Towards a well-formed Mongolian specification that allows interoperable implementations

(标题大意: 走向一份格式良好的蒙古文规范, 让可互通的实作成为可能)

To: MWG #3 (Unicode Mongolian Working Group Meeting 3), Ulaanbaatar
From: Liang Hai / 梁海 <lianghai@gmail.com>
Date: 4 April 2019

1 Introduction

Two decades ago, Unicode/UCS added support for the Mongolian script with a seemingly sensible character set. Little was known at the time about how exactly these characters were meant to be rendered. Various vendors have since struggled to make their own senses of the characters, while users have been suffering from both the poor interoperability between vendor implementations and the consequent lack of native support on major platforms.

Experts have been uncovering issues of the Mongolian encoding and have proposed various patches, especially over the last few years. However, until very recently (Liang Jinbao 2018), few complete specifications have ever been available for the community to discuss and evolve in order to eventually reach an agreement.

This specification

Although the architectural defects cannot be resolved without switching to another encoding model, a rigorous specification can help eliminate unwanted differences between vendor implementations.

This document demonstrates and proposes how to specify both comprehensive encoding guidelines for text representation and coherent shaping rules for text rendering. It mainly deals with the first three layers in the overall technical architecture of the Hudum encoding support:

- The Unicode Standard and ISO/IEC 10646 (UCS), a synchronized pair of standards, specify the encoded characters and standardized variation sequences.

- The Unicode Standard and its various supplementary standards also provide the characters with additional behavioral specifications, including character properties (general category, cursive joining type, etc.) and algorithms (normalization, collation, line breaking, text segmentation, bidirectional, vertical text layout, etc.).

- Fonts and shaping engines with OpenType or an equivalent font technology (e.g., AAT and Graphite) are relied for implementing the required complex shaping.

- Hudum is treated inline as a horizontal, left-to-right writing system, while layout engines are responsible for setting lines vertically and arranging multiple lines with the preferred left-to-right order.

A temporary scope

As a beginning, the scope has been restricted to the typical style of Hudum (ᠬᠤᠳᠤᠮ *xudum*; Худам *xudam*) writing system that is contemporarily used.

Eventually a single specification should cover all the major writing systems unified under the Unicode Mongolian encoding (i.e., Hudum, Todo, and Manchu–Sibe), as well as their Sanskrit–Tibetan extensions (Ali Gali letters and writing systems) and historical forms of the writing systems (early Hudum ones, early Todo, Old Manchu, etc.).

Acknowledgements

2 Encoding principles

To a limited extent, Hudum is encoded in a way similar to how other cursive scripts and writing systems are encoded in Unicode. That is to say, instead of directly encoding written units that are already cursively joined to each other, certain underlying abstract units are considered to have the potential of *cursive joining* and are encoded as characters, then the displayed form of a character is contextually determined.

There is also an obligatory contextual variation process (which involves so-called *bowed consonants*) that is beyond cursive joining, but it still within the usual scope of text encoding (cf., obligatory lam–alef ligatures of Arabic).

For more information about these ordinary contextual shaping mechanisms, see Section 3.2, *Cursive joining*, and Section 3.3, *Graphemes*. Figure 1 shows a simple word what only involves these two mechanisms, where abstract characters with cursive joining potential and indefinite shapes contextually interact with each other, and get resolved into certain positional forms, then exhibit some additional interaction between a bowed consonant and its following letter.

