Rebuttal of Background of Indic Segmentation

Submitter: **Richard Wordingham**
Date: **30 April 2017**

Introduction

Submission L2/17-094 contains several unreliable statements, and I feel obliged to make corrections or warning.

The definition offered for the Indic Orthographic syllable offered is erroneous and inadequate. The formulation offered is:

V[m] | {CH}C[v][m] | CH

where

V = independent vowel

m = anusvara, visarga, chandrabindu

C = consonant, or consonant + nukta

v = dependent vowel

H = halant / virama

Even for the above constituents and for the Devanagari script, a more general formulation is required, namely

V[m] | {CH}C{v}[m] | {CH}CH

The extension of the third alternative is obvious; even Sanskrit has a few words that end in two consonants. Even then, at the code point level, this ignores the fact that Microsoft has long acknowledged the sequence of repha and independent vowel.

Some vowels in Devanagari are expressed by sequences of two vowels, for example <U+094E DEVANAGARI VOWEL SIGN PRISHTHAMATRA E, U+093E DEVANAGARI VOWEL SIGN AA>, which conveys the same vowel sound as U+094B DEVANAGARI VOWEL SIGN O.

The expression above ignores the use of ZWJ and ZWNJ. These are required for the proper display of Hindi when a font might otherwise use conjuncts considered appropriate for Sanskrit but inappropriate for Hindi,

e.g. द्ग <U+0926 DEVANAGARI LETTER DA, U+094D DEVANAGARI SIGN VIRAMA, U+0917 DEVANAGARI LETTER GA> as opposed to द्‌ग <U+0926, U+094D, U+200D ZERO WIDTH JOINER, U+0917>.

There is also the issue that even in Devanagari, a virama does not always combine consonants into a single orthographic cluster. For example, according to 'Devanagari VIP Team Issues Report' (Unicode document reference L2/11-11370), a derived form from Nepali श्रीमान् should be written श्रीमान्को <U+0936 DEVANAGARI LETTER SHA, U+094D DEVANAGARI SIGN VIRAMA, U+0930 DEVANAGARI LETTER RA, U+0940 DEVANAGARI VOWEL SIGN II, U+092E DEVANAGARI LETTER MA, U+093E DEVANAGARI VOWEL SIGN AA, U+0928 DEVANAGARI LETTER NA, U+094D, U+200C ZERO WIDTH NON-JOINER, U+0915 DEVANAGARI LETTER KA, U+094B DEVANAGARI VOWEL SIGN O> and not श्रीमान्को <U+0936, U+094D, U+0930, U+0940, U+092E, U+093E, U+0928, U+094D, U+0915, U+094B>. As the explicit virama is chosen to preserve the shape of the base form, it would seem that the preferred form should be considered as having 4 orthographic syllables SH.RII, MAA, N, KO, and the dispreferred form as having 3 orthographic syllables SH.RII, MAA, N.KO.

It also appears, though confirmation from Tamils would be good to have, that U+0BCD TAMIL SIGN VIRAMA normally does not combine consonants into orthographic clusters.

L2/17-094 urges that UAX#29 take the expression for the orthographic syllable into account. On the contrary, I would urge that, if it does, it should instead modify the expression for grapheme clusters to yield a new concept roughly corresponding to the orthographic syllable. The primary question to be addressed is when two grapheme clusters lie in the same unit of this type. This may actually be a better topic for CLDR. The orthographic cluster may even be too large a sequence for some of the applications. For example, it has been suggested that hyphenation should not split an orthographic syllable, but I have seen non-Devanagari manuscripts which break lines between the consonant and a vowel written on the right, corresponding to an extended grapheme cluster boundary.