JTC 1/SC 2/WG 2 N3934

Title: A Comment on the proposal to unify Khitan script and CJK Ideograph, in N3925.
Date: 2010-10-04
Source: suzuki toshiya (mpsuzuki@hiroshima-u.ac.jp)
Status: an individual contribution

Original Chinese proposals to encode Khitan script in ISO/IEC 10646, N3820 and N3918 are designed to allocate a separated block for Khitan script. Considering the strong similarities between CJK ideographs and Khitan script, an idea to unify them is proposed by Berkeley experts, N3925.
The similarities between CJK ideographs and Khitan script is quite strong (in comparison with the preceding example of Hanzi-inspired script, Tangut script; it cannot share many structures with CJK ideograph). Furthermore, several characters are supposed to be borrowed from Hanzi. So it is reasonable to speculate the possibility to unify CJK ideograph and Khitan script.
N3925 proposes to discuss this issue in JTC1/SC2/WG2, but I have concern such unification may have unexpected impact to CJK ideograph users, so I propose to collect the comments about this unification idea from CJK ideograph users (via IRG) and from Khitan script users (via experts in WG2) for first. After the collection of their comments, WG2 should decide whether new Khitan script block should allocated and all Khitan script should be coded in the block.

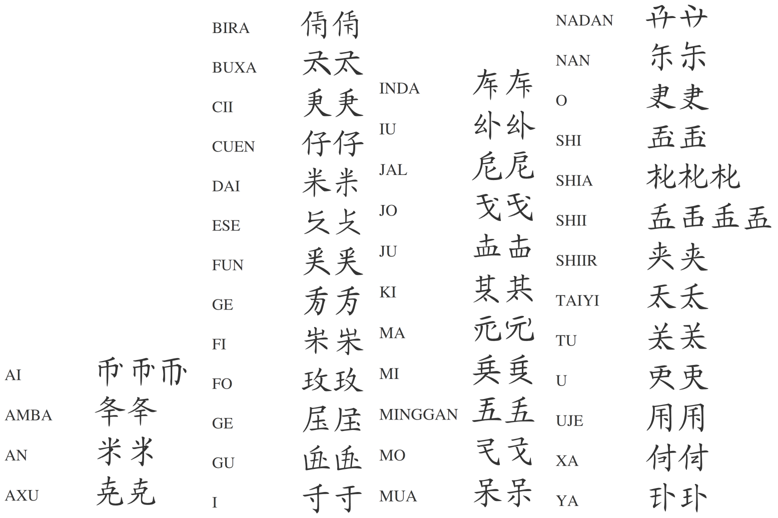
The impacts that I'm afraid are following:

1. Consistency with other Hanzi-inspired script(s)

Jurchen script may have quite similar position with Khitan script, because it can share many structures with CJK ideographs. Therefore, if Khitan script should be unified with CJK ideographs, the unification between Jurchen script and CJK ideographs should be speculated too. In fact, Jurchen script was used in longer time than Khitan script, the calligraphic similarities between Jurchen script and CJK ideographs in archaeological materials seem stronger than that between Khitan script and CJK ideographs.
Also, China NB is expected to provide the list of Hanzi-inspired historical scripts which are expected to be coded *out of* CJK ideograph block.

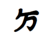
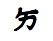
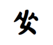
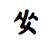
2. Consistency of Unification rule for CJK Unified/Compatibility Ideographs

According to Jurchen script expert's document, the criteria of unifiable glyphic variants may be different between Hanzi and Hanzi-inspired scripts. For example, N3817 has an analysis by Andrew West, for possible glyph variants that Chinese experts want to distinguish at character encoding level.



Jurchen characters with similar shapes but proposed to be coded separately (N3817)

In Khitan script, there might be similar problem. Some pairs/trios of Khitan characters (e.g. J-0251/J-0252, J-0298/J-0299, J-0313/J-0314, etc) in N3918 show quite similar shapes.

J-0251		岈	Unkown-105	200		
J-0252		岘	Unkown-106	201		
J-0298		彷	ô	252		
J-0299		徂	ô	253		
J-0313		衢	ŋ	264		
J-0314		彡	Unkown-123	265		

Khitan characters with similar shapes but proposed to be coded separately (N3918)

If they are proposed as CJK ideographs, they would be unified, as far as they don't have different meaning or pronunciations and regarded as non-cognate characters. Unfortunately, Khitan script is not decoded completely, it is difficult to determine

J-0251/J-0252 and J-0298/J-0299 are non-cognate or not (for many characters, meanings and pronunciations are described as *unknown*).

3. Script identification in the plain text mixing CJK Ideograph and Khitan script

The scholars of Khitan script who are interested in the standardization of its coding are mainly from China, so the typesetting mixing Chinese and Khitan scripts would be popular and important use-cases. Some Khitan characters are quite similar shape with existing CJK ideograph, but different meaning. For example, J-0051 is almost impossible to distinguish from "十" (U+5341, numeral 10 in CJK ideograph) by its shape. But according to N3918, its meaning is WEST. Therefore, in the typesetting mixing Khitan and Chinese text, the differentiation by different typefaces would be expected. In fact, Chinese proposals of Jurchen and Khitan script encodings, the referential glyphs are designed in KaiTi-style, not in SongTi-style (it is reasonable, because most archaeological materials of Jurchen and Khitan script show KaiTi-style glyphs, because both scripts are developed and obsoleted before mass printing). If Khitan and CJK ideographs are coded in same block, it is difficult for plain text to assign different typefaces to each script. Some markup languages are required. It increases the difficulties to digitize the archaeological materials for Khitan scripts. It is expected to hear the comments from the experts of the intelligent font and text layout systems.

4. Collation

Although ISO/IEC 10646 does not define any collation rule for CJK ideographs, the codepoints in each block are ordered to imitate KangXiZiDian. It seems that Jurchen and Khitan script scholars use different radical systems (e.g. 十-like J-0051, 支-like J-0055, 木-like J-0065, 皮-like J-0076, 土-like J-0079 are classified to same radical). Ordering Khitan characters by KX-like rules can introduce some inconvenience to search. It is expected to hear the comments from the users & authors of the dictionary of Khitan script.

By these concerned impacts, I think JTC1/SC2/WG2 should not decide the unification of CJK ideographs and Khitan script quickly within Busan meeting. The feedbacks from IRG and the experts from Khitan, Jurchen scripts (and other Hanzi-inspired scripts in East Asia) should be collected as the information for WG2 decision.

end of document