L2/09-069

Malayalam: Adopting CLDR collation tailoring by DUCET

Cibu Johny cibu@google.com

2009-01-27

[This request is also filed in CLDR bug tracking as bug #2005[4]]

Malayalam is the only language that uses Malayalam script as its primary script[1, 2]. So it is might be prudent to incorporate the CLDR collation tailoring rules for Malayalam directly in DUCET. That would remove the need for using tailoring by software components to get dictionary order for Malayalam words.

Rules

The proposed rules and their rationale are described below. Please see the appendices for these rules in specific syntax.

Malayalam Visarga and Avagraha may not have any primary weight:

```
The first primary ignorable << U+0D03
U+0D03 << U+0D3D
```

The Malayalam AU length marker may differ from AU-sign only in tertiary weight and may be placed in between OO-sign and Samvruthokaram sign (U-Sign and Virama):

```
U+0D4B < U+0D57
U+0D57 <<< U+0D4C
U+0D57 < U+0D41 U+0D4D
```

Chillu encoding prevailing before 5.1 and their corresponding visible virama forms may differ only by secondary weights. Also, 5.1 Chillus and pre-5.1 Chillus may defer only by tertiary weights for existing data compatibility. Following 6 rules sets correspond to Chillus:

NNA

```
U+0D23 U+0D4D << U+0D23 U+0D4D U+200D
U+0D23 U+0D4D U+200D <<< U+0D7A
```

NA

```
U+0D28 U+0D4D <<< U+0D7B U+0D4D
```

/nta/ may be treated as <NA + VIRAMA + RRA>

```
U+0D28 U+0D4D << U+0D28 U+0D4D U+200D
U+0D28 U+0D4D U+200D <<< U+0D7B
```

RA

```
U+0D30 U+0D4D << U+0D30 U+0D4D U+200D
U+0D30 U+0D4D U+200D <<< U+0D7C
```

LA
```
U+0D32 U+0D4D << U+0D32 U+0D4D U+200D
U+0D32 U+0D4D U+200D <<< U+0D7D
```

LLA
```
U+0D33 U+0D4D << U+0D33 U+0D4D U+200D
U+0D33 U+0D4D U+200D <<< U+0D7E
```

KA
```
U+0D15 U+0D4D << U+0D15 U+0D4D U+200D
U+0D15 U+0D4D U+200D <<< U+0D7F
```

Anuswara may be considered as a MA_dead:
```
U+0D2E U+0D4D << U+0D02
```

The order of last few letters may be: HA, LLA, LLLA, RRA
```
U+0D39 < U+0D33
U+0D33 < U+0D34
U+0D34 < U+0D31
```

Note: This includes the fix for bug #2000[3] proposed.

Reference

1. Search in Ethnologue for Malayalam script.
2. http://en.wikipedia.org/wiki/Malayalam_script
3. http://www.unicode.org/cldr/bugs/locale-bugs/incoming?id=2000
4. http://www.unicode.org/cldr/bugs/locale-bugs/incoming?id=2005
5. Online version of this document

Appendix A: The CLDR version of the rules

Please refer to online version for characters[5]. PDF version might miss some conjuncts and newly added Unicode characters.

```
<rules>
    <reset><first_primary_ignorable/></reset>
    <s>ം</s>
    <reset>ം</reset>
    <s>?</s>

    <reset>ക്ാ</reset>
    <p>്ര</p>
    <reset>്ര</reset>
    <t>ന്ര</t>
    <reset>്ര</reset>
    <p>്ര്യ</p>


    <reset>ൻ്റ</reset>
```

```
    <s>ൺ് </s>
    <reset>ൺ് </reset>
    <t>?</t>


    <reset>ന് </reset>
    <t>?ൊ</t>
    <reset>ന് </reset>
    <s>ൻ് </s>
    <reset>ൻ് </reset>
    <t>? </t>

    <reset>ര് </reset>
    <s>ർ് </s>
    <reset>ർ് </reset>
    <t>? </t>

    <reset>ല് </reset>
    <s>ൽ്</s>
    <reset>ൽ് </reset>
    <t>?</t>

    <reset>ള് </reset>
    <s>ൾ്</s>
    <reset>ൾ്</reset>
    <t>?</t>

    <reset>ക് </reset>
    <s>ക്്</s>
    <reset>ക്് </reset>
    <t>?</t>

    <reset>മ് </reset>
    <s>ം</s>

    <reset>ഹ </reset>
    <p>ള</p>

    <reset>ള </reset>
    <p>ഴ</p>

    <reset>ഴ </reset>
    <p>റ</p>
  </rules>
```

Appendix B: The ICU version of the rules

```
&[first primary ignorable] << ഃ
&ഃ << ?

&ൊ < ൗ
```

&ൗ <<< ൊൗ
&ൗ < ൂ്

&ൺ് << ൺ്
&ൺ് <<< ?

&ന് <<< ?്
&ന് << ന്
&ന്് <<< ?

&ര് << ര്
&ര്് <<< ?

&ല് << ല്
&ല്് <<< ?

&ള് << ള്
&ള്് <<< ?

&ക് << ക്
&ക്് <<< ?

&മ് << ം

&ഹ<ള
&ള<ഴ &ഴ<റ

Experimentation
link:
http://demo.icu-project.org/icu-bin/locexp?_=ml&d_=en&x=col