L2/06-334

Proposal on Handling Reph in Gurmukhi and Telugu Scripts

Nagarjuna Venna

August 1, 2006

1 Introduction

Chapter 9 of the Unicode standard [1] describes the representational model for encoding Indic scripts. Devanagari is described in Section 9.1; the principles of Indic scripts are covered in some detail in the introduction to Devanagari. The descriptions of the remaining Indic scripts were abbreviated highlighting any differences from Devanagari where appropriate. Some of the problems in this description were clarified by Public Review Issue #37 [2] which focused on consistent handling of Zero Width Joiner (ZWJ) in Indic scripts. That proposal put forth a set of rules for handling ZWJ and ZWNJ that are applicable across all Indic scripts.

The formation of Reph is defined in Section 9.1, Rules for Rendering, R2 of [1]. Reph is defined as a nonspacing combining mark glyph form of U+0930 `DEVANAGARI LETTER RA` positioned above or attached to the upper part of a base glyph form. Basically, Reph is formed when a RA which has the inherent vowel killed by the virama begins a syllable. Not all scripts have Reph; if the script in question has a Reph form, the sequence <RA, VIRAMA, C> is rendered with Reph on C. Also, for Devanagari, the sequence <RA, VIRAMA, ZWJ, ...> is always rendered as eyelash-RA instead of Reph.

Devanagari, Bengali, Gujarati, Oriya, and Kannada are listed in [1] and [2] as scripts that have a Reph form. Gurmukhi, Tamil, and Telugu are listed as scripts that do not have a Reph form. Malayalam is described as a script that has Reph in the traditional orthography but not in modern usage. However, both Telugu and Gurmukhi are similar to Malayalam in that Reph was used in ancient texts, but is not used in contemporary writings. While several scripts consistently use Reph (both in modern and historic usage), Gurmukhi, Malayalam, and Telugu have variable usage with respect to Reph and there are special scenarios where users may need to display Reph, typically for reproducing old documents. Unfortunately, there is no mechanism in the standard for users to indicate to a renderer that Reph should be displayed, if possible, in one of these scripts.

The intent of this proposal is to specify an encoding mechanism that allows users of Gurmukhi and Telugu to indicate that Reph should be displayed by

Figure 1: Examples of Gurmukhi Reph extracted from [5]

a rendering engine, if supported by the font that is being used. Malayalam, however, is not addressed in this proposal.

2 Reph

2.1 Gurmukhi

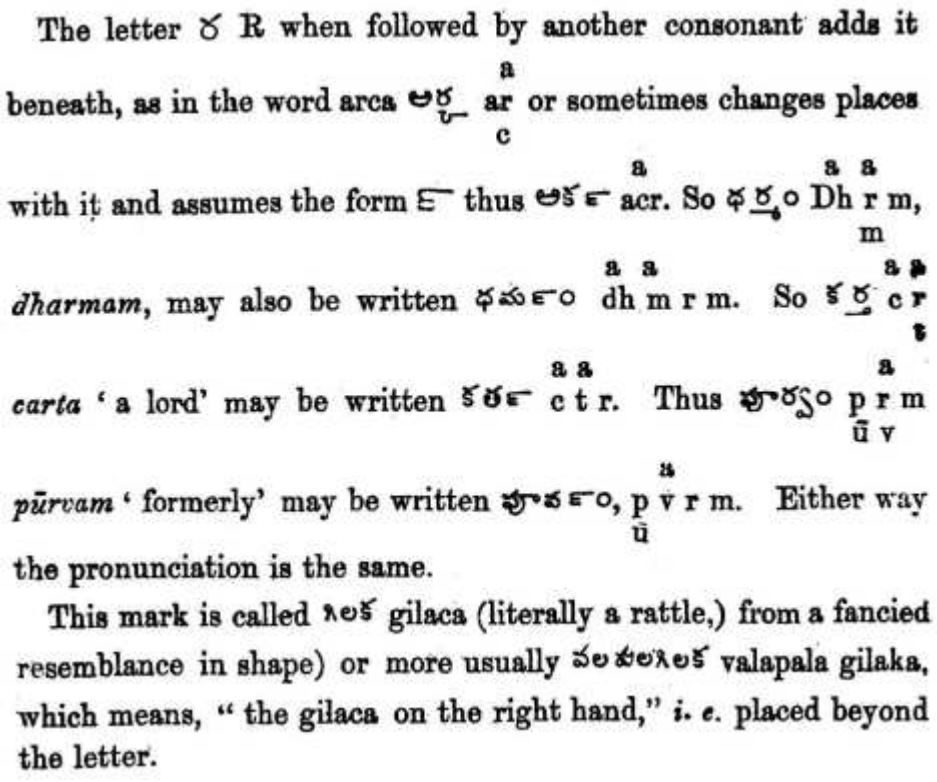
Reph in Gurmukhi is traditionally used when transliterating from Indic scripts that use complex consonant clusters beginning with RA. Examples of such usage can be found in Mahan Kosh [5], an encyclopedia on Sikhism. Reph is used there largely for Sanskrit terms.

