L2/14-033

Comments on the Zanabazar Square proposal

Shriramana Sharma, jamadagni-at-gmail-dot-com, India

2014-Jan-26

This is w.r.t. Anshuman Pandey's proposal L2/14-024 N45__ for Zanabazar Square.

Encoded representation of Mongolian diphthongs

Anshuman has documented the use of the vowel signs for *ai* and *au* as secondary vowel markers for diphthongs. Conceptually, they are used "in addition to" the vowel markers for the basic vowel signs since obviously the vowel they denote comes in second place. Thus in encoding all these diphthongs, Anshuman suggests the placement of the secondary vowel marker after the representation of the primary in encoded text, whether that be as a single character for short vowels or two for long vowels including the length mark. For instance:

oi <LETTER A, VOWEL SIGN O, VOWEL SIGN AI>

ōi <LETTER A, VOWEL SIGN O, VOWEL LENGTH MARK, VOWEL SIGN AI>

ou <LETTER A, VOWEL SIGN O, VOWEL SIGN AU>

In this connection this passage from TUS 6.2 ch 9.3 on p 301 (331 of PDF) should be noted:

> *More generally, when a consonant or independent vowel is modified by multiple vowel signs, the sequence of the vowel signs in the underlying representation of the text should be: left, top, bottom, right.*

This was mentioned in connection with use of dual vowel signs in Gurmukhi but since Zanabar Square is also an Indic script with vowel signs carrying CCC=0 and having Indic syllabic and matra categories, I understand that the above rule applies here too.

I understand that this guideline is due to the CCC=0 preventing reordering during normalization whereby VOWEL SIGN O + VOWEL SIGN AU is *not* canonically equivalent to VOWEL SIGN AU + VOWEL SIGN O though both may be graphically equivalent. However, I would think that the guideline is specifically needed only when two vowel signs are attached above and below the base, since in an LTR script a user is unlikely to not know the correct order of a vowel sign that is attached to the left or right of the base.

Order of Vowel Sign AU in encoding

Now the specific concern now is w.r.t. the only diphthong with *u* as secondary vowel seems viz. *ou*. The parallel diphthong *oi* with *i* as secondary is encoded as VOWEL SIGN O + VOWEL SIGN AI and likewise it is suggested to encode *ou* as VOWEL SIGN O + VOWEL SIGN AU. Linguistically this is fine, but the above rule suggests that the vowel sign placed on the left viz VOWEL SIGN AU should occur first in encoded text. This would mean that the sequence for *ou* should be VOWEL SIGN AU + VOWEL SIGN O even though this is linguistically wrong.

Now the UTC should decide whether it is important to follow the above rule and recommend that VOWEL SIGN AU + VOWEL SIGN O going against linguistic sensibility. Whatever is decided should be documented clearly in the relevant chapter of a future version of the standard to aid proper understanding and usage of the encoding.

Usage of Length Mark

An additional concern would arise in the case of diphthongs where the first component is long as seen in the case of *ōi* (see prev page). The sequence recommended is: FIRST VOWEL SIGN + LENGTH MARK + VOWEL SIGN AI. (No diphthongs of long vowels with a secondary *u* seem to be attested.)

Now the LENGTH MARK attaches to the bottom right of the base and the VOWEL SIGN AI attached to the top right, but the the IndicMatraCategory is given as “right” for both on p 24 of the proposal. Indeed, IndicMatraCategory.txt does not seem to distinguish these diagonal directions from the main ones (for good reasons no doubt), and it is not clear which direction such a diagonally positioned mark should be considered to stand on. However, between a top right and bottom right mark one might consider the latter to fall under the “bottom” category. In which case, the above guideline would mean that the LENGTH MARK should be positioned after the VOWEL SIGN AI, once more against linguistics.

Of course, it is not clear that these are IndicMatraCategory.txt actually defines the directions that are to be used in the guideline above, and the guideline may simply be taken to be that, and not a hard rule. In which case, the UTC may simply recommend that these are the recommended sequences for these diphthongs, that is, in case of diphthongs, the encoding should follow linguistic and not visual order.

Reversed Consonants

In the previous version of this proposal L2/13-198 N4471, Anshuman had proposed three consonants to be encoded separately:

	REVERSED DA		REVERSED NA	𑨫	REVERSED SHA

These represent the retroflex sounds of voiceless stop, (voiced) nasal and voiceless fricative respectively. As the names suggest, these are the laterally reversed forms of 𑨙, 𑨝 and 𑨪 which were labeled DA, NA and SHA.

In the current proposal, 𑨙 DA is renamed to TA as per Sanskrit/Tibetan usage. Its reverse form however has been removed. So has REVERSED NA . However, REVERSED SHA 𑨫 is proposed but renamed to SSA.

The logic behind removing these letters seems to be that there are other "legitimate" letters for denoting the retroflex stop and nasal i.e. 𑨔 and 𑨘 respectively, and hence the forms and should be considered "glyphic variants". However, just because the linguistic value is the same does not mean that the alternate representation is a glyphic variant. The forms and are derived by reversing the letters for the dentals TA and NA viz 𑨙, 𑨝, and do not bear any orthographic relationship to 𑨔 and 𑨘. As such, they cannot be considered glyphic variants of the latter (or of the former, since they contrast in usage).

If the reversed forms of TA and NA i.e. and are removed, why not reversed SHA i.e. 𑨫 also, given that that is also not a part of the original script created by Zanabazar just these other two (as noted in p 9 of the proposal)? The only reason seems to be the requirement of a letter to represent the fricative. However, encoding is not done to provide a letter for each sound but to provide a codepoint for each written character; and characters should be identified based on orthographic identity and not linguistic identity.

Anshuman dismisses and as "scribal idiosyncrasies". Why are these thus dismissable, but 𑨫 not? Whether it is a Lama who reversed the letter or somebody else, there exist attested documents using such these forms distinctively. Thus there is a requirement for encoding all three to digitally represent those letter forms.

As such, I submit that all three reversed forms should be encoded as:

	REVERSED TA		REVERSED NA	𑨫	REVERSED SHA

-o-o-o-