

## 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 —, then the result of font stacking is usually egregiously incorrect and is aesthetically unacceptable, as the following figure (the document, p.10) shows:

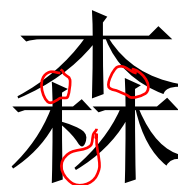


The correct character is in square, while the computer representation is 1:2 rectangular.

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 | 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.

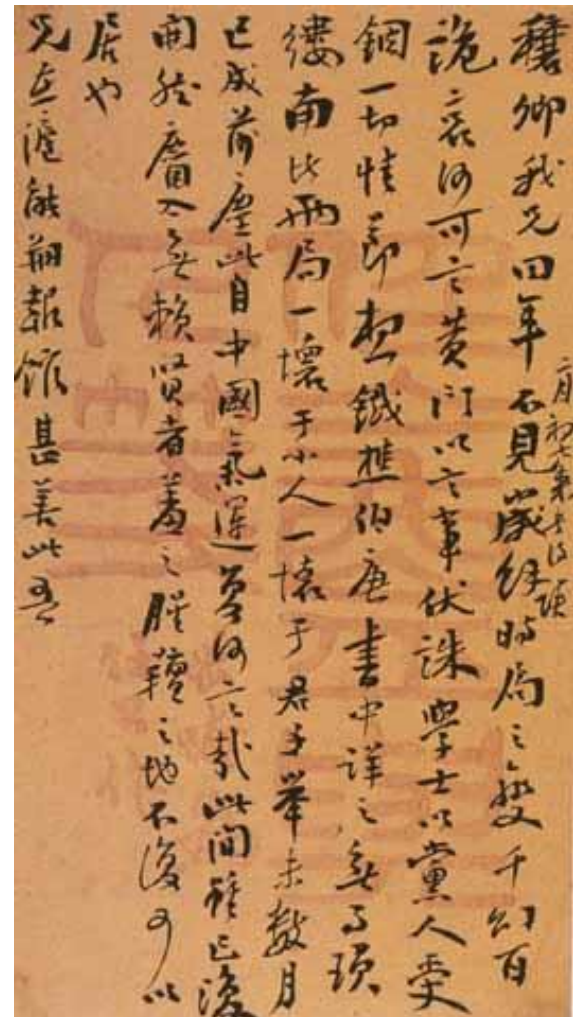
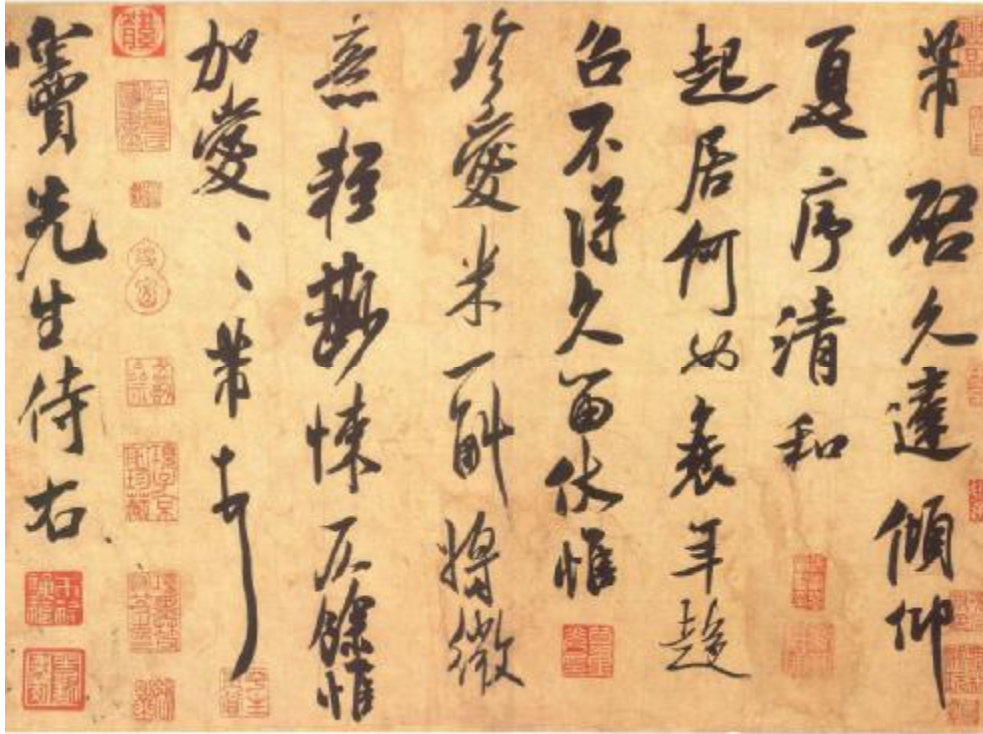
## Inter-character spacing

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.

