L2/09-264

Date: July 29, 2009

To: Unicode Technical Committee

From: Unicode Editorial Committee (per meeting June 25) and CLDR (per meetings July 1, July 29)

Subject: Proposal for CLDR to specify preferred character sequences

Currently there are several places in which the Unicode standard or related documents specify preferred or well-formed character sequences for certain scripts. For example:

- In *The Unicode Standard 5.0,*
  - Section 11.3 Myanmar, p. 381, Ordering of Syllabic Components; includes Table 11-3, Myanmar Syllabic Structure.
  - Section 11.4 Khmer, p. 390, Ordering of Syllable Components; includes Figure 11-3, Examples of Syllabic Order in Khmer.
- Portions of Proposed Draft Unicode Technical Report on Indic Scripts in Unicode (in development).
- Perhaps some of the material from Unicode 5.2 on use of Malayalam Chillu Characters.

Post-Unicode-5.2, it would make sense to move the specifications for such sequences to CLDR, and to remove the specifications from the Unicode standard. CLDR would develop a standard format (using e.g. regular expressions) for representing such information consistently across scripts/languages.

Advantages:

- Could result in APIs (e.g. in ICU) for sequence validation (e.g. using regex) and testing for problems.
- We would then have CLDR testing of the sequences, and be able to test CLDR data.
- The specification in the Unicode standard is not machine readable, but in CLDR data it would be.
- Having the specifications in CLDR would make it easy for them to be language-based or language-specific, if needed.