A mixed encoding scheme for the Mongolian block

Weizhe Zheng, with contributions from the Script Ad Hoc Group (especially Greg Eck, Hai Liang, Yilei Shen, Yihua Wang, and Shi Yan)

September 18, 2017

1. Introduction

We describe an encoding scheme for the Mongolian block based on a mixture of graphetic and phonetic considerations, consistent with the Script Ad Hoc recommendations (L2/17-243 and L2/17-328). Since the four scripts of the block follow different principles, we adopt strategies adapted to the individual scripts:

- graphetic for Mongolian;
- graphetic and phonetic for Todo;
- phonetic for Sibe and Manchu, with a couple of graphetic modifications.

1.1. Features

Under this scheme normal texts, including historical and Ali Gali texts, are encoded without format characters. Use of format characters is mostly limited to the following situations:

- Presentation forms can be encoded using ZWJ and/or variation selectors.

- If desired, some spelling mistakes can be encoded using variation selectors. For example, Manchu Ali Gali ᡍᠣ (mdo) is sometimes mistakenly written ᡍᠣ, which can be encoded using variation selectors. We note however that there are many types of spelling mistakes, and not all of them can be encoded in this way.

Most characters are simple, subject to cursive joining and having no variant forms. There are a total of 16 complex characters (1 for Todo, 15 for Sibe/Manchu) that do have variant forms. The automatic variant selection for complex characters is local, depending only on the following character (if any), the preceding character (if any) and the joining position of the preceding character. A complete description is given in Section 3.

1.2. Detailed strategies

For Mongolian we incorporate 14 new code points (N1..N14) of the Maximal Graphetic Approach from the Script Ad Hoc recommendations (L2/17-243). We decompose E long stem ᡐ as N2+180A(nirugu). Extending the graphetic analysis to Mongolian Ali Gali and historical forms in the current standard yields two more new code points (NA1 ᠌, NH2 ᠌).

For Todo/Sibe/Manchu, major improvements to the current phonetic encoding model are:

- Use of Syllable Boundary Marker (1807) in Todo. The current standard considers the syllable boundary marker in Todo as part of the following vowel, which is a major source of FVS usage. See 1820, 1844..1849.

- Disunification of the velar series (ᡍ ᡎ ᡏ ᡐ) from the uvular series (184D ᡍ, 184E ᡎ, 1863 ᡏ, 1864 ᡐ, 1865 ᡑ, 1874 ᡨ). The Todo velar letters are then unified with Ali Gali letters of the same glyphs and sounds (183B, 1889). The Sibe/Manchu velar letters are then merged with letters of the same sounds and similar glyphs used in borrowings (1858, 186C, 186D). The disunification of the uvular and velar series is especially important for Manchu, as both series can occur before the vowels e, i, ū in Manchu and Manchu Ali Gali.

For the following aspects of Todo/Sibe/Manchu, we adopt graphetic encoding:

- One new letter (N16) for the Sibe/Manchu genitive marker i ᡳ, eliminating the need for NNBSP or any other enclitic marker. The Todo enclitic -ni ᢏ is encoded with 180A (nirugu).

- In the two cases where automatic selection of variants could fail (NA and Manchu Ali Gali SA), graphetic encoding is used, resulting in two new code points (N15, NA7). For an alternative on the encoding of NA, see Section 4.1.

We have also included new code points (NA2..NA6, NH1, NH3) for Ali Gali and historical forms missing in the current standard. Most of these new code points (NA3..NA6, NH3) are needed regardless of encoding model.

2. Table of characters

Legends:

- N1..N17: new code points for modern texts; NA1..NA7: new code points for Ali Gali; NH1..NH3: new code points for historical texts.
- New code points, new variants, new usage are highlighted in yellow.
- Deprecated code points, deprecated variants, deprecated usage are highlighted in aqua.
- Glyph changes are shown in red.
- Unicode 10.0 glyphs in unattested positions are shown in silver.
- Ali Gali (AG) usage is indicated in an extended Wylie transliteration of Tibetan.