Figure 1. Cursive joining and bowed consonants
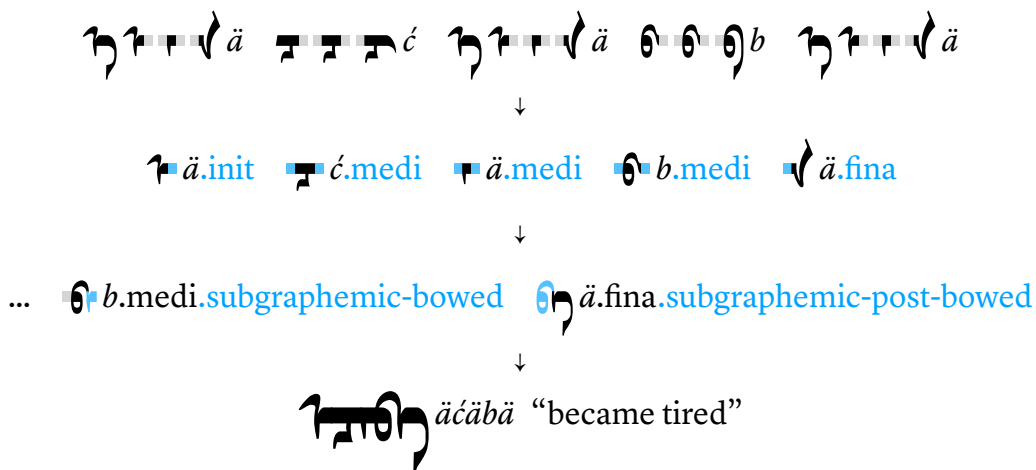
What actually make the Hudum encoding unusual are the additional principles that extend far beyond ordinary text encoding:

- Phonetic letters
- Orthographical shaping
- Grammatical enclitics
- Manual overriding

Specific rules are then derived from these principles for *how text should be encoded* (see Section 3, *Text representation*) and *how encoded text should be displayed* (Section 4, *Text rendering*).

2.1 Phonetic letters

Hudum does not have a well-received system of typical letters (i.e., user-perceived primary units of writing) that is common to other writing systems. Instead, users are

3

accustomed to identify letters on a much more phonetic level where letters are not reliably related to writing, and are thus considered *phonetic letters* in this specification. Although the exact alphabet (i.e., the set and order of letters) varies considerably, Table 1 shows a typical version.

Table 1. Hudum phonetic letters

a	*ä*	*e*	*i*	*o*	*u*	*ö*	*ü*	
⟨glyph⟩	⟨glyph⟩	⟨glyph⟩	⟨glyph⟩	⟨glyph⟩	⟨glyph⟩	⟨glyph⟩	⟨glyph⟩	↖ *Transliteration* ← *Representative written form* ↙ *Phoneme*
a	ə	ə	i	ɔ	ʊ	o	u	

n	*ŋ*	*b*	*p*	*x*	*g*	*m*	*l*	*s*	*ś*	*t*	*d*	*ć*	*j*	*y*	*r*	*w*
⟨glyph⟩	⟨glyph⟩	⟨glyph⟩	⟨glyph⟩	⟨glyph⟩	⟨glyph⟩	⟨glyph⟩	⟨glyph⟩	⟨glyph⟩	⟨glyph⟩	⟨glyph⟩	⟨glyph⟩	⟨glyph⟩	⟨glyph⟩	⟨glyph⟩	⟨glyph⟩	⟨glyph⟩
n	ŋ	p	pʰ	x	k	m	l	s	ʃ	tʰ	t	tʃʰ	tʃ	j	r	w

f	*k*	*c*	*z*	*h*	*ř*	*ł*	*ž*	*č*
⟨glyph⟩	⟨glyph⟩/⟨glyph⟩	⟨glyph⟩	⟨glyph⟩	⟨glyph⟩	⟨glyph⟩	⟨glyph⟩	⟨glyph⟩	⟨glyph⟩
f	kʰ	tsʰ	ts	x	ɻ	ɬ	tʂ	tʂʰ

Notes:

1. In this specification, phonetic letters are referred to with their single-letter transliterations that are always in *italics*.

2. The listed written forms here are not the full set, but are merely what commonly used as written representatives of phonetic letters.

3. Typical phonetic transcriptions of corresponding phonemes in the standard Chakhar phonology are also provided for reference.

4. The seven native vowel letters are often referred to as *vowel one* to *vowel seven*, in order to distinguish the two visually identical pairs, *o/u* and *ö/ü*, in writing.

5. The written form of *k* has a regional difference: the Manchu ka letterform (which also appears to be similar to Ali Gali ga and Todo velar ga) is preferred in China, while the Ali Gali kha letterform is preferred in Mongolia.

6. Letters *e* and *h* are sometimes considered disambiguating variants of *ä* and *x*, respectively.

7. Light gray highlighted ones are loanword letters, thus often excluded from the alphabet. *p* and *w* are early introduced loanword letters, but today often considered to be native letters. *ŋ* is often excluded because it cannot occur on syllable onset positions. *ŋ* and *ł* are often excluded together because they are considered as letter sequences *ng* and *lh*.

