JTC1/SC2/WG2 N3489
L2/08-274
2008-08-04

Universal Multiple-Octet Coded Character Set
International Organization for Standardization
Organisation Internationale de Normalisation
Международная организация по стандартизации

Doc Type: Working Group Document
Title: On the inputting model for Batak
Source: Michael Everson and Uli Kozok
Status: Individual Contribution
Action: For consideration by JTC1/SC2/WG2 and UTC
Date: 2008-08-04

N3320R lays out the rendering rules for Batak reordering as follows:

> The main peculiarity of Batak rendering has to do with the way vowel glyphs are re-ordered when the killer (PANGOLAT or PANONGONAN) is used to close the syllable by killing the inherent vowel of a final consonant. This re-ordering is entirely regular and there are no exceptions to it.
>
> | | | | | | | | | | | | |
> |---|---|---|---|---|---|---|---|---|---|---|---|
> | ᯖᯇ᯲ | tap | = | ᯖ | ta | | | | + ᯇ | pa | + ◌᯲ | PANGOLAT |
> | ᯖᯇᯧ᯲ | tĕp | = | ᯖ | ta | + | ◌ᯧ | -ĕ | + ᯇ | pa | + ◌᯲ | PANGOLAT |
> | ᯖᯩᯇ᯲ | tep | = | ᯖ | ta | + | ◌ᯩ | -e | + ᯇ | pa | + ◌᯲ | PANGOLAT |
> | ᯖᯇᯪ᯲ | tip | = | ᯖ | ta | + | ◌ᯪ | -i | + ᯇ | pa | + ◌᯲ | PANGOLAT |
> | ᯖᯇᯬ᯲ | top | = | ᯖ | ta | + | ◌ᯬ | -o | + ᯇ | pa | + ◌᯲ | PANGOLAT |
> | ᯖᯇᯮ᯲ | tup | = | ᯖ | ta | + | ◌ᯮ | -u | + ᯇ | pa | + ◌᯲ | PANGOLAT |
>
> So although the backing store for *tip* is TA + I + PA + PANGOLAT, the display is not *ᯖᯪᯇ᯲ (which cannot occur) but rather ᯖᯇᯪ᯲. One way a font might implement this would be with a set of triplets, *Vowel + Consonant + Killer = glyph-CVK*. In the event that a visual order were entered in the text stream, an error state could be indicated with the retention of the dotted circle, thus:
>
> | | | | | | | | | |
> |---|---|---|---|---|---|---|---|---|
> | ᯖᯇᯪ᯲ | tip | = | ᯖ | ta | + ◌ᯪ -i | + ᯇ pa | + ◌᯲ | PANGOLAT (correct) |
> | ᯖᯇᯪ◌᯲ | tapiK | = | ᯖ | ta | + ᯇ pa | + ◌ᯪ -i | + ◌᯲ | PANGOLAT (incorrect) |
>
> Another way of putting this is to say that the PANGOLAT cannot follow a VOWEL SIGN, but only a LETTER.
>
> This regular re-ordering poses no new architectural challenges for the Brahmic model; indeed glyph reordering in complex syllables in Tai Tham is far more complex. There are moreover a number of reasons for preferring logical order for Batak. Both open and closed syllables are very frequent in the languages which use Batak: ᯇᯒᯬ᯲ᯂᯘᯪ᯲ *por-kis*, ᯔᯉᯬᯝᯘᯬ᯲ᯂᯉᯬ᯲ *ma-no-ngos-kon*, ᯔᯉ᯲ᯑᯇᯖᯬ᯲ᯂᯉᯬ᯲ *man-da-pot-kon*, ᯔᯒᯬ᯲ᯂᯒᯬ᯲ᯐ *mor-kor-ja*, ᯖᯒᯇᯮ᯲ᯂᯮ *ta-rup-ku*. Phonetic syllable structure is easier to process, to sort, to search, if logical ordering is used, because these cannot be mis-identified as ᯇᯒᯬᯂᯘᯪ *paro\kasi\*, ᯔᯉᯬᯝᯘᯬᯂᯉᯬ *manongaso\kano\*, ᯔᯉᯑᯇᯖᯬᯂᯉᯬ *mand\dapato\kano\*, ᯔᯒᯬᯂᯒᯬᯐ *maro\karo\ja*, ᯖᯒᯇᯮᯂᯮ *tarapu\ku*—all of which have valid syllable structures. Moreover, like other languages of Indonesia, most speakers are literate in Bahasa Indonesian, and their experience with computing is with that language, which has an extremely phonetic orthography. Their expectation will be to input their language by sound. Similar discussion held with users of the Balinese and Javanese scripts likewise indicated that phonetic input was their expectation. Visual order in the UCS is used with Thai and Lao for reasons of legacy, and with Tai Tham because of its similarity to Thai. All other Brahmic scripts in the UCS use logical order, and Batak need be no exception.

