On the Indic_Syllabic_Category of vowel carriers

To: SAH, UTC
From: Robin Leroy
Date: 2024-08-27

Summary

The Indic_Syllabic_Category assignment of the vowel-carrier letter (often “A”, “O” in Thai and related scripts) in Indic scripts that otherwise have no independent vowel has been inconsistent; it is most often Consonant (Ahom, Buginese, Kharoshthi, Lepcha, Makasar, Marchen, Rejang, Tibetan; as well as Lao, Tai Viet, and Thai), but in two cases it is Vowel_Independent (Soyombo, Zanabazar square), and in one case Consonant_Placeholder (Limbu).

Proposal

1. Change the Indic_Syllabic_Category of U+11A50 SOYOMBO LETTER A and U+11A00 ZANABAZAR SQUARE LETTER A from Vowel_Independent to Consonant, and that of U+1900 LIMBU VOWEL-CARRIER LETTER from Consonant_Placeholder to Consonant.
2. Going forward, scripts should either have more than one character with InSC=Vowel_Independent, or should have none, with the vowel-carrier being classified as InSC=Consonant. An invariant test should enforce this.

Alternatives considered

Consonant_Placeholder seems inappropriate; none of the other characters this category function as ordinary vowels or consonants.

One major problem with giving these letters Indic_Syllabic_Category=Vowel is that they can often be analysed as /ʔa/ rather than /a/. The Unicode Technical Committee should not get into the business of deciding the unanswerable question of whether it is appropriate to analyse any of the languages written using a given script as having an initial /ʔ-/ phoneme. As an example of the kind of argument we should avoid, *The Unicode Standard*, Section 13.14, *sub* “Vowels”, states that “Initial vowels are represented by the neutral letter U+1C23 LEPCHA LETTER A”, whereas Richard Ishida’s orthography page for Lepcha transcribes it as /ʔ/. Conversely, Section 16.7, *sub* “Consonants” states that “U+1A4B TAI THAM[1] LETTER A represents a glottal consonant”, whereas Richard Ishida’s orthography page for Northen Thai states that “[t]here is an incomplete set of independent vowels, and standalone vowel sounds are typically written using vowel signs applied to ᩣ U+1A4B LETTER A”. Such questions of phonology are explicitly out of scope for Indic_Syllabic_Category; as is noted in IndicSyllabicCategory.txt, “rules for syllabic structure in Indic scripts may differ significantly from how phonological syllables are defined”.

In contrast, the goal of Indic_Syllabic_Category is to provide data that can assist in “[s]pecification of syllabic structure” and “[s]pecification of segmentation rules”; the latter being already used by *The Unicode Standard*, since Indic_Syllabic_Category is used in the derivation of Indic_Conjunct_Break, itself used in the definition of Extended Grapheme Clusters. For those purposes, treating the vowel carriers as consonants is helpful, as it

[1] Note that Tai Tham *does* have other independent vowels, so it is out of scope for the narrow principle proposed in this document.

simplifies the expression of the "typical" syllable[2] in those scripts: the dependent vowels are only expected to occur after consonants.

Finally, adding yet another InSC value for those would retain all the difficulties of /ʔ/, would add to the ever-growing maintenance burden on the UCD side, and, while it would technically allow for finer handling in syllable processing, it would in practice make all the regular expressions more complicated for little tangible benefit.

[2] There are of course cases of independent vowels that are expected to carry dependent vowels in scripts with multiple independent vowels, such as the Kokborok diphthongs. Nevertheless, considering the default to be that dependent vowels occur on consonants simplifies the general case. If supported, these exceptions need special handling anyway.