L2/22-132

On the encoding model of the Gurung Khema script

Eduardo Marín Silva

25/05/2022

Foreword. This is a response to the document L2/22-096 that proposes a new script for inclusion. In this document I give my feedback on the encoding model and propose some changes. A copy of this document has been sent to the email address of the author.

Decomposable Vowel Signs. The current model uses 11 code-points for the vowel signs with 7 of them having decompositions to other "non-composed" vowel signs. This at first glance make sense; it is not uncommon for vowel signs in Indic scripts to have decompositions in order to both aid in rendering and input, and maintain canonical equivalence. I however, disagree that this is beneficial for this particular script.

Other decomposable vowel signs in Indic scripts are typically made of parts that are both on the left and the right of the base, but Gurung Khema only has vowel signs that are above (and below, if one uses the 1995 orthography) the base, so the same concerns do not apply.

Furthermore, if one examines the glyphs for the vowel signs, we realize that two of them: 16120 ◌̈ VOWEL SIGN II and 16121 ◌̄ VOWEL SIGN U could also be composed; the first being composed of two instances of 1611F ◌̇ VOWEL SIGN I and the second composed of two instances of 1611E ◌̄ VOWEL SIGN AA. So if we are being consistent, these introduce 5 new valid decompositions to AI, O, and AU (1,1 and 3 respectively).

Furthermore, there is an apparent mistake in the decompositions of 16124 VOWEL SIGN EE and 16127 ◌̋ VOWEL SIGN OO, where 1611F ◌̇ VOWEL SIGN I and 1612D ◌̵ (or the "LENGTH MARK") seem to have switched places.

Adding these corrections, we end with the following table:

Vowel Signs	Proposed Decompositions	"Corrected" Decompositions
1611E AA	---	---
1611F I	---	---
16120 II	---	1611F 1611F (I-I)
16121 U	---	1611E 1611E (AA-AA)
16122 UU	1611E 1612D (AA-L)	1611E 1612D (AA-L)
16123 E	1611E 1611F (AA-I)	1611E 1611F (AA-I)
16124 EE	1611F 1612D (I-L)	1612D 1611F (L-I)
16125 AI	1611E 16120 (AA-II)	1611E 16120 (AA-II) 1611E 1611F 1611F (AA-I-I)
16126 O	16121 1611F (U-I)	16121 1611F (U-I) 1611E 1611E 1611F (AA-AA-I)
16127 OO	1611E 1611F 1612D (AA-I-L)	1611E 1612D 1611F (AA-L-I)
16128 AU	16121 16120 (U-II)	16121 16120 (U-II) 1611E 1611E 1611F 1611F (AA-AA-I-I) 16121 1611F 1611F (U-I-I) 1611E 1611E 16120 (AA-AA-II)
1612D "LENGTH MARK"	---	---

If we assume that the order of the vowels in EE and OO is merely an editorial mistake, we can still see that it is arbitrary to treat II and U as if they don't have decompositions. Making such an exclusion out of convenience is unnecessary compared to the alternative.

Furthermore, we can see that this model forces the separate encoding of a so-called “length mark” in order to have decompositions for EE and OO, but this is merely a graphical primitive that isn’t used in isolation (see section 4).

The decomposable model, would make more sense if the users wanted to represent arbitrary combinations of vowel signs, but this is clearly not the case.

Atomic vowel signs. Having no decompositions has many upsides, that I list now:

1. All vowel signs are treated equally, making implementation simpler.
2. No need to add the “length mark” making the encoding more closely aligned with the actual orthography, and again, making implementation simpler.
3. No need to make the arbitrary exclusions of decompositions for II and U in order to avoid the presence of multiple valid decompositions for other signs.
4. Allows support for glyph variants, that don’t look like the standard ones [see section 3.4], making the script more robust to changes in user preferences.

On the MEDIAL RA and LAIHOMA signs. On section 3.2, a few new signs are introduced, but two of them are proposed as unifiable with other characters; The LAIHOMA sign (used for nasalization) with 030C ◌̌ COMBINING CARON and the MEDIAL RA sign with U+032D ◌̭ COMBINING CIRCUMFLEX ACCENT BELOW, due to their visual similarity.

This unification makes sense at first glance, but once has to remember that this is an Indic script and disunifications of visually similar characters is common, due to complex rendering requirements and specific glyph variants not present in the generic characters.

If users in the future, wanted to change the shape of these characters, unifying them would make implementing those variants very problematic. Furthermore, 030C and 032D don’t have (indeed, **should not** have) Indic Syllabic Categories, making implementations that rely on them, more complex.

It is my opinion that disunifying these characters has more pros than cons and has undeniable precedent on the encoding of Indic scripts. This is also apparent on the decision to encode the THOLHOMA separately and not unify it with 0316 ◌̖ COMBINING GRAVE ACCENT BELOW, because the same rationale for disunification, applies to the signs in question.

On the 1995 orthography. On section 3.1, the oldest version of the script is discussed (the 1995 version). The most salient difference is the use of vowel signs that go below the base (or both above or below). This was changed in the year 2000, where all vowel signs were made to appear above the base (see section 3.3).

We can assume that in that 5 year period, there would be a good number of important documents using the old model; therefore there would be an interest in digitization. Users could merely replace the old forms with the new ones, but that would corrupt the original appearance of the text, defeating the point somewhat. The old forms cannot be treated as variants of the new, due to the different positioning.

I propose to include an extra vowel sign: ◌ GURUNG KHEMA VOWEL SIGN OLD U, to support this old orthography. This sign could then be used with other signs to complete the set of the old forms. Proposing it at the same time as the entire script, is better overall as compared to proposing it later.

On the naming of the letters. I have noticed that the naming of the letters goes like “GURUNG KHEMA [CONSONANT/VOWEL] [*sound*]”, but this is unlike other Indic scripts that call independent vowels and consonants as “LETTER”. Although not really necessary, following the scheme of “GURUNG KHEMA LETTER [*sound*]” would fit better in Unicode.

In summary. I have argued against vowel signs with decompositions, in favor of atomic signs and removal of the unnatural addition that is the “LENGHT MARK”. I have also argued for the disunification of the MEDIAL RA and LAIHOMA, as well as the addition of an extra vowel sign (OLD U) to support the 1995 orthography.

###