Except for a few code points (1868, 1869, 1873), glyphs are ordered according to the Unicode 10.0 code chart when applicable.

code point	complex?	name	glyphs				usage				notes
			isol	init	medi	fina	Mongolian	Todo	Sibe	Manchu	
N1		separate left tail	⟩				a, e				Non-joining. For an alternative, see Section 4.2.
N2		Alef-NA	ꠙ	ꠞ	ꠟ	ꠡ	a, e, ʔ, n	AG a.isol			Historical final forms handled by font. See Note 1.
N3		I-YA-JA	ꠀ	ꠁ	ꠂ	ꠃ	i, y, j				
N4		gedesue		ꠄ	ꠅ	꠆	o, u, ö, ü				
N5		bitegueue seguel	ꠇ			ꠈ	o, u, ö, ü, w				
N6		OE-UE			ꠉ	ꠊ	ö, ü				
N7		WA-EE		ꠋ	ꠌ	ꠍ	w, é	f	w	w	
N8		YA hooked		ꠎ	ꠏ		y		y	y	
N9		NA dotted		ꠐ	ꠑ	ꠒ	n	n	n	n	Historical final forms handled by font. See Note 1.
N10		QA-GA		ꠓ	ꠔ	ꠕ	q, ɣ(coda)				
N11		QA-GA dotted		ꠖ	ꠗ	ꠘ	ɣ				
N12		Taw		ꠙ	ꠚ	ꠛ	t, d, AG th				
N13		Taw coda			ꠜ	ꠝ	d				Historical final forms handled by font. See Note 1.
N14		Damed		ꠞ	ꠟ	ꠠ	t, d				
N15		Todo dotless NA			ꠡ	ꠢ		n (coda)	n (coda)	n (coda)	

code point	complex?	name	glyphs				usage				notes
			isol	init	medi	fina	Mongolian	Todo	Sibe	Manchu	
N16		Sibe small I	ʿ						i (genitive marker)	i (genitive marker)	Non-joining. Smaller than final form of Sibe/Manchu i in some sources
N17		Sibe MA	ᚑ	ᚑ	ᚑ				m	m	final form differs from 182E
NA1		Ali Gali left tail				ᚑ	AG a.fina	AG a.fina			
NA2		Ali Gali virama				(	AG virama (following k, g)				
NA3		Todo Ali Gali reversed I				ᚑ		AG I (rI for ṛ, lI for ḷ)			
NA4		Todo Ali Gali PA		ᚑ	ᚑ			AG p			
NA5		Manchu Ali Gali alternate RA		ᚑ	ᚑ					AG R (Ri/iRi/ ĭRi for ṛ)	
NA6		Manchu Ali Gali alternate LA		ᚑ	ᚑ					AG L (Li/iLi/ ĭLy for ḷ)	
NA7		Manchu Ali Gali SA with tail				ᚑ				AG s	
NH1		left dot					(used with i, u, ü in older documents)				
NH2		KE-GE dotted		ᚑ	ᚑ	ᚑ	k, g (old)			Old Manchu coda k (rare)	
NH3		CHA with two dots			ᚑ		Buryat sh (in shi)				
1807		Sibe syllable boundary marker		ᚑ				?	?	?	
180A		nirugu	▪	▪	▪		stem, e.medi	stem	stem	stem	
180E		vowel separator									deprecated
1820		A	ᚑ ᚑ	ᚑ	ᚑ ᚑ ᚑ	ᚑ ᚑ		a, AG a	a	a	Mongolian usage deprecated