Members of the Script Subcommittee requested an investigation as to whether the user community really wanted to input in phonetic order or if they wanted to input in visual order. Michael Everson reported this to Uli Kozok in a GTalk chat session, excerpted below. The conversation was informal.

Everson: The UTC wants me to "prove" that logical ordering as opposed to visual ordering is better for the script. Obviously for linguistic purposes logical ordering is better. So one writes **ba+na+** (\ for **virama**) for *ban*, **ba+e+na+** for *ben*… Sorting is supported this way and any kind of linguistic lookup. The alternative is to write **ba+na+** for *ben*, and **ba+na+e+** for *ben*, but still **ba+e+na+e** for *bene*. *Ben* and *bene* would be written differently, namely **BNE** vs **BENE**. I tried to explain that for the other scripts of Indonesia it was certainly the case that because of Bahasa Indonesia people expected logical input in general. This is certainly true of Javanese and Balinese. Logical = phonetic. So they want me to make a Unicode implementation to "prove" that people can deal with phonetic input.

Kozok: Admittingly this is a challenge, but there s not much we can do except of designing a complicated algorithm that automatically converts phonetic input into the correct form. Actually we have done this already, but this is as an Internet based option only. I don't have the money to pay the guy who implemented the online version in designing a Windows based program.…

Everson: Phonetic input is easy. Getting the font to do the right thing with phonetic input is easy. What is not easy is visual order input. (Well, visual order input is also easy with a dumb font. But then your data is in bad shape.) So I favour phonetic input. **ba-e-na-** = *ben*, **pa-i-na-** = *pin*. Any Batak who also types Indonesian will be familiar with that: **b-e-n**, **p-i-n**. Some UTC people wanted to know if Bataks would prefer **b-n-e-** and **p-n-i-**.

Kozok: Well, all Batak that know the script know that *ben* has to be written **ba+ne+e+** so we shouldn't tell them that they are wrong in the some 1000 Batak manuscripts are equally wrong. Of course typing it is a bit of a nuisance and a little computer program implementing a smart algorithm that automatically converts typed **b+e+n** into **ba+ne+e+** would be desirable. Together with some friends in Medan we're thinking of getting some funds to implement such a program, but so far not much progress has been made.

Everson: Yes, we know that *ben* has to be *displayed* as **ba+ne+e+**. We will of course support that. It is not difficult to do that. So if they can type **b+e+n** and if the font does the right thing, that is what they want? A dumb font would force them to type **b+n+e+** which would play havoc with data. You don't need a smart program to turn **b+e+n** into the right visual display. You just need to have the right tables and glyphs in your font. You tell it that when a string **Vowel + Consonant + Killer** occurs, display a glyph which is CVK (not VCK). I could implement that in a font in about a day. I assume they want a QWERTY-based keyboard layout. We could do a project to implement such a font (multi-platform) and keyboard software (multi-platform (Mac, Linux, Windows)

Kozok: Yes, if it were possible to type **b+e+n** and the output would be **b+n+e+** that would clearly be favored. As I said I have worked with a German guy on this and it was quite complicated. I am trying to find the website and my communication with him. It's only then that I can tell you where we got stuck.…

It is clear from this discussion that the user community does in fact already prefer logical input to visual input, since they are looking for ways to implement that already. Accordingly we do not believe that it is necessary to test the user preference—it is already known. However, an exciting opportunity presents itself:

Kozok: My font set has now been officially adopted by a number of organisations and is used in schools all around the Batak lands to teach the Batak script.

Michael is prepared to make a font implementation of Uli's fonts, because to do so is straightforward—all we need is for Batak to get assigned to a ballot and this implementation can begin.