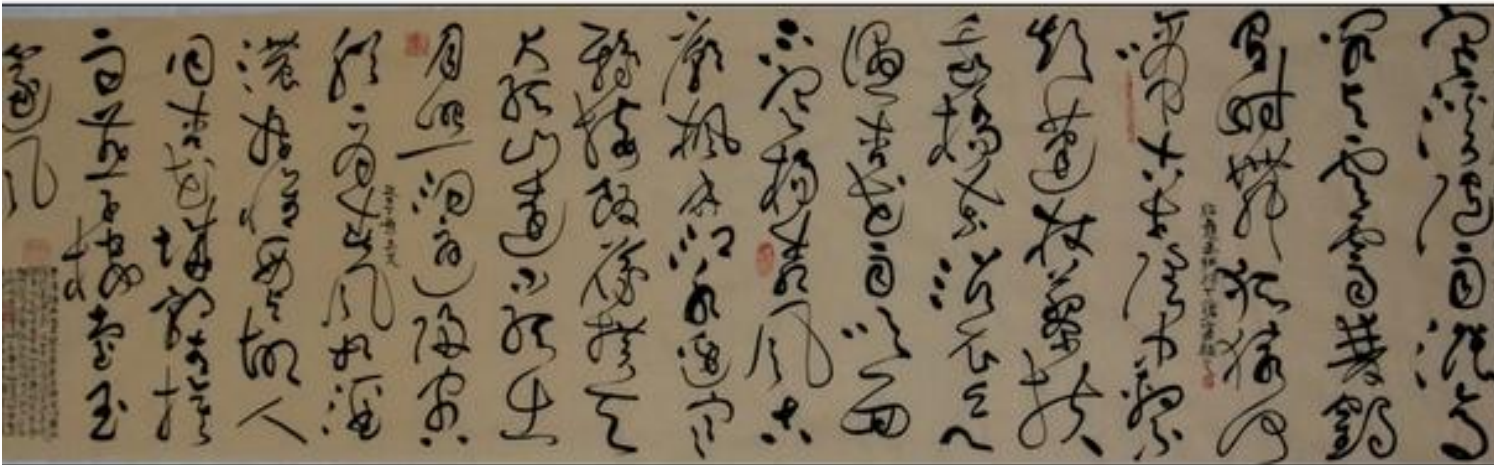
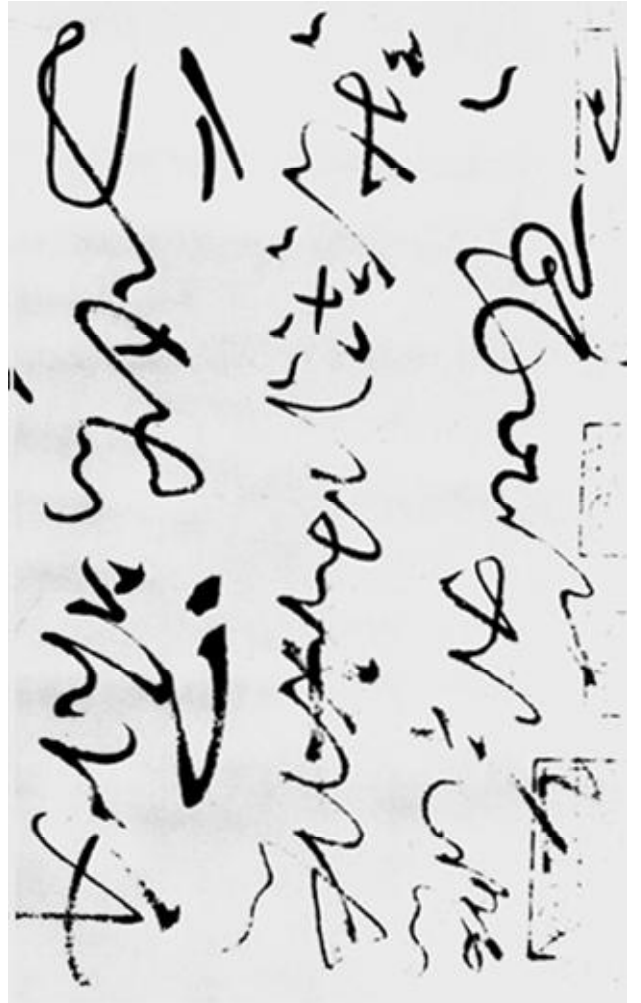
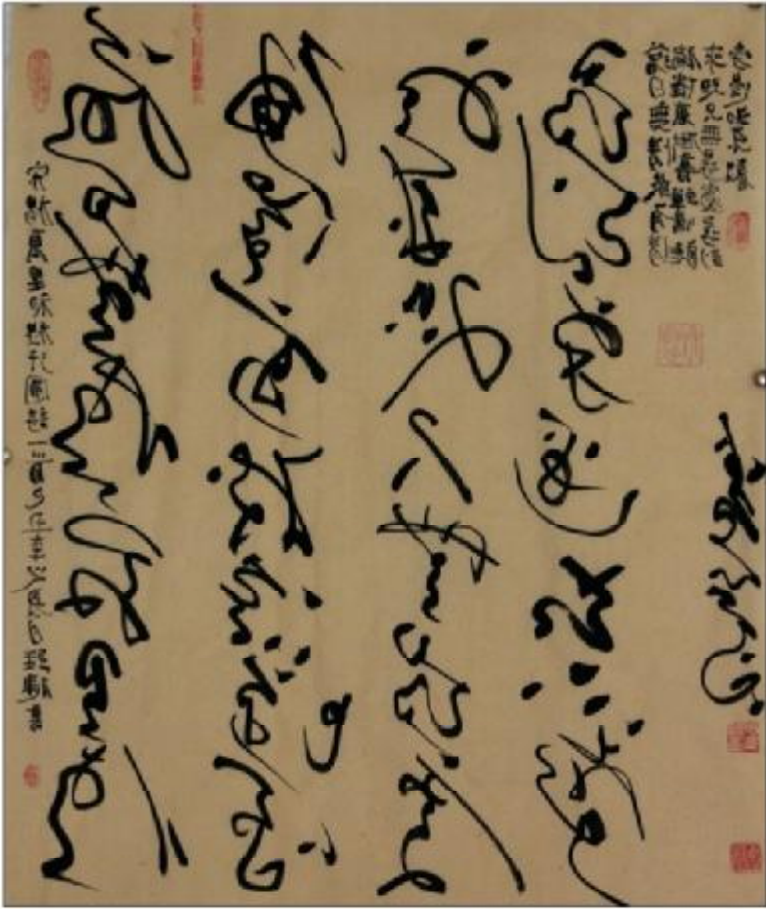