Figure 1 shows examples of Reph extracted from [5].

2.2 Telugu

Reph in Telugu is known as *valapalagilaka*. The word *valavala* or *valapala* means the right side [4]. *valapalagilaka* is defined as the letter RA which is always

Figure 2: Description of Telugu Reph from [3]



written on the right side of the syllable before which it is pronounced [4]. This letter can be found in books published until about 150 years ago. Reph has pretty much disappeared from modern Telugu writing though some people still use it.

Figure 2 is a description of Reph with some examples from [3].

3 Rules for conjoining consonants

Conjoining of consonants in Indic scripts follows a three-level precedence heirarchy: a dead consonant C_d followed by a consonant C2 can be displayed in three ways [2]:

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

In general, the highest level available is used. In scripts that have Reph, the sequence <RA, VIRAMA, C2> takes precedence over other conjoining forms; it is always rendered with Reph on C2. The sequence <RA, ZWJ, VIRAMA, C> is rendered as full RA + sub/post-base form of C, if C is a C2-conjoining consonant. In Devanagari script, the sequence <RA, VIRAMA, ZWJ, C2> is rendered as half form of RA ("eyelash RA") + C2; no specific behavior is defined for other scripts.

4 Possible Solutions

Three possible plain text encoding solutions to the problem of indicating Reph in Gurmukhi and Telugu are considered here. Solution 1 specifies the use of sequence <RA, VIRAMA, C2> to render Reph in Gurmukhi and Telugu like other Indic scripts that have Reph; Solution 2 specifies the use of the sequence <RA, VIRAMA, ZWJ, C2> to render Reph; Solution 3 specfies the use of the sequence <RA, ZWJ, VIRAMA, C2> to render Reph. The advantages and drawbacks of each of the solutions are also discussed.

It should be noted that a font based solution is possible by specifying a modification to the rendering engine behavior. For Gurmukhi and Telugu, if a font does not form Reph in the 'rphf' feature, the rendering engine can be modified to not analyze the clusters as having Reph behavior; but if the font does form Reph in the 'rphf' feature the engine will analyze the clusters accordingly. As such, a font created for historic Telugu usage (that includes Reph) and a font created for modern usage (that does not include Reph) would simply work. A font that is designed for both modern and historic usage can potentially be accommodated using an OpenType LanguageSystem tag. While this solution may be adequate, it does not give a document writer the chance to indicate if Reph should be displayed by a rendering engine if possible. In addition, a font based approach can co-exist with an encoding solution as described in Appendix A.

The following description uses Telugu script as an example on occassion; the description applies to Gurmukhi as well unless noted otherwise.

4.1 Solution 1

The simplest solution is to add Gurmukhi and Telugu to the list of scripts that have Reph. This implies the sequence <RA, VIRAMA, C2> is always rendered with Reph and the sequence <RA, ZWJ, VIRAMA, C2> is rendered as full RA + sub/post-base form of C2. No specific behavior is defined for <RA, VIRAMA, ZWJ, C2>; the rendering follows precedence rules.

This solution however suffers from a major drawback: it changes the default rendering of existing Gurmukhi and Telugu text. Most users of modern Gurmukhi and Telugu do not recognize Reph and the most common form of writing conjuncts of RA will require using the sequence <RA, ZWJ, VIRAMA, C2> instead of <RA, VIRAMA, C2>.

4.2 Solution 2

In this solution, the sequence <RA, VIRAMA, C2> will render conjuncts of RA using the modern orthography (both in Gurmukhi and Telugu, full form of RA followed by sub/post-base form of C2); the sequence <RA, VIRAMA, ZWJ, C2> will render Reph. Finally, <RA, ZWJ, VIRAMA, C2> will continue to be rendered identical to <RA, VIRAMA, C2>.

The main advantage of this solution is that the sequence <RA, VIRAMA, ZWJ, C2> can be defined as Reph displaying for all scripts (except for Devanagari) that have some form of Reph. This doesn't change the existing behavior that <RA, VIRAMA, C2> displays Reph in scripts that use it in modern orthography. This solution goes with the pattern set in [2], a consistent mechanism for requesting Reph in all scripts.

4.3 Solution 3

In this solution, the sequence <RA, VIRAMA, C2> will render conjuncts of RA as in Solution 2; the sequence <RA, ZWJ, VIRAMA, C2> will render Reph. No specific behavior is defined for <RA, VIRAMA, ZWJ, C2> as in Solution 1.

The main appeal of this solution is that it is the mirror of the solution used in Kannada for the same problem. There, the sequence <RA, VIRAMA, C2> is rendered using Reph and <RA, ZWJ, VIRAMA, C2> is rendered without Reph. Following the precedence hierarchy specified in [2], in a font that features the complete set of sub/post-base forms of the consonants in Telugu (all vattu forms), the sequence <C1, ZWJ, VIRAMA, C2> and the sequence <C1, VIRAMA, C2> must always be rendered identically[1]. This solution will break that relationship for all sequences that start with RA. This is highly undesirable because it goes against the pattern set in [2] where the sequence <C1, ZWJ, VIRAMA, C2> is always used to get C2-conjoining irrespective of the script.

In addition, for a C2-conjoining consonant, the sequence <NBSP, ZWJ, VIRAMA, C> can be used to render the sub/post-base form of C in isolation. In Telugu, inserting RA in the place of NBSP in the above sequence will alter the rendered text substantially (C goes from being sub/post-base form to full form) compared to replacing NBSP with any other Telugu consonant. This is in contrast to replacing NBSP with RA in Kannada, the full form of RA will be rendered before C and C remains unaltered.

5 Proposal

In the interest of retaining backward compatibility with existing Gurmukhi and Telugu text, it is proposed that Solution 1 be not adopted. Solution 3 introduces a new pattern for Gurmukhi and Telugu in addition to changing the display drastically when NBSP is replaced by a consonant other than RA in the sequence

[1] With the exception of K.SSA, the only ligature in Telugu

<NBSP, ZWJ, VIRAMA, C>. In light of this, this proposal recommends that Solution 2 be adopted for handling Reph in Gurmukhi and Telugu scripts.

Adoption of Solution 2 requires a reclassification of Indic scripts for Reph handling purposes into three categories instead of the current two categories: scripts that primarily use Reph, scripts that do not have Reph and scripts in which Reph is derpecated. These three categories are named Regular Reph, No Reph, and Deprecated Reph respectively. Devanagari, Bengali, Gujarati, Oriya, and Kannada belong to Regular Reph category; Tamil belongs to No Reph category and Gurmukhi, Telugu belong to the Deprecated Reph category.

This also introduces a slight variation in previously stated Indic rendering rules. Earlier, Reph formation was given precedence over C2-conjoining; this would remain true for scripts that belong to Regular Reph category. For Deprecated Reph scripts, Reph formation wouldn't have precedence unless ZWJ is used. In addtion, the sequence <RA, VIRAMA, ZWNJ, C> is always displayed with an overt halant on RA and the sequence <RA, ZWJ, VIRAMA, C2> will display the full form of RA with conjoining form of C2 if that exists in the font. This part remains unchanged from [2].

Table 1 shows the relevant sequences and their rendering. This is an update to Table 12 in [2]. This table does not assume any changes to the rendering engine (for example, checking if a font forms Reph in the 'rphf' feature for all scripts). Please note that when a cell says 'Level n', it means that is the highest precedence level that can be applied to render that sequence. The precedence levels are specified in [2] and repeated in Section 3 of this proposal. A '-' indicates that there is no special processing associated with that sequence.

Table 1: Disambiguation of Sequences

Sequence	Regular Reph	No Reph	Deprecated Reph
<RA, VIRAMA, C2>	Reph on C2	-	-
<RA, VIRAMA, ZWJ, C2>	eyelash RA + C2 for Devanagari, else Reph on C2.	-	Reph on C2
<RA, ZWJ, VIRAMA, C2>	-	-	-
<C1, VIRAMA, C2>	Level 1	Level 1	Level 1
<C1, VIRAMA, ZWJ, C2>	Level 2	Level 2	Level 2
<C1, ZWJ, VIRAMA, C2>	Level 2	Level 2	Level 2

6 Acknowledgements

I'm grateful to Peter Constable for his careful review and thoughtful comments. I'm also thankful to Suresh Kolicala and Sukhjinder Sidhu for their feedback on the drafts.

References

[1] The Unicode Consortium. The Unicode Standard, Version 4.0.0, defined by: *The Unicode Standard, Version 4.0* (Boston, MA, Addison-Wesley, 2003. ISBN 0-321-18578-1)

[2] Constable, Peter. *Proposal on Clarification and Consolidation of the Function of ZERO WIDTH JOINER in Indic Script*, 30 June 2004, <`http://www.unicode.org/review/pr-37.pdf`> (1 August 2006)

[3] Brown, Charles Philip. 2006(1856). *The Grammar of the Telugu Language, 2nd Edition*. Chennai: Asian Educational Services. ISBN 82-206-0041-X.

[4] Brown, Charles Philip. 2004(1903). *Dictionary Telugu - English, 2nd Edition*. Revised by M, Vemkata Ratnam, Campbell, W H, Kandukuri, Viresalingam. Chennai: Asian Educational Services. ISBN 81-206-0037-1.

[5] Nahba, Kahan Singh. 1926. *Mahan Kosh*.

A Interaction with the font based solution

If a rendering engine chooses to support font based solutions by checking if the font forms Reph in 'rphf' feature for all scripts, Table 2 shows how the above proposal will co-exist with such a solution.

Table 2: Disambiguation for font based approach

Font Type	<RA, VIRAMA, C2>	<RA, VIRAMA, ZWJ, C2>
Modern Only	Full-RA + halant + C2	Full-Ra + halant + C2
Historic only	C2 + Reph	C2 + Reph
Dual Behavior	Full-RA + halant + C2	C2 + Reph