Layman’s comments on the encoding proposal Khitan small script

To Unicode Consortium

I am an amateur in linguistics and Unicode from China. I read the encoding proposal (link) of the Khitan small script into Unicode (henceforth referred to as the document), co-authored by Andrew West, Viacheslav Zaytsev, Michael Everson, dated 2016-05-21. I wish to share my thoughts on that proposal.

This article was first sent to Andrew West, and later sent to the Unicode Consortium following West’s advice.

For the ease of typing and rendering, I write phonogram blocks as [XX/XX], [X/XX] etc, and logograms as {X}, {Y} etc. The document considered three font types: vertical, horizontal, linear. For the sake of clarity, I only consider vertical representation, which is used by the actual Khitan scripts.

Shape adjustment

The document proposes to encode the individual phonogram components (yuanzi 原字) instead of encoding each phonogram block. I support this encoding scheme, but we should be very careful about how the font actually reproduces the blocks.

In all transcriptions of Khitan small script by Sinologists in computer typesetting, I found that they simply put the yunazi together without adjusting their shapes. This is not the correct way to write the script. The Khitan small script is more like Chinese hanzi instead of Mongolian, Phags-pa, in which each letter has a fixed shape. When Khitan phonograms are composed together to form a block, their shapes undergo a significant adjustment: they are narrowed and/or shortened.

Consider a phonogram block [X/YZ]. When typeset in Menksoft font, the top yuanzi X is placed in the center and the yuanzi Y, Z are aligned horizontally below X, creating an upward-pointing triangular shape. But in real Khitan script, the shape of all three yuanzi are adjusted so the block is always rectangular, not triangular. The top yuanzi X is in normal width and is shortened in height (compressed), while the bottom yuanzi Y, Z are always narrowed to half width and are usually also shortened. If the top X is 一, then the result of font stacking is usually egregiously incorrect and is aesthetically unacceptable, as the following figure (the document, p.10) shows:

The correct character is in square, while the computer representation is 1:2 rectangular.
A satisfactory font must take into consideration the adjustment scheme for the yuanzi, depending where they are placed, and the number of stacks in a block. I do not know if such a feature can be achieved by a font, or whether it should be part of the Unicode.

In the current encoding model, the space between two stacks in a block and the space between two logograms are the same, which could create a serious ambiguity, since we cannot distinguish a block \([X/Y]\) from two consecutive logograms \([X][Y]\) in computer typesetting. Though in real script, the two are clearly distinguishable because \([X/Y]\) is crammed together, while \([X][Y]\) are not. Comparison: In Chinese, 女子 (woman, female) [two “logograms”] is completely different from 好 (good) [a “phonogram block”].

Another issue and challenge is the inter-stack kerning. When a phonogram block is written, one cannot draw a horizontal line separating the two adjacent stacks. In the left block below, the vertical stroke 丨 in the bottom 木 protrudes the top stack. Same for the stroke 丿 in the bottom 火 in the middle block.

This phenomena is common in Chinese hanzi. For example, look at the character 森 (magnified above on the right); the top 木 and the bottom 林 should not be separated into two disjoint rectangular boxes. The same applies to the two 木 in 林. But since Unicode encodes individual CJK characters, the “inter-stack kerning” issue is not present in CJK.

**Inter-character spacing**

The document claims there is a white space between two phonograms, and between a phonogram block and a logogram. I find it difficult to agree. The Khitan script was modeled on Chinese hanzi, and blocks are formed by conjoining the yuanzi together. There is a natural gap between two blocks, but that does not mean a white space, like in English or in Mongolian. If we look at traditional Chinese scripts, especially calligraphy works (Figures 1, 2) and tablet carvings (Figures 3, 4), we see gaps that are identical to Khitan script gaps. Given the size of individual hanzi and Khitan small script, this gap is readily negligible.

The upright writing (usually used in tablet carvings) is square, and the calligraphic hand is rectangular, both comparable with Khitan small script. The gap between characters is caused by jumping from one rectangle (or square) to the adjacent rectangle (square), and is intended to separate characters to make it clear for reading.

The gap-keeping is generally ignored by cursive calligraphy (草书). In particular, in the wild cursive hand (狂草), the ink brush is intended not to be lifted from the paper as much as possible, which often results in many consecutive characters being “connected”. (Figures 5, 6, 7) This further shows that there is no white space between characters. Modern computer typesetting does not insert a
space between Chinese characters, or between two words. The same should apply to the Khitan
small script.

Figures 1, 2
Figures 3, 4
Figures 5, 6, 7
Comments on variants

The document (p. 110) proposes to encode separately the erroneous forms of some yuanzi, identified by Jiruhe & Wu 2009, given in the following list:

- 77 ereotype  = 78 ereotype
- 80  stereotype = 342  stereotype
- 406  stereotype = 149  stereotype
- 460  stereotype = 223  stereotype
- 320  stereotype = 321  stereotype
- 468  stereotype = 467  stereotype
- 154  stereotype = 155  stereotype
- 384  stereotype = 167  stereotype

If there is more or less a consensus between scholars that these are indeed variants in Khitan script, then they should be encoded as variants in Unicode, which can be selected by a variant selector, just like the way CJK ideograms are encoded. (See below for example) The underlying philosophy is the following. The academia is neither the creator nor the user of this ancient script, thus we should not add erroneous characters to the repertoire to remedy the mistakes made by the academia a thousand years later.

U+9AA8: 骨 (Mainland China) 骨 (Japan)

(Prepared, written by Snow Listener, 2016-11-17)