Request to Redefine the Scope of the Ideographic Variation Database

Andrew West

25 October 2016

In "Preliminary proposal to define 357 variation sequences for Tangut ideographs" (L2/16-111), I proposed the definition of a large number of standardized variation sequences for Tangut ideographs. This proposal was discussed at UTC #147 (May 2016), and there was no consensus to accept this proposal.

Given that there is a demand from the user community to be able to differentiate significant glyph variants of Tangut characters in plain text (see L2/16-243, "Summary of Meeting on Khitan Scripts, 20 August 2016 (Yinchuan, China) - Ad Hoc Report #1"), but the UTC is not willing to maintain a list of standardized variants for Tangut ideographs, a compromise solution would be to allow for the registration of Tangut glyph variants in the same way that CJK ideographic variants may be registered in the Ideographic Variation Database (IVD).

I believe that a few small changes to the definition of the IVD in UTS #37 "Unicode Ideographic Variation Database" would allow the IVD to accept registrations of Ideographic Variation Sequences (IVS) for Tangut and any ideographic script, and not just be limited to CJK unified ideographs. The main change would be to modify the definition of an IVS in section 2 to specify that the initial character for an IVS must have the **Ideographic** property rather than the **Unified_Ideograph** property. As the Unicode Standard specifies that the initial character for a variation sequence may not be canonically decomposable (http://www.unicode.org/versions/Unicode9.0.0/ch23.pdf p. 839), it would also be necessary to qualify the definition to exclude canonical decomposable characters (i.e. CJK compatibility ideographs), as suggested below:

> An Ideographic Variation Sequence (IVS) is a sequence of two coded characters, the first being a character with the ~~Unified_Ideograph~~ Ideographic property that is not a canonical decomposable character, the second being a variation selector character in the range U+E0100 to U+E01EF.

Additional explanatory text would also be required, and most occurrences of 'unified ideograph' would need to be changed to 'ideograph', but the above change is the only formal change to definition that would be needed.

This change would expand the list of allowable IVS initial characters to the following Unicode 9.0 characters (those not currently allowed are highlighted in yellow).

Allowable IVS Initial Characters under the New Definition

Block	Code Points	Character Names
CJK Symbols and Punctuation	3006	IDEOGRAPHIC CLOSING MARK
	3007	IDEOGRAPHIC NUMBER ZERO
	3021	HANGZHOU NUMERAL ONE
	3022	HANGZHOU NUMERAL TWO

Block	Code Points	Character Names
	3023	HANGZHOU NUMERAL THREE
	3024	HANGZHOU NUMERAL FOUR
	3025	HANGZHOU NUMERAL FIVE
	3026	HANGZHOU NUMERAL SIX
	3027	HANGZHOU NUMERAL SEVEN
	3028	HANGZHOU NUMERAL EIGHT
	3029	HANGZHOU NUMERAL NINE
CJK Unified Ideographs Extension A	3400..4DB5	CJK UNIFIED IDEOGRAPH-3400.. CJK UNIFIED IDEOGRAPH-4DB5
CJK Unified Ideographs	4E00..9FD5	CJK UNIFIED IDEOGRAPH-4E00.. CJK UNIFIED IDEOGRAPH-9FD5
CJK Compatibility Ideographs	FA0E	CJK COMPATIBILITY IDEOGRAPH-FA0E
	FA0F	CJK COMPATIBILITY IDEOGRAPH-FA0F
	FA11	CJK COMPATIBILITY IDEOGRAPH-FA11
	FA13	CJK COMPATIBILITY IDEOGRAPH-FA13
	FA14	CJK COMPATIBILITY IDEOGRAPH-FA14
	FA1F	CJK COMPATIBILITY IDEOGRAPH-FA1F
	FA21	CJK COMPATIBILITY IDEOGRAPH-FA21
	FA23	CJK COMPATIBILITY IDEOGRAPH-FA23
	FA24	CJK COMPATIBILITY IDEOGRAPH-FA24
	FA27	CJK COMPATIBILITY IDEOGRAPH-FA27
	FA28	CJK COMPATIBILITY IDEOGRAPH-FA28
	FA29	CJK COMPATIBILITY IDEOGRAPH-FA29
Tangut	17000..187EC	TANGUT IDEOGRAPH-17000.. TANGUT IDEOGRAPH-187EC
Tangut Components	18800..18AF2	TANGUT COMPONENT-001.. TANGUT COMPONENT-755
CJK Unified Ideographs Extension B	20000..2A6D6	CJK UNIFIED IDEOGRAPH-20000.. CJK UNIFIED IDEOGRAPH-2A6D6
CJK Unified Ideographs Extension C	2A700..2B734	CJK UNIFIED IDEOGRAPH-2A700.. CJK UNIFIED IDEOGRAPH-2B734
CJK Unified Ideographs Extension D	2B740..2B81D	CJK UNIFIED IDEOGRAPH-2B740.. CJK UNIFIED IDEOGRAPH-2B81D
CJK Unified Ideographs Extension E	2B820..2CEA1	CJK UNIFIED IDEOGRAPH-2B820.. CJK UNIFIED IDEOGRAPH-2CEA1

I think that it is acceptable to allow the eleven characters in the CJK Symbols and Punctuation block with Ideographic property to act as IVS initial characters.

The following two characters do not have the ideographic property, and would be excluded. However, it seems to me that they should be classified as ideographic, and allowed to act as an initial character in an IVS sequence. Therefore, I suggest giving them (and 16FE1 NUSHU ITERATION MARK in Unicode 10.0) the Ideographic property.

3005 IDEOGRAPHIC ITERATION MARK

16FE0 TANGUT ITERATION MARK

Scripts that are currently unencoded or are in the process of being encoded which should be given the ideographic property, and therefore should be allowed to act as initial characters in IVS sequences, include:

- Nushu
- Khitan Large Script
- Jurchen
- Small Seal Script
- Oracle Bone Script