code point	complex?	name	glyphs				usage				notes
			isol	init	medi	fina	Mongolian	Todo	Sibe	Manchu	
1821		E	㇆		↑	✓ ↲					deprecated
1822		I	↗	↟	↲ ↲	↲					deprecated
1823	✓	O	ㅂ	ㅂ	ㅂ ㅂ	ㅂ ㅁ			o	o	Mongolian usage deprecated
1824		U	ㅂ	ㅂ	ㅂ ㅂ	ㅂ					deprecated
1825		OE	ㅂ	ㅂ	ㅂ ㅂ ㅂ	ㅂ ㅂ					deprecated
1826		UE	ㅂ ㅂ	ㅂ	ㅂ ㅂ ㅂ	ㅂ ㅂ					deprecated
1827		EE	㇆	㇆	㇆	㇆					deprecated
1828		NA	㇆ ㇆	㇆ ㇆ ✓ ㇆	✓					deprecated
1829		ANG		㇆	㇆	㇆				ng	Mongolian usage deprecated (=N2+1889)
182A		BA	ㅂ	ㅂ	ㅂ ㇆	b		b	b	the historical final form ㇆ handled by font
182B		PA	ㅂ	ㅂ	ㅂ	p				
182C		QA	㇆ ㇆	㇆ ㇆	㇆ ㇆ ㇆ ㇆	㇆					deprecated
182D		GA	㇆ ㇆	㇆ ㇆	㇆ ㇆ ㇆ ㇆	㇆ ㇆					deprecated
182E		MA	㇆	㇆	㇆	m				Sibe/Manchu letter encoded as N18
182F		LA	㇆	㇆	㇆	l	l	l	l	
1830		SA	㇆	㇆	㇆ ✓ ㇆	s	s	s	s	the historical final form ✓ handled by font; ㇆ encoded as NA7
1831		SHA	㇆	㇆	㇆	sh	sh			

code point	complex?	name	glyphs				usage				notes
			isol	init	medi	fina	Mongolian	Todo	Sibe	Manchu	
1832		TA		ᠲ	ᠲ ᠲ	ᠲ					deprecated
1833		DA		ᠳ ᠳ	ᠳ ᠳ	ᠳ ᠳ					deprecated
1834		CHA		ᠴ	ᠴ	ᠴ	c, AG ch	z	c	c, AG ch	
1835		JA		ᠵ	ᠵ	ᠵ				j	Mongolian usage deprecated
1836		YA		ᠶ ᠶ	ᠶ ᠶ ᠶ	ᠶ					deprecated
1837		RA		ᠷ	ᠷ	ᠷ	r	r	r		
1838		WA		ᠸ	ᠸ	ᠸ ᠸ					deprecated
1839		FA		ᠹ	ᠹ	ᠹ	f, AG ph				
183A		KA		ᠺ	ᠺ	ᠺ	k', AG g	g', AG k			script usage rearranged with 1858
183B		KHA		ᠻ	ᠻ	ᠻ	k'(Mongolia), AG kh	k, AG kh			
183C		TSA		ᠼ	ᠼ	ᠼ	ts, AG tsh				
183D		ZA		ᠽ	ᠽ	ᠽ	dz				
183E		HAA		ᠾ	ᠾ	ᠾ					deprecated
183F		ZRA		ᠿ	ᠿ	ᠿ	rr				
1840		LHA		ᡀ	ᡀ	ᡀ					deprecated. = 182F LA + 1841 ZHI
1841		ZHI		ᡁ	ᡁ	ᡁ	zh, h.medi, h.fina	AG voiced aspiration			
1842		CHI		ᡂ	ᡂ	ᡂ	ch				
1843		Todo long vowel sign		ᡃ	ᡃ	ᡃ		lengthening of preceding vowel			could be analyzed as transparent to cursive joining
1844		Todo E	ᡄ	ᡄ	ᡄ ᡄ	ᡄ		e			ᡄ encoded as 1807(SBM)+1844
1845		Todo I	ᡅ	ᡅ	ᡅ ᡅ	ᡅ		i			
1846		Todo O	ᡆ	ᡆ	ᡆ ᡆ	ᡆ		o			

code point	complex?	name	glyphs				usage				notes
			isol	init	medi	fina	Mongolian	Todo	Sibe	Manchu	
1847		Todo U						u			the forms ᠊ ᠊ encoded as 1849; stylistic isolate variant handled by font
1848		Todo OE						ö			
1849		Todo UE						ü, u			stylistic isolate variant handled by font
184A		Todo ANG						ng			
184B		Todo BA						b			
184C		Todo PA						p, AG ph			
184D		Todo QA						q			velar letter ᠻ encoded as 183B
184E	✓	Todo GA						ɣ			(velar letter ᠻ encoded as 1889)
184F		Todo MA						m			
1850		Todo TA						t, AG th			
1851		Todo DA						d			᠊ in some sources
1852		Todo CHA						c			
1853		Todo JA						j			
1854		Todo TSA					j.medi/fina, AG c(rare)	ts			
1855		Todo YA						y			the stylistic variant ᠊ handled by font
1856		Todo WA						w			
1857		Todo KA						k'			
1858	✓	Todo GAA							k, k'	k, k', AG kh	script usage rearranged with 183A and 1863/1874
1859		Todo HAA						h			

code point	complex?	name	glyphs				usage				notes
			isol	init	medi	fina	Mongolian	Todo	Sibe	Manchu	
185A		Todo JIA						j (historical)			in fact a ligature: 184B Todo BA + 18A7 Ali Gali half YA
185B		Todo NIA					AG ny	AG ny			
185C		Todo DZA						dz			
185D	✓	Sibe E							e	e	
185E	✓	Sibe I							i		Sibe dzi treated as ligature
185F		Sibe IY							ǐ	ǐ	
1860	✓	Sibe UE							u	u	
1861		Sibe U							ū	ū	
1862		Sibe ANG							ng		
1863	✓	Sibe KA							q		(velar letter encoded as 1858)
1864		Sibe GA							ɣ	ɣ, AG g	Manchu AG ge treated as ligature (velar letter encoded as 186C)
1865		Sibe HA							χ	χ	(velar letter encoded as 186D)
1866		Sibe PA							p	p, AG ph	
1867		Sibe SHA							sh	sh	
1868	✓	Sibe TA							t	t, AG th	
1869	✓	Sibe DA							d	d	
186A		Sibe JA							j		

code point	complex?	name	glyphs				usage				notes
			isol	init	medi	fina	Mongolian	Todo	Sibe	Manchu	
186B		Sibe FA		ᆮ	ᆮ	ᆮ			f		
186C	✓	Sibe GAA	ᆑ	ᆑ ᆑ	ᆑ			g, g'	g, g', AG k	
186D	✓	Sibe HAA	ᆓ	ᆓ ᆓ	ᆓ			h, h'	h, h'	
186E		Sibe TSA		ᆨ	ᆨ	ᆨ			ts	ts, AG tsh	
186F		Sibe ZA		ᆩ ᆩ	ᆩ ᆩ	ᆩ			dz	dz	Sibe dzi ᆩ treated as ligature
1870		Sibe RAA		ᆪ	ᆪ	ᆪ			rr	rr	
1871		Sibe CHA		ᆫ	ᆫ	ᆫ			ch	ch, AG Th	
1872		Sibe ZHA		ᆬ	ᆬ	ᆬ			zh		
1873	✓	Manchu I	ᆭ	ᆭ	ᆭ ᆭ / ᆭ ᆭ	ᆭ ᆭ / ᆭ				i	Manchu ʔi ᆭ encoded as 1807(SBM)+1873
1874	✓	Manchu KA		ᆮ	ᆯ ᆯ / ᆯ ᆯ	ᆯ / ᆯ ᆯ				q	velar letter ᆮ encoded as 1858; dotted velar letter encoded as NH2
1875		Manchu RA		ᆰ	ᆰ	ᆰ				r	
1876	✓	Manchu FA		ᆱ ᆱ	ᆱ ᆱ	ᆱ				f	
1877		Manchu ZHA		ᆲ	ᆲ	ᆲ				zh, AG D	
1880		Ali Gali anusvara	ᆳ ᆳ				AG ṃ (at the end)	AG ṃ (at the end)		AG ṃ (at the end)	stylistic variants handled by fonts
1881		Ali Gali visarga	ᆴ ᆴ				AG ḥ	AG ḥ		AG ḥ	stylistic variants handled by fonts
1882		Ali Gali damaru	ᆵ				AG ḥ	AG ḥ		AG ḥ	
1883		Ali Gali ubadama	ᆶ				AG ḥ	AG ḥ		AG ḥ	
1884		Ali Gali inverted ubadama	ᆷ				AG ḥ	AG ḥ		AG ḥ	

code point	complex?	name	glyphs				usage				notes
			isol	init	medi	fina	Mongolian	Todo	Sibe	Manchu	
1885		Ali Gali baluda					AG 3	AG 3		AG 3	
1886		Ali Gali three baluda					AG 9	AG 9		AG 9	
1887		Ali Gali A								AG ﬁ	Mongolian usage deprecated
1888		Ali Gali I									deprecated. is a stylistic variant of 1822; is encoded as 1873.
1889		Ali Gali KA					k, g, AG k	g, AG g			
188A		Ali Gali NGA					AG ng	AG ng			AG nga treated as ligature
188B		Ali Gali CA					AG ts				
188C		Ali Gali TTA					AG T	AG T			Initial form evolved into 1841.
188D		Ali Gali TTHA					AG Th	AG Th			Should be unified with 1842.
188E		Ali Gali DDA					AG D	AG D			
188F		Ali Gali NNA					AG N	AG N		AG N	
1890		Ali Gali TA					AG t				
1891		Ali Gali DA					AG d				In fact a historical form of Damed. Should be unified with N14.
1892		Ali Gali PA					AG p			AG p	
1893		Ali Gali PHA									deprecated. Stylistic variant of 182B, not used in Ali Gali.
1894		Ali Gali SSA					AG S	AG S			

code point	complex?	name	glyphs				usage				notes
			isol	init	medi	fina	Mongolian	Todo	Sibe	Manchu	
1895		Ali Gali ZHA		ᢕ	ᢕ	ᢕ	AG zh				In fact a historical form of ZRA. Should be unified with 183F.
1896		Ali Gali ZA		ᢖ	ᢖ	ᢖ	AG z				
1897		Ali Gali AH		ᢗ	ᢗ	ᢗ	AG ẖ	AG ẖ			
1898		Todo Ali Gali TA		ᢘ	ᢘ	ᢘ		AG t			ᢘ in some sources
1899		Todo Ali Gali ZHA		ᢙ	ᢙ	ᢙ		AG zh			ᢙ in some sources
189A		Manchu Ali Gali GHA		ᢚ	ᢚ	ᢚ				AG gh	Manchu AG ghe ᢚ treated as ligature
189B		Manchu Ali Gali NGA		ᢛ	ᢛ	ᢛ				AG ng	
189C		Manchu Ali Gali CA		ᢜ	ᢜ	ᢜ				AG ts	
189D		Manchu Ali Gali JHA		ᢝ	ᢝ	ᢝ				AG dzh	
189E		Manchu Ali Gali TTA		ᢞ	ᢞ	ᢞ				AG T	
189F		Manchu Ali Gali DDHA		ᢟ	ᢟ	ᢟ				AG Dh	
18A0	✓	Manchu Ali Gali TA		ᢠ	ᢠ	ᢠ				AG t	
18A1	✓	Manchu Ali Gali DHA		ᢡ	ᢡ	ᢡ				AG dh	
18A2		Manchu Ali Gali SSA		ᢢ	ᢢ	ᢢ				AG S	
18A3		Manchu Ali Gali CYA		ᢣ	ᢣ	ᢣ				AG c	
18A4		Manchu Ali Gali ZHA		ᢤ	ᢤ	ᢤ				AG zh	

code point	complex?	name	glyphs				usage				notes
			isol	init	medi	fina	Mongolian	Todo	Sibe	Manchu	
18A5		Manchu Ali Gali ZA	ᢅ	ᢅ	ᢅ					AG z	
18A6		Ali Gali half U	ᢆ	ᢆ	ᢆ	AG wa					
18A7		Ali Gali half YA	ᢇ	ᢇ	ᢇ		AG y/ya				
18A8		Manchu Ali Gali BHA	ᢈ	ᢈ	ᢈ					AG bh	
18A9		dagalga									deprecated
18AA		Manchu Ali Gali LHA	ᢊ	ᢊ	ᢊ					AG lh	

Notes:

1. The final forms of certain Mongolian letters have historical stylistic variants, some inherited from Sogdian. The following quotation from Pandey's revised proposal to encode Sogdian L2/16-371r2 applies to these letters:

   > The terminal strokes may be oriented in different directions, even within the same line. The orientation of terminals vary according to the whim of the scribe or the space available on a page. Terminal variation occurs most often at the end of a line for filling space or for compensating for lack of space at a margin. These stroke variations are stylistic and there is no semantic difference between final forms with different terminals. … Alternate final forms are considered glyphic variants and may be controlled through fonts.

2. In Todo Ali Gali, the ts series and c series differ from source to source. These are not identified in the table.

3. Automatic variant selection

There are 16 complex characters that have variant forms, including 11 consonant characters and 5 vowel characters.

3.1. Consonants

The 11 complex consonant letters have variants in the initial or medial positions. The automatic variant selection depends on the following character and whether the preceding character (if any) is a vowel.

For the 3 consonant letters that have variants in the medial position only, the selection depends only on syllabification.

character	onset or stray	coda
184E (medi)	ᡎ	ᡎ
1863 (medi)	ᡣ	ᡣ
1874 (medi)	ᡴ	ᡴ

For the 8 consonant letters that have variants in both the initial and medial positions, the selection depends on the following character, and on syllabification for 1868 and 1858.

character	onset or stray			coda (medi)
	condition: followed by	conditional form	default form	
1868 (init/medi)	e/u/ū	᠊ᡐ ᡐ	᠊ᡒ ᡒ	ᡖ
1869 (init/medi)		᠊ᡑ ᡑ	᠊ᡓ ᡓ	
18A1 (init/medi)		ᢉ᠊᠅ ᢉ᠅	ᢎ᠅ ᢎ᠅	
18A0 (init/medi)		ᢏ	ᢏ	
1858 (init/medi)		᠘	᠘	᠘
186C (init/medi)	e/i/u/ū	ᡌ	ᡌ	
186D (init/medi)		ᡍ	ᡍ	
1876 (init/medi)	i/o/u/ū	ᡶ	ᡶ	

3.2 Vowels

The 5 complex vowel letters have variants in the medial or final positions. The automatic variant selection depends on the preceding character and the joining position of the preceding character. In the table below, for each variant form, the class of the preceding character is given. T stands for 1868/1869/18A0/18A1.

character	form 1	form 2	form 3	form 4
185D (medi/fina)	(default) ᠎ᡝ	T ᠎ᡝ	1807/1887/189B ᠎᠊	
1860 (medi)	(default) ᠎ᡠ	T ᠎ᡠ		
1860 (fina)	(non-T).medi ᠎ᡠ	T.medi ᠎ᡠ	(non-T).init ᠎ᡠ	T.init ᠎ᡠ
1823 (fina)	medi ᠎ᡳ	init ᡳ		
185E (medi)	(default) ᡝ	vowel ᡝ		
1873 (medi)	(default) ᡝ	vowel (except 185E/1873) ᡝ		
1873 (fina)	(default) ᠎ᡳ	186F/189C/189D ᠎᠊		

4. Alternatives

4.1. Phonetic NA in Todo/Sibe/Manchu

Liang Hai, Shen Yilei, and Yan Shi suggested encoding the letter NA in Todo/Sibe/Manchu phonetically. Instead of using N9 and introducing N15, they propose to modify the current letter NA as follows.

code point	complex?	name	glyphs				usage				notes
			isol	init	medi	fina	Mongolian	Todo	Sibe	Manchu	
1828	✓	NA	᠘	᠈ᠨ	᠈ᠨ᠈ ᠈ᠨ᠊	᠊ᠨ ᠊ᠨ		n	n	n	Mongolian usage deprecated

The two medial forms are selected automatically.

character	onset or stray	coda
1828 (medi)	ᠨ	ᠨ

The downside of this alternative is that the final forms cannot be selected automatically: The dotted final form has to be selected manually using variation selectors.

4.2. Unification of two left tail characters

Greg Eck suggested that the non-joining left tail (N1) be unified with the joining left tail (NA1). The non-joining left tail would then be represented by prepending a ZWNJ or a space.

4.3. More new characters for maximal backward compatibility

The mixed scheme modifies the behavior of 8 Todo/Sibe/Manchu characters: 1845, 1847, 184D, 184E, 1863..1865, 1874. Although Unicode usage in these scripts is limited, Yilei Shen suggested deprecating these code points and introducing new characters instead, in order to ensure maximal backward compatibility.