Figures 1, 2

司武也仕或陽即身大國除儀同三司道國公之苗慶此  
 授襄陽將軍隨格州括蒼縣丞大唐世子蕭舍曹懷  
 袖方傾襟榜平刃唯善為米挺秀主民胡陳水鏡脫  
 履千乘拂衣獨往塞惟方里夫人任氏舊蔭冠冕實  
 閨膏腴孝敬表於閨門信義洽於州里四時無以過  
 其信百金不取比其諾五志無違出言必踐因心披  
 揚自邑升天恭倫之性教自天默孝悌之心由於  
 志有規規之度為璋璋之寶內無聲色之好德給舖  
 藻若春秋八年有二貞觀歲甲午之年七月八日  
 殞於隆政之裏夫人春秋八十有二已亥之歲十月  
 廿三日薨于私第公則先祖必境夫人教誨長生  
 元輪迴悲忻永存今以庚子之歲正月己亥月廿三  
 日辛酉合葬於長安城西南十里土名高陽之禮禮  
 也歌其夫也如賓如相愛其子也有禮可觀其詞曰  
 長河帶地高岸擢木繁行不窮碑臚相繼百承樹意  
 沈沈蘭高門國於清規自然永辭世道長閨泉門  
 沈沈錄寒空丘墳春偷已暮隴樹哀吟天長地久  
 古往今來鑄石不磨千載方代嗚呼哀哉

Figures 3, 4

夫抗音投澗美惡必酬振服依河  
 長短交自斯乃德音道俗水鏡古  
 今法生傲蓬孝文皇帝專心於三  
 寶又遇北海母子崇信於二京妙  
 演之際屢叨未遑一降淨心忝充  
 五戒思樹芥子庶幾須彌今為  
 孝文并北海母子造像表情以申  
 接遇法生攝始王家身終夙霄締  
 敬歸功帝在万品衆生一切同福  
 魏景明四年二月一日比丘法  
 生并孝文皇帝并北海王母子造



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

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)