L2/12-164

Further editorial updates for Indic

Shriramana Sharma, jamadagni-at-gmail-dot-com, India

2012-May-02

§1. Devanagari

Some editorial updates which I had requested back in L2/10-471 §1 have still not been effected. I am hence once more summarizing them so that they may be quickly disposed of. The reasoning and evidence behind the request has been explained in my previous document and Anshuman Pandey's L2/09-320 before that. Therefore I only list the changes that need to be made. All these changes apply only to the names list.

1) DEVANAGARI LETTER characters 090E SHORT E, 0912 SHORT O, and the corresponding vowel sign characters 0946 and 094A are currently annotated as used for transcribing Dravidian short vowels. They however have native use in Kashmiri, and in the Bihari languages Bhojpuri, Magadhi and Maithili. They should first be annotated for their native use and then for the secondary transcription use. The recommended wording is: "used in Kashmiri and in Bihari languages"
2) 093A and 093B DEVANAGARI VOWEL SIGN characters OE and OOE are under the sub-heading "Dependent Vowel Signs for Kashmiri". They are used not only in Kashmiri but also in the Bihari languages. Therefore the sub-heading should be changed to just "Dependent Vowel Signs" and the following text should be added in italic below that: "used in Kashmiri and in Bihari languages".
3) 094F DEVANAGARI VOWEL SIGN AW has the annotation "Kashmiri". This should be replaced by "used in Kashmiri and in Bihari languages".
4) 0973 to 0977 are under the sub-heading "Independent Vowels for Kashmiri". That sub-heading applies only to 0976 and 0977 and should hence be moved to stand before 0976. The sub-heading for 0973, 0974 and 0975 should be just "Independent Vowel Signs" and the following text should be added in italic below that: "used in Kashmiri and in Bihari languages".

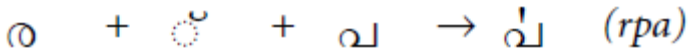
Now there might be some doubt as to what exactly "Bihari languages" refer to, but instead of repeating the specific names in all four places, text in italic can be added at the head of the block or below the first occurrence clarifying the specific languages referred to.

§2. Malayalam

In L2/12-106 I had requested updates to the Malayalam chapter description of the dot reph. I now request further updates for Malayalam based on issues raised over private mail by Santosh Thottingal. From my conversation with him it is clear that there is difficulty on part of the Malayalam community in understanding the encoding model used for some rare Malayalam written forms. It is hoped that the proposed text will address that difficulty.

§2.1. Inappropriate entry in Table 9-26

Santosh Thottingal pointed out in a draft document he sent me via private mail that TUS 6.0 p 308 (p 340 of PDF) lists in Table 9-26:



... and this is inappropriate because after the encoding of the DOT REPH character, RA + VIRAMA cannot represent the dot reph written form. **Therefore this entry should be removed from the table.**

§2.2. Corrections to description of samvruthokaram

Below the above table, TUS continues:

> When the candrakala sign is visibly shown in Malayalam, it usually indicates the suppression of the inherent vowel, but it sometimes indicates instead a reduced schwa sound [ə], often called “half-u” or samvruthokaram. In the later case, there can also be a -u vowel sign, and the base character can be a vowel letter.

It is better to avoid implying that the frequency of usage of the candrakala as the virama is greater than its usage for the samvruthokaram.

Further, feedback from native users (Santosh Thottingal) disputes the usage of the candrakala with independent vowel letters. Linguistically also, the Malayalam samvruthokaram, derived from the old Tamil kurriyalukaram, cannot occur at word-initial position (where independent vowels would be used). It is hence recommended to remove the reference to and example of independent letters.

It is hence requested to change the wording to:

When the candrakala sign is visibly shown in Malayalam, it may indicate the suppression of the inherent vowel, or it may indicate instead a reduced schwa sound [ə], often called "half-u" or samvruthokaram. In the later case, there can also be a -u vowel sign preceding the candrakala.

§2.3. Correction to table 9-27

Further, **the table 9-27 should be renamed** "Usage of candrakala for samvruthokaram" as it only gives examples of this particular usage of candrakala. Further, as we need to remove the example with independent vowel, **it is sufficient to provide a single example of a word using candrakala for samvruthokaram** with and without the vowel sign U:

പാല്, പാലു് /pālŭ/ milk

In passing, **the sentence "The anusvara can be seen after vowel letters"** following the above table **should also be corrected to read** "The anusvara can be seen multiple times after vowels". The word "multiple times" is needed to indicate the specific nature of the usage and it should also be recognized that it is not only limited to independent vowels.

§2.4. Description of special cases involving RRA

Below table 9-28 upto the end of table 9-30, many changes are required to the text dealing with special cases of RRA. Rather than specify each problem with the text (as there are many), I will merely provide the text to replace all content after table 9-28 upto and including table 9-30. As the replacement text is worded carefully with specific attention to diacritics and spelling, it is recommend to take it as it is:

Special cases involving RRA: There are some special written forms involving the MALAYALAM LETTER RRA ഠ which should be carefully handled. The relevant issues are briefly discussed here and the encoded sequences recommended for use are clearly stated.

The letter RRA ഠ is normally pronounced as a trill with the inherent vowel: /r̲a/. Two consecutive occurrences of the letter would then be naturally pronounced as two trills with two inherent vowels: ഠഠ /r̲ar̲a/. However, in older Malayalam orthography (until about 1960) the above written form was used to represent two alveolar stops with only one

inherent vowel: /tta/. In effect, ൿ functioned as a digraph. It also behaved as a single unit for any required vowel signs, with vowel marks to the left being written before the digraph ൿ and those to the right being written after the digraph ൿ.

In modern Malayalam orthography, this digraph has been replaced by the corresponding stack ൎ which is the only written form now used to represent /tta/. This is encoded as റ + ് + റ. As such, when the modern orthography is represented in Unicode, vowel signs can be used after the sequence റ + ് + റ and they would correctly position themselves with respect to it. For example: the word /māttoli/ “echo” is written as മാറ്റൊലി and encoded as മ + ാ + റ + ് + റ + ൊ + ല + ി.

However, in older orthography the same word would be written as: മാൺറൊലി. This however cannot be encoded using a single vowel sign coming after the digraph ൿ as such digraphs are not recognized in Indic Unicode. Unicode encodes written forms and not the sounds thereof. Thus words using the ൿ digraph should be encoded by ignoring the digraph. For example to encode the word മാൺറൊലി, one must apply the left vowel mark to the first RRA and the right vowel mark to the second: മ + ാ + റ + ൺ + റ + ൊ + ല + ി. This specifically affects the usage of reordrant and two-part vowel signs:

/tte/	ൺറ	റ + ൺ + റ
/ttē/	ൽറ	റ + ൽ + റ
/ttai/	ൻറ	റ + ൻ + റ
/tto/	ൺറൊ	റ + ൺ + റ + ൊ
/ttō/	ൽറൊ	റ + ൽ + റ + ൊ
/ttau/	ൺറൗ	റ + ൺ + റ + ൗ

Another special case involving RRA is when it combines with chillu-N ൻ. The sound /ṉḍa/ is written as ൯. Again, as Unicode encodes written forms and not their sounds, this written form must be encoded using chillu-N and the virama: ൻ + ് + റ. Thus the proper name ആന്റൊനി /āṉḍoni/ is encoded as: ആ + ൻ + ് + റ + ൊ + ന + ി. The virama character is needed even though ൻ is itself vowelless, as sometimes there is no stack of ൻ and a following റ: ഹെൻറി /henri/ which would be encoded just as ഹ + െ + ൻ + റ + ി.

-o-o-o-