 16.2 Page 551

this sequence requests the sub- or post-base form of C2 **that is disconnected from C1**, to be used with the full form of C1. If C2 has no **disconnected** sub- or post-base form, then fallback to the level 3 display is used.

- 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, ZWNJ, VIRAMA, C>** can be used to display the **respective** sub- or post-base form of C in isolation.

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

Sequence	S-KA	P-RA	Y-VA	Y-YA
Reformed rendering of <C1, VIRAMA, C2>	സ്ക	പ്ര	യ്വ	യ്യ
Reformed rendering of <C1, **ZWJ**, VIRAMA, C2>	സ്ക	പ്ര	യ്വ	യ്യ
Reformed rendering of <C1, **ZWNJ**, VIRAMA, C2>	സ്ക	പ്ര	യ്വ	യ്യ
Traditional rendering of <C1, VIRAMA, C2>	സ്ക	പ്ര	യ്വ	യ്യ
Traditional rendering of <C1, **ZWJ**, VIRAMA, C2>	സ്ക	പ്ര	യ്വ	യ്യ
Traditional rendering of <C1, **ZWNJ**, VIRAMA, C2>	സ്ക	പ്ര	യ്വ	യ്യ

Implication for current rendering technologies

The behavior of the two popular traditional orthography fonts Rachana and Meera with Harfbuzz[13], is to cursively connect C2-subjoining forms, irrespective of whether the joiner is ZWJ or ZWNJ.

പ്ര

Meera[14] traditional font rendering <PA, ZWJ/ZWNJ, VIRAMA, RA> with Harfbuzz

Uniscribe[15] also does not distinguish between ZWJ or ZWNJ. It produces disconnected C2-conjoining form that is not correctly reordered.

---

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

പ(

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

These tests indicate that there is no pre-existing, well established <ZWJ/ZWNJ, VIRAMA> usage in Malayalam. Also, Harfbuzz rendering of <ZWJ, VIRAMA> is in agreement with this proposal.

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 is important to users.

Moreover, some traditional conjuncts are used to represent numerals in once popular alternate number systems. Numerals needs to be represented in plain text.

Lastly, some words are required to use a specific C2-conjoining form. Examples:

പോയ്ക്രു
പറയ്യോാ

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 connected C2-conjoining form, the conjuncts used for a numeral can get split up in a reformed orthography rendering, obscuring its numeral semantics.

Also, the spelling differences like that between the following words needs to be represented in plain text:

പോയ്ക്രു
പറയ്യോാ

Moreover, MAL and MLR Language System tags are part of a specific rendering technology. Unicode specification cannot rely on such a specific technology.

3. 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.

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

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

5. In traditional rendering, many <Consonant, VIRAMA, RA> sequences modify the base consonant forming a deep ligature. Forming a deep ligature with <ZWJ, VIRAMA> is against PRI 37.

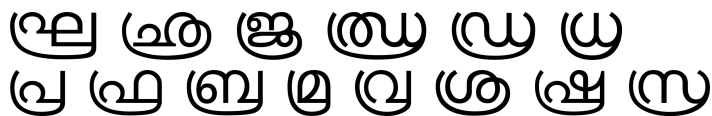All traditional renderings of the form <Consonant, VIRAMA, RA> that modify the base are the following:

ക്ര ഗ്ര ട്ര ത്ര ദ്ര ന്ര ഭ്ര ഹ്ര

Their corresponding base letters are:

ക ഗ ട ത ദ ന ഭ ഹ

We can see that the base is almost intact and cannot be really called deep conjunct like in the case of Devanagari K-SHA:

क + ष = क्ष

Remaining consonant letters not at all change the base on <VIRAMA, RA>:

ഘ്ര ഛ്ര ജ്ര ഡ്ര ഢ്ര ധ്ര
പ്ര ഫ്ര ബ്ര മ്ര യ്ര ശ്ര ഷ്ര സ്ര

Of course, C2-conjoining form of RA assume different shapes; however, they still do not modify the base consonant.

6. In Grantha and Telugu there are instances of various C2-conjoining forms are used interchangeably. Doesn't that mean the described differences be implemented in rich text?

This proposal does not require all Indic scripts to always choose a C2-conjoining from from multiple. It only provides a mechanism only if the writer wishes to do so; otherwise, it is upto the rendering system to choose the right form as it is today. The argument of this proposal is that in Malayalam script there is a need for differentiating between some C2-conjoining forms. That does not mean that, this need is universal to Indic. So the scripts like Grantha or Telugu which do not have this need, can safely ignore the proposed additional sequences.

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

See the answer to question 3. Also, unless an Indic script has multiple C2-conjoining forms, this proposal does not have any implications.

8. 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.

9. 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.

10. 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.