L2/07-056

Comments on Viet Tay Proposal (L2/07-039)

Peter Constable

2007-02-01

Comments about the document

These don't particularly pertain to the technical merits of the encoding proposal:

1) Given the context of a Brahmi-derived script, the discussion about syllable structure can be misleading: a reader might start thinking in terms of what clustering behaviour is, but in fact the proposal is for something that would be comparable to Latin in terms of cluster behaviour: the initial consonants, medial /w/, spacing vowels, final consonants, and spacing tones are all letters. The discussion of syllables is still useful, though, in relation to line breaking (I'm assuming that, even if modern users have adopted word spaces, there will be people who want to set documents without word spacing, e.g. to reflect traditional style or in archival of historic texts.)

2) In discussing vowels, I would have characterized O LOW as a consonant that does double duty as a vowel. (At least, I'm assuming that these language communities do consider these to be one or the other, and I'm assuming they'd do this the same as is done for Thai, Lao and Lanna. The fact that it has a tone association and pairs with O HIGH seems to leave no option about this.) It's something to be aware of if you're trying to recognize syllables.

3) The description of digraph vowels is a tiny bit confusing in that, in showing character sequences, it isn't clear where the consonant goes. The case of /a:w/ suggests that dotted circle in all of these represents the consonant, but the first two seem to have only two components in the sequence – the two vowel components.

4) I wouldn't have described letters like AY, AN, etc. or even the combination for /-ap/ as ligatures. Except for the combination used for /-ap/ the others are letters; they just happen to stand for vowel-consonant rhymes rather than vowel rhymes. Even the representation of /-ap/ isn't a ligature; it's a letter-mark combination used to write a rhyme. As a result, the discussion of whether "ligatures" need to be encoded is completely unnecessary in relation to these. For instance, even the suggestion that AW or AY might be encoded as sequences completely ignores the historical derivation: these text elements are atomic entities that derived from atomic entities in earlier forms of writing. It would make as much sense to ask whether "x" should be encoded as a sequence of other characters.

The ligature question is still worth addressing wrt symbols KON, NWNG, etc., however. These things are potentially analyzable. (I think the conclusion is the right one, though.)

5) The statement is made that Tai scripts do not have an established standard for sorting. While that may be strictly true, there are some principles that apply to every other Tai script I've looked at. Specifically, sort order of consonants is based on three attributes (which may be measured in terms of an earlier form of the language different from the modern language):

- Point of articulation, starting with velar and moving forward. (This is carried over from practice in Indic scripts from which these derived.)

- Manner of articulation. (This is also carried over from practice in Indic scripts from which these derived.)
- Tone class.

(The tone classes are historically related to the manners of articulation: four kinds of articulation split into two groups, leaving two manners x two tones. That's simplifying, but it's a fair generalization for the 30-second summary. This applies to obstruents; the short story on nasals and other sonorants is that you end up with two tone classes of them as well.)

Now, there may not be a universal principle on how those are prioritized. I recall when New Tai Lue was being discussed for encoding, it turned out that different sources prioritized these three factors in different ways. One can always analyze the ordering in terms of these three factors, and I believe the same ordering within each factor has always been used in cases I've looked at.

6) There's something important in relation to sorting that is there if you're paying attention but I think should be called out explicitly: in applying weights to vowel signs, their weight within a syllable needs to be applied according to their *logical* order, and that applies the same for all vowels regardless of their visual order. Since encoding is in visual order and not logical order, sorting requires preprocessing to re-order vowels within the syllable (either that or treating each syllable as an atomic whole when assigning a weight).

Comments on the encoding proposal

Given my comments above regarding sorting of Tai scripts, I wonder if the consonants aren't coded in an inconsistent order: in the velar set, voiceless unaspirated stops come first, but in dental and bilabial sets, voiced stops come first.

My biggest concern has to do with the canonical combining classes of combining marks: everything is assigned to class 0, even though there are only above and below marks. Having all the marks be class 0 means additional guidance will be needed for implementers to deal with questions of equivalence since normalization isn't being used at all. For instance, these sequences would all look the same but be non-equivalent:

< KO LOW, U, MAI KANG, MAI EK >
< KO LOW, MAI KANG, U, MAI EK >
< KO LOW, MAI KANG, MAI EK, U >

If we simply assigned the below mark to class 220 and the above marks to class 230, equivalences would be completely taken care of, just as they are for Latin.

Other than those two issues, I haven't noticed anything else that concerns me with the encoding proposal. I expect someone will comment and perhaps even object to the fact that logical order isn't used. I'm not overly concerned about that – after all, I was the one a few years back suggesting that that should be done for New Tai Lue. I still think that would have been a better choice for New Tai Lue, and I think it's a reasonable choice here. In spite of the slight complication added to sorting, it saves a greater amount of complexity when it comes to rendering and user editing experience.