8. Letters *ř*, *ž*, and *č* are often excluded because they are considered as disambiguating variants of *r*, *j*, and *ć* for transcribing Chinese syllables each in only one syllable: ⟨glyph⟩, ⟨glyph⟩, and ⟨glyph⟩. *ř* actually has extended usage today for other Chinese r-starting syllables and other loanword [ɻ~ʒ] sounds.

Multi-to-one confusability. The system of phonetic letters is largely based on historical phonemes of the Mongolian language reflected in the conservative orthography, instead of how the under-differentiated Hudum writing system actually works with its limited set of graphemes. Many phonetic letters therefore do not have distinct written forms and can be confusable in writing, but are still identified as distinct letters because they are meant to be the abstract representatives of distinct phonemes.

One-to-multi unpredictability. Furthermore, the yellow highlighted phonetic letters in Table 1 can be written with multiple different graphemes and/or grapheme sequences.

As which written form exactly is used to represent a phonetic letter is determined with a combination of complex predictive rules and arbitrary variations, many phonetic letters have largely unpredictable correspondences with their written forms.

Ordinary letters. For the sake of specification, an ordinary letter is defined as a consonant letter that does not involve a bowed grapheme.

2.2 Orthographical shaping

Typically, orthographical features of a writing system are directly reflected in encoding (e.g., the English writing system requires the final consonant letter of certain words to be doubled when a suffix is joined). However in the Hudum encoding, as the encoded phonetic letters do not directly represent written forms, a number of orthographical rules are utilized to predict most written forms in order to minimize required manual controls in encoding. See Section 3.5, *Additional variation patterns*.
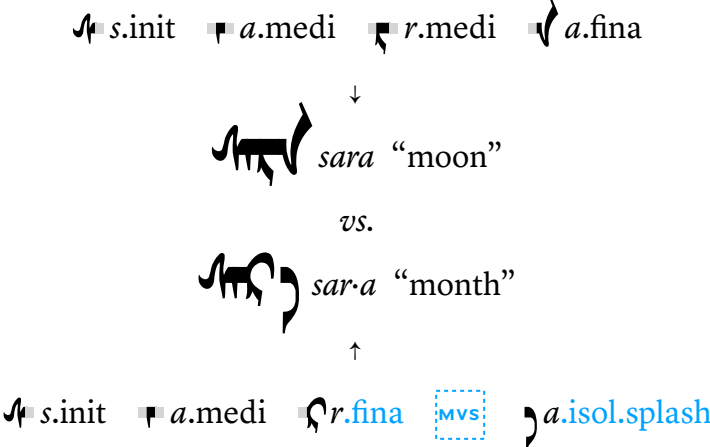
Figure 2. Orthographical shaping

s.init   *a*.medi   *i*.medi.offglide   *n*.fina.coda

↓

sain "good"
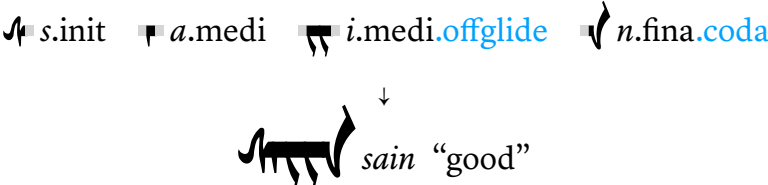
Subjective and incomplete. The one-to-multi ambiguous nature of phonetic letters means the usage of various written forms is ultimately unpredictable from each word's phonetic letter sequence alone. Orthographical rules are also naturally subjective and incomplete because they are scholars' summary of their observation on the writing system. The result is, manual overriding on the predictive orthographical shaping rules is inevitable.

The splash. *The splash* (ćaćulg·a), a non-joining leftward tail form of *a* and *ä* that only appears at the end of certain words, is a common lexical variation. Certain letters take special forms when followed by it. For more information about how the splash is requested with a dedicated format control, see "The splash" in Section 3.5.

