Comments on positional mismatches in Mongolian encoding

Author: SHEN Yilei (沈逸磊) (shenyileirob@qq.com)

Date: 2017-08-11

[The author would like to thank LIANG Hai (梁海), MA Xudong (马旭东), WANG Yihua (王奕桦), YAN Shi (嚴實), and ZHENG Weizhe (郑维喆), among others, for their insightful discussion into this issue.]

This paper discusses the problems of positional mismatches that exibit in the current Mongolian encoding of the Unicode Standard. It is argued that these mismatches should be rectified at a later point, and a full list of mismatches is given.

0 Preliminary notes

Owing to the pervasive editorial errors in the transfer of the variant specification to the code charts since TUS7.0, the Mongolian variant specification in the present discussion is based on StandardizedVariants.html (henceforth SV.html) prior to TUS8.0, the Mongolian section of which has remained unchanged since TUS3.2. We shall not blame the editors as SV.html did not overtly list no-FVS forms, and when attempting to recover these forms in the code charts, errors are almost inevitably made on whether there is a no-FVS form and what it is. In view of this, I have resorted to TR170, the document most conforming to SV.html yet having a full specification of the variants, for the no-FVS forms. A full chart of the present Mongolian variant specifications in TUS without editorial errors is given in Appendix A.

In addition, it should be noted that the correspondence between SV.html and TR170 is based on FVS assignments rather than on the names of variants. There are several naming discrepancies between SV.html and TR170: to name a few, "1st medial" and "2nd medial" are swapped for U+1825, U+1826, and U+1836 respectively. Therefore it is rather confusing to refer to the variants by their names, and labels "no-FVS", "FVS1", "FVS2", and "FVS3" are used instead throughout this paper.

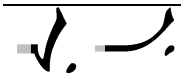
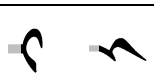
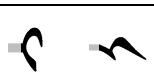
1 Introduction

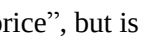
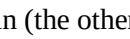
The **positional mismatch** to be discussed in this paper is the mismatch between the genuine cursive *glyph types* (cursive positions) of Mongolian variant forms and the stipulated counterparts in the current Unicode Standard. These positional mismatches are problems inherited since the finalizing of the Mongolian encoding project, but various implementations (notably the two major shaping engines Uniscribe and HarfBuzz) have ever since unanimously assumed the genuine positions disregarding the standard. Having suffered from the headache caused by the mismatch for ages, the W3C Mongolian forum agreed upon the identification of 7 notorious mismatch cases in 2015, and were going to ask the UTC to fix them, but the proposal unfortunately came to nothing in the end. Nevertheless, the gist of the proposal is embodied in their latest documents (L2/17-124 and L2-17/128).

These 7 noted cases of positional mismatches are:

Glyph[1]	Code point	Xlit.	Current spec.	Usage	Proposed change[2]

[1] The Mongolian script is rendered first in printing style (White) and then in handwriting style (Hawang) throughout this paper.
[2] As agreed upon in the W3C Mongolian forum (public-i18n-mongolian@w3.org). See https://lists.w3.org/Archives/Public/public-i18n-mongolian/2015JulSep/0273.html.

		1820	*a*	FVS2 medial	Post-NNBSP	FVS1 initial
		1828	*n*	FVS2 medial	Pre-MVS or word-final	FVS1 final
		182C	*x*	FVS2 medial	Pre-MVS (archaic)	FVS1 final
		182C	*x*	FVS3 medial	Pre-MVS	No-FVS final
		182D	*g*	FVS2 medial	Pre-MVS	FVS3 final
		1835	*ǰ*	FVS1 medial	Pre-MVS (archaic)	FVS1 final
		1836	*y*	FVS2 medial	Pre-MVS	No-FVS final

a's FVS2 medial occurs only after NNBSP, as the first letterform in the Hudum Mongolian (henceforth Hudum) ablative enclitic *–ača*  . The remaining 6 cases are characterized by their occurring (nearly) solely before MVSes. In particular, *n*'s FVS2 medial is mostly used in Hudum as an onset consonant before a MVS, as in Hudum *ün_e*  "price", but is also used in transcribing loan words occasionally both in Hudum and Manchu, as in Manchu *han"*  (transcription of Mandarin syllable *han*). As the glyphs suggest, these forms either end (*a*'s FVS2 medial) or begin (the other 6 cases) in a cursively disjoined stroke, and are intuitively cursively initial or final respectively.

2 Problems

2.1 Contradicting the general cursive joining rules and making the implementation complicated

The first argument for wiping out the mismatches is that they contradict the general cursive-joining rules of the standard. The general cursive joining rules of Mongolian, as specified under the heading of *Cursive joining*, is essentially identical to the Arabic model, a simplified version of which is illustrated below:[3]

R4:	Dual_Joining → Dual_Joining.medi / {Dual_Joining, Join_Causing} __ {Dual_Joining, Join_Causing}
R5:	Dual_Joining → Dual_Joining.init / __ {Dual_Joining, Join_Causing}
R6:	Dual_Joining → Dual_Joining.fina / {Dual_Joining, Join_Causing} __
R7:	Dual_Joining → Dual_Joining.isol

Being mutually exclusive, the transformational rules R4~R7 apply in this disjunctive order. For example, R5 applies only when R4 is not applicable and the contextual condition of R5 is satisfied, and its application in turn blocks R6 and R7.

Let's apply the general cursive joining rules to the above-mentioned mismatch cases.

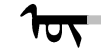
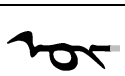
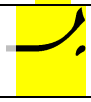
–ača  :

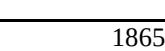
Code point	200D	1820	1834	1820
Char. name	NNBSP	ML. A	ML. CHA	ML. A
Joining type	Non_Joining	Dual_Joining	Dual_Joining	Dual_Joining
Joining rule		R5	R4	R6
Resultant position		Initial	Medial	Final

[3] Joining types Left_Joining, Right_Joining and Transparent, which are largely irrelevant here, are omitted in the formulation.

Post-shaping glyph				
				
Mismatched position		Medial	(No mismatch)	(No mismatch)

ün_e :

Code point	1826	1828	180E	1821
Char. name	ML. UE	ML. NA	MVS	ML. E
Joining type	Dual_Joining	Dual_Joining	Non_Joining	Dual_Joining
Joining rule	R5	R6		R7
Resultant position	Initial	Final		Isolate
Post-shaping glyph				
				
Mismatched position	(No mismatch)	Medial		Final[4]

han"  :

Code point	1865	1820	1828	180C
Char. name	MLS. HA	ML. A	ML. NA	FVS2
Joining type	Dual_Joining	Dual_Joining	Dual_Joining	Transparent
Joining rule	R5	R4	R6	
Resultant position	Initial	Medial	Final	
Post-shaping glyph				
				
Mismatched position	(No mismatch)	(No mismatch)	Medial	

The original cursive joining rules are fairly intuitive, and various implementations have stuck to the general Arabic model. However, the original simplicity is disrupted as two bizarre provisions are introduced later in the section of Mongolian when addressing NNBSP and MVS, which read:

> NNBSP affects the form of the preceding and following letters. The final letter of the stem or suffix preceding the NNBSP takes the final positional form, whereas the first letter of the suffix following NNBSP may take the normal initial form, a variant initial form, a medial form, or a final form, depending on the particular suffix. (Core Spec. of TUS9, p. 533)
>
> The MVS has a twofold effect on shaping. On the one hand, it always selects the forward tail form of a following letter *a* or *e*. On the other hand, it may affect the form of the preceding letter. The particular form that is taken by a letter preceding an MVS depends on the particular letter and in some cases on whether traditional or modern orthography is being used. (ibid., p. 534)

These two provisions open up the possibility of positional mismatch. But is such stipulation desirable from a technical perspective? Not at all. If one were to faithfully implement this scheme, they would postulate additional rules preempting the general rules, which should be built into the engine in an OpenType framework:

R4:	Dual_Joining → Dual_Joining<medi> / {Dual_Joining, Join_Causing} __ {Dual_Joining, Join_Causing}
	a → a<medi> / NNBSP __ {Dual_Joining, Join_Causing}
R5:	Dual_Joining → Dual_Joining<init> / __ {Dual_Joining, Join_Causing}
	n → n<medi> / {Dual_Joining, Join_Causing} __ {FVS2, MVS}
	x → x<medi> / {Dual_Joining, Join_Causing} __ MVS
	x → x<medi> / {Dual_Joining, Join_Causing} __ FVS3 MVS
	g → g<medi> / {Dual_Joining, Join_Causing} __ MVS

[4] The disjoint tail is another case of mismatch to be addressed later in this paper.

	ǰ → ǰ<medi> / {Dual_Joining, Join_Causing} __ MVS
	y → y<medi> / {Dual_Joining, Join_Causing} __ MVS
R6:	Dual_Joining → Dual_Joining<fina> / {Dual_Joining, Join_Causing} __
R7:	Dual_Joining → Dual_Joining<isol>

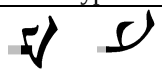
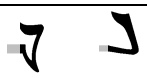
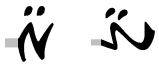
However, as the rules above would serve only tagging the glyphs with positional features, the font designers would still need to partially duplicate these rules in the GSUB table as below, so as to specify the required specific variants within the underdetermined variant paradigms:

a<medi> → a.medi3 / NNBSP __
n<medi> → n.medi3 / __ {FVS2, MVS}
x<medi> → x.medi3 / __ MVS
x<medi> → x.medi4 / __ FVS3
g<medi> → g.medi3 / __ MVS
ǰ<medi> → ǰ.medi2 / __ MVS
y<medi> → y.medi3 / __ MVS

That being the case, it is obvious why no engine would follow the standard. Other things being equal, no one would favor a two-step scheme that can be equivalently carried out in one go.

2.2 Contradicting both users' intuition and the grammatical tradition

Things get worse when this complication is exposed to users in metalanguage, where one may wish to render these mismatched variants out of context. If one wants to list all pre-MVS consonant forms with MVS absent, they are faced with the following paradigm of representation:

Xlit.	Current spec.	Char. sequence	Glyph
m	No-FVS final	ZWJ, __	
l	No-FVS final	ZWJ, __	
s	No-FVS final	ZWJ, __	
	FVS1 final	ZWJ, __	
š	No-FVS final	ZWJ, __	
r	No-FVS final	ZWJ, __	
w	FVS1 final	ZWJ, __	
n	FVS2 medial	ZWJ, __, FVS2, **ZWJ**	
	No-FVS final	ZWJ, __	
g	FVS2 medial	ZWJ, __, FVS2, **ZWJ**	
	No-FVS final	ZWJ, __	
x	FVS2 medial	ZWJ, __, FVS2, **ZWJ**	

	FVS3 medial	ZWJ, __, FVS3, **ZWJ**		
ǰ	FVS1 medial	ZWJ, __, FVS1, **ZWJ**		
y	FVS2 medial	ZWJ, __, FVS2, **ZWJ**		

Not only are letters that occur before MVS divided into unmismatched (*m*, *l*, *s*, *š*, *r*, *w*) and mismatched (*x*, *ǰ*, *y*) ones, but discrepancies arise within a single letter (*n*, *g*) as well, though the variants of *n* and *g* in question differ only in dotting. This highly irregular pattern defies all mnemonics.

It is more irrational that the first letterform in the masculine ablative enclitic *–ača*  is medial yet the first letterform in the feminine ablative enclitic *–eče*  is initial, though the two forms look exactly the same. Owing to the mismatch, one has to type <*e*, *č*, *e*> for *–eče* while <**ZWJ**, *a*, FVS2, *č*, *a*> for *–ača*. Jirimutu has commented in the W3C Mongolian mailing list as follows:[5]

> I cannot understand why Professor Quejingzhabu insist this A before [sic: after] NNBSP as medial form. We strongly disagree this definition.
> If anybody insist this as medial form. I would like ask add one more medial form to all of the other characters which is possible to use before NNBSP!!!

These mismatches should be attributed to the standard-setters' attempt to bring grammatical wordhood to the identification of positions. For example, the six pre-MVS forms occur mostly word-medially in Hudum, so they are identified as medial forms. However, this is clearly a misconstruction of the Unicode terms *isolate*, *initial*, *medial*, and *final*, which are instances of the *glyph type*. The *glyph type* concerns only joinedness of strokes at character junctures: a form is initial only when it begins with a disjoined juncture and ends with a joined juncture, etc. This is the case with the Unicode specifications for Arabic script, where most letters have one final and one isolate each. Had the Mongolian practice been applied to Arabic encoding, the isolates of most Arabic letters would be identified as FVS1 finals, apart from a few (*r*, *f*, *q*, *l*) that constitute well-formed words in isolation. Thus we would have:

Code point	Xlit.	(1st) final	"2nd final" (mismatched isolate)	Isolate
0631	*r*	ر	(Undefined)	ر
0632	*z*	ز	ز	(Undefined)

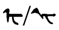
As a result, one would have to type something like <ZWJ, __, FVS1> to get most of the letters in isolation but to type directly the letters alone for *r*, *f*, *q*, and *l*, in a standard-conformant way. It seems utterly ridiculous, but is what is going on in the Mongolian encoding. The point is that linguistic wordhood should have no bearing on the cursive joining model.[6]

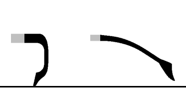
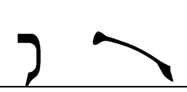
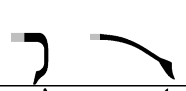
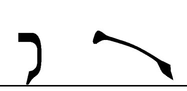
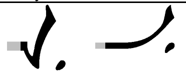
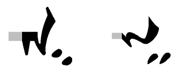
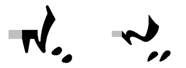
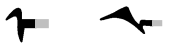
So far the readers are likely to get the impression that there is an established grammatical tradition of Mongolian in which the graphemic analysis of the Hudum script shall be done primarily with regard to the word boundary. As far as we know, however, things may well be the opposite. Chinggeltei's *蒙古语语法* (*Mongolian Grammar*, published in Chinese in 1991), a classic in this realm, groups all pre-MVS finals with ordinary finals rather than the medials. An influential dictionary *蒙汉词典* (*Mongolian–Chinese Dictionary*, published in Chinese in 1999) is similar in this respect. In fact, the present author has no material at hand which goes against this practice. Moreover, traditional Mongolian teaching has followed the same practice, as Myatav Erdenichimeg, the author of TR170, has pointed out that these mismatched cases are taught as intuitive initials or finals rather than mismatched medials.[7] Hence it is questionable whether the graphemic analysis primarily concerning the word boundary has ever gained any currency.

[5] https://lists.w3.org/Archives/Public/public-i18n-mongolian/2016JanMar/0017.html
[6] It may, however, be considered in line breaking, word counting or so, which is irrelevant here.
[7] https://lists.w3.org/Archives/Public/public-i18n-mongolian/2015JulSep/0198.html

3 A full list of mismatches

We have reached the conclusion in the previous section that the identification of positions should not take into consideration anything other than graphetic joinedness. In a thorough investigation into SV.html with this principle, we find 30 cases of mismatch in total. These mismatches fall into 5 classes according to their causes of mismatch:

- Type A: disjointed tails. Disjointed tails, as variants of U+1820 and U+1821, occurs only word-finally after MVS. They are isolates instead of finals as they are disjoined at both junctures.
- Type B: pre-MVS consonants forms. Most pre-MVS consonant forms are joined at the beginning juncture and disjoined at the ending juncture, and are thus finals. The only exception is  as in Hudum *bui-ǰ_a*   (particle), which is disjoined at both junctures, and is thus an isolate of *ǰ* (U+1835).
- Type C: post-NNBSP vowel forms. Post-NNBSP vowel forms are disjoined at the beginning juncture, and are thus initials and isolates rather than medials or finals.
- Type D: consonant isolates. Consonant except U+1835 have no genuine isolates. The so-called consonant isolates in the current standard are in fact initials and medials.
- Type E: straight tails in Todo. Two Todo letters, *e* (U+1844) and *d* (U+1851), end in the straight tail. In the current standard, *e*'s isolate  is subsumed by no-FVS initial  , *e*'s final  by no-FVS medial  , and *d*'s medial  by no-FVS final  , simply because the difference between a straight tail and a broken joined juncture is minimal in some fonts. These pairs should be separated.

Type	No.	Code point	Char. name	Slot	Glyph	Used in*	Should be	Subsumed glyph	Used in*	Should be	Note
A	1	1820	ML. A	FVS1 final		AA__			H___	Isolate	
	2	1821	ML. E	FVS1 final		H___			H___	Isolate	
B	3	1828	ML. NA	FVS2 medial		H_SM	Final				
	4	182C	ML. QA	FVS2 medial		H___	Final				
	5	182C	ML. QA	FVS3 medial		H___	Final				a
	6	182D	ML. GA	FVS2 medial		H___	Final				
	7	1835	ML. JA	FVS1 medial		H___	Final		H___	Isolate	
	8	1836	ML. YA	FVS2 medial		H___	Final				a
C	9	1820	ML. A	FVS2 medial		H___	Initial				
	10	1822	ML. I	No-FVS final		H___			H___	Isolate	
	11	1822	ML. I	No-FVS medial		H___			H___	Initial	
	12	1824	ML. U	No-FVS final		H___			H___	Isolate	
	13	1824	ML. U	No-FVS medial		H___			H___	Initial	

	14	1826	ML. UE	No-FVS final			H___				H___	Isolate	
	15	1826	ML. UE	No-FVS medial			H___				H___	Initial	
	16	1828	ML. NA	FVS3 medial			_T__	Initial					
	17	185E	MLS. I	No-FVS final			__S_				__S_	Isolate	
	18	1873	MLM. I	No-FVS final			___M				___M	Isolate	
	19	182C	ML. QA	No-FVS isolate			H___	Initial			H___	Medial	b c
	20	182C	ML. QA	FVS1 isolate			H___	Initial			H___	Medial	d
	21	182D	ML. GA	No-FVS isolate			H___	Initial					b e
	22	184E	MLT. GA	No-FVS isolate			_T__	Initial			_T__	Medial	b
D	23	1863	MLS. KA	No-FVS isolate			__S_	Initial			__S_	Medial	b
	24	1864	MLS. GA	No-FVS isolate			__SM	Initial			__SM	Medial	b
	25	1865	MLS. HA	No-FVS isolate			__SM	Initial			__SM	Medial	b
	26	1874	MLM. KA	No-FVS isolate			___M	Initial					b
	27	1889	MLAG. KA	No-FVS isolate			AA__	Medial					b
	28	1844	MLT. E	No-FVS initial			_T__				_T__	Isolate	
E	29	1844	MLT. E	No-FVS medial			_T__				_T__	Final	f g
	30	1851	MLT. DA	No-FVS final			_T__				_T__	Medial	h

[*] The four slots HTSM indicate usages in daily Hudum, Todo, Sibe, and Manchu respectively; A's indicate Galic-only usages.
[a] Duplicated as no-FVS final in TUS10 code chart by editorial error.
[b] Omitted in TUS10 code chart by editorial error.
[c] Moved to FVS1 isolate in the Chinese standards (GB/T 26226—2010 and GB/T 25914—2010).
[d] Moved to FVS2 isolate in the Chinese standards (GB/T 26226—2010 and GB/T 25914—2010).
[e] Moved to no-FVS initial in the Chinese standards (GB/T 26226—2010 and GB/T 25914—2010).
[f] No-FVS final added in the Chinese standard (GB/T 26226—2010).
[g] No-FVS final added in TUS10 code chart by editorial error.
[h] No-FVS medial added in TUS10 code chart by editorial error.

4 Representative glyphs and letter citation

It is noted that there are proposals that wishes to introduce more mismatches to the standard, notably Greg Eck's *DS01* (as of Dec 16, 2016). These mismatches are intended for showing representative glyphs and consonant citations. In the

former case, MONGOLIAN LETTER UE (U+1826) for example, the representative glyph originally as no-FVS initial (᠊/᠊ᠣ᠊) is proposed as FVS2 isolate, in addition to no-FVS isolate (᠊/᠊ᠣ) and FVS1 isolate (᠊/᠊ᠥ). In the latter case, MONGOLIAN LETTER TODO ANG (U+184A) for example, no-FVS medial is proposed as no-FVS isolate because the medial is used as a stand-in in letter citation in absence of a genuine isolate. These proposals go blatantly against the cursive joining model and should be dismissed immediately.

5 Summary

The positional mismatches in current Mongolian encoding:

- are illogical from a technical perspective;
- contradict users' intuition; and
- are not underpinned by a grammar tradition.

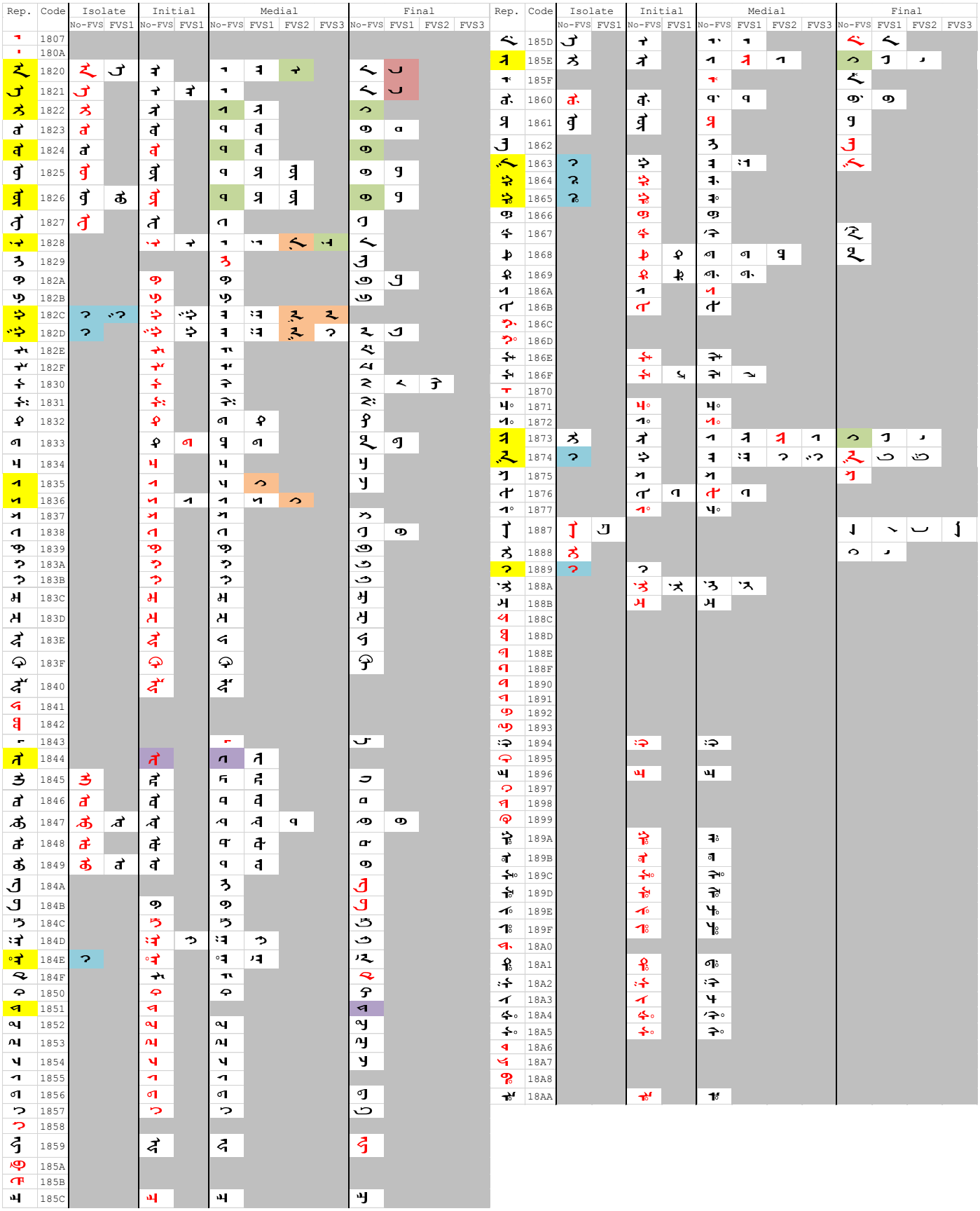
To clear up the mess, I request that these mismatches as listed in Section 3 be rectified at a later point when we have reached a consensus on the potential reassignments of variants, and that no more mismatches should be introduced in the future.

An excerpt of the resultant chart of Mongolian variants is shown below: (affected cells highlighted; colors indicating mismatch types; deleted cells rendered in grey)

Rep.	Code	Isolate			Initial				Medial						Final					
glyph	point	No-FVS	FVS1	New1	No-FVS	FVS1	New1	New2	No-FVS	FVS1	FVS2	FVS3	New1	New2	No-FVS	FVS1	FVS2	FVS3	New1	New2
᠊	1820	᠊	᠊	᠊	᠊		᠊		᠊	᠊	᠊				᠊	᠊				
᠊	1821	᠊		᠊	᠊	᠊			᠊						᠊	᠊				
᠊	1822	᠊		᠊	᠊		᠊		᠊	᠊					᠊					
᠊	1824	᠊		᠊	᠊		᠊		᠊	᠊					᠊					
᠊	1826	᠊	᠊	᠊	᠊		᠊		᠊	᠊	᠊				᠊	᠊				
᠊	1828				᠊	᠊	᠊		᠊	᠊	᠊	᠊			᠊				᠊	
᠊	182C	?	᠊?		᠊	᠊?	?	᠊?	᠊	᠊	᠊	᠊	?	᠊?					᠊	᠊
᠊	182D	?			᠊	᠊	?		᠊	᠊	᠊	?			᠊	᠊			᠊	
᠊	1835			᠊	᠊				᠊	᠊					᠊				᠊	
᠊	1836				᠊	᠊			᠊	᠊	᠊								᠊	
᠊	1844			᠊	᠊				᠊	᠊									᠊	
᠊	184E	?			᠊		?		᠊	᠊			?		᠊					
᠊	1851				᠊								᠊		᠊					
᠊	185E	᠊		᠊	᠊				᠊	᠊	᠊				᠊	᠊	᠊	᠊		
᠊	1863	?			᠊		?		᠊	᠊			?		᠊					
᠊	1864	?			᠊		?		᠊				?							
᠊	1865	?			᠊		?		᠊				?							
᠊	1873	᠊		᠊	᠊				᠊	᠊	᠊	᠊			᠊	᠊	᠊	᠊		
᠊	1874	?			᠊		?		᠊	᠊	?	᠊?			᠊	᠊	᠊			
?	1889	?			?								?							

A Full chart of the present Mongolian variant specifications in TUS without editorial errors

A full chart of the present Mongolian variant specifications in TUS without editorial errors (yet containing positional mismatches), based on SV.html completed with TR170 according to my arranging of the data, is given below for reference: (mismatches highlighted; references of representative glyphs in red)



B List of editorial errors in TUS10 Mongolian code chart

A full list of the editorial errors in TUS10 Mongolian code chart is given in the attached .xlsx file.