Figure 3. The splash

s.init   *a*.medi   *r*.medi   *a*.fina

↓

sara "moon"

vs.

sar·a "month"

↑

s.init   *a*.medi   *r*.fina   MVS   *a*.isol.splash

2.3 Grammatical enclitics

Enclitics are a special group of postposition words that are grammatically comparable to suffixes (and are thus also considered a part of the modified word) but are written separately from their modified word. Only some of them exhibit enclitic-specific orthographical features, thus the scope of such words is grammatical. A special whitespace is used to both connect an enclitic to its preceding word and to trigger the special shaping required by some enclitics. See "Enclitics" in Section 3.5.
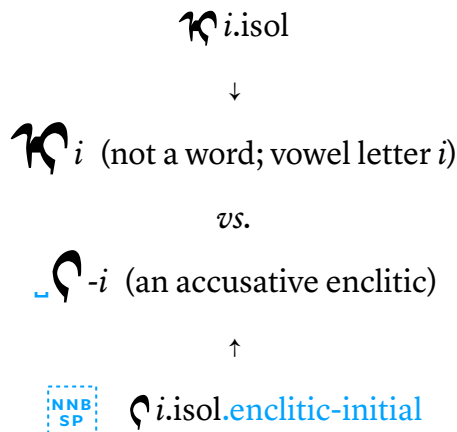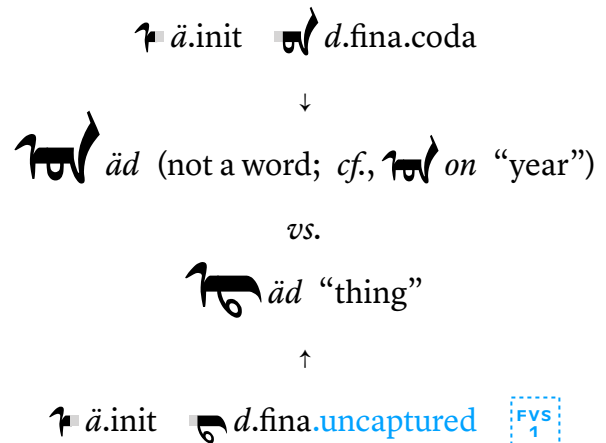
Figure 4. Enclitics

𐎣 *i*.isol

↓

𐎣 *i* (not a word; vowel letter *i*)

vs.

 ̣𐎣 -*i* (an accusative enclitic)

↑

[NNBSP] 𐎣 *i*.isol.enclitic-initial

Figure 5. Manual overriding

𐎡 *ä*.init   𐎣 *d*.fina.coda

↓

𐎣 *äd* (not a word; *cf.*, 𐎣 *on* "year")

vs.

𐎣 *äd* "thing"

↑

𐎡 *ä*.init   𐎣 *d*.fina.uncaptured [FVS 1]

2.4 Manual overriding

When an expected written form is not captured by predictive shaping rules, one of the three last-resort format controls, Free Variation Selectors (FVSes), is used to request the desired written form. Unlike MVS and NNBSP which mark complex lexical or grammatical features then have the expected variations derived, an FVS only affects the base character it is applied to, having no extra effects. See "The shaping step for uncaptured forms" in Section 4.2, *Hudum-specific shaping phase*.

Theoretically, as all written forms on a given cursive position are assigned with an FVS, the predictive rules would be merely syntactical sugar built upon the FVS mechanism for minimizing the number of FVSes used. However, in order to avoid the pollution of unnecessary format controls, FVSes are explicitly rendered invalid wherever the predictive shaping is adequate.

3 Text representation

This section specifies how Hudum text should be encoded. The following aspects are introduced:

- Required characters
- Cursive joining mechanism
- Graphemes as a concrete low-level abstraction for describing written forms
- Phonetic letters and their written forms
- Variation patterns considered for reducing phonetic letters to written forms

Text representation of a word is thus determined by:

- Identifying the underlying sequence of phonetic letters either by consulting a dictionary, or by analyzing possible written forms and variation patterns.
- Apply FVSes wherever phonetic letters are not reduced to the desired written forms by predictive rules.

3.1 Minimal character set

A typical Hudum implementation requires the characters shown in Table 2, including Mongolian-specific ones and characters that are shared with other scripts.

Table 2. Required characters

Script	Type of characters	Character or range of characters	Note
General	Space (1)	U+0020 SPACE	
	Format controls (3)	U+200C ZERO WIDTH NON-JOINER	
		U+200D ZERO WIDTH JOINER	
		U+202F NARROW NO-BREAK SPACE	
	Misc	·()«»‹›0123456789 *etc.*	
Mongolian	Punctuation (6)	U+1800 MONGOLIAN BIRGA..	BIRGA and FOUR DOTS have limited modern usage
		U+1805 MONGOLIAN FOUR DOTS	
	Format controls (5)	U+180A MONGOLIAN NIRUGU..	
		U+180E MONGOLIAN VOWEL SEPARATOR	
	Digits (10)	U+1810 MONGOLIAN DIGIT ZERO..	Limited modern usage
		U+1819 MONGOLIAN DIGIT NINE	
	Phonetic letters (35)	U+1820 MONGOLIAN LETTER A..	
		U+1842 MONGOLIAN LETTER CHI	

Yellow highlighted characters are involved in the complex shaping of Hudum. All the Mongolian-specific characters listed in the table above are encoded in the main Mongolian block (U+1800..U+18AF). The other block, Mongolian Supplement (U+11660..U+1167F), currently only have 13 characters for variants of the *birga* sign (U+1800 MONGOLIAN BIRGA), which are not typically used in day-to-day text.

For information about phonetic letters, see Section 3.3, *Graphemes* and Section 3.4, *Encoded phonetic letters*. The following subsections introduce other types of characters.

Format controls

Zero Width Non-Joiner (ZWNJ), Zero Width Joiner (ZWJ), and nirugu. U+200C ZERO WIDTH NON-JOINER and U+200D ZERO WIDTH JOINER are Unicode's standard cursive joining controls. Note that ZWJ also breaks interaction (such as ligation) between consecutive two letters as it is treated as an invisible letter. U+180A MONGOLIAN NIRUGU is a Mongolian-specific modifier letter that behaves exactly like ZWJ but is visible as a piece of stem stroke. See Section 3.2, *Cursive joining*.

In particular, for the Hudum-specific shaping steps, ZWJ and nirugu act like medial forms of an ordinary letter (defined in Section 2.1), while ZWNJ acts like an ordinary space (U+0020).

Nirugu instead of ZWJ is recommended for average users' need of causing joining in day-to-day text. A common use case is terminating a patronymic abbreviation, which the initial syllable body (i.e., an optional onset plus the first vowel) or merely the initial consonant letter of one's father's name.

Vowel Separator (MVS) and Narrow No-Break Space (NNBSP). MVS is a Mongolian-specific format control for requesting the splash variation. It is transcribed as "·" (a middle dot). See "The splash" in Section 3.5, *Additional variation patterns*. NNBSP is a Mongolian-specific whitespace and format control for marking and shaping enclitics, and note that it is also used as a general whitespace by other scripts. See "Enclitics" in Section 3.5.

In terms of cursive joining behavior, MVS and NNBSP are both non-joining inline characters, like an ordinary space.

Free Variation Selectors (FVSes). Mongolian-specific format controls. As combining marks, they are applied to certain characters for requesting the forms not captured by predictive shaping rules. See "Uncaptured forms" in Section 3.5.

Punctuation and miscellaneous characters

Mongolian-specific punctuation characters do not yet have well-defined spacing behavior. Fore example, it is inconsistent in implementations if U+1802 MONGOLIAN COMMA, U+1803 MONGOLIAN FULL STOP, and other punctuation characters (as they tend to have significant and balanced spacing on both sides) are both preceded and followed by whitespace characters (e.g., U+0020 SPACE) or just have the preceding spacing as part of their glyphs.

The choices of non-Mongolian-specific punctuation characters have been heavily influenced by what characters are used in Chinese text, and are often not the ideal choices. Certain CJK punctuation characters are indeed beneficial because of their upright, non-rotated appearance in vertical text, but it is a problem to actually clarify how spacing and positioning of these fullwidth characters should be adapted for Mongolian usage.

3.2 Cursive joining

Written forms exhibit the cursive joining mechanism. Both sides of a written form can be either joined to a neighboring written form or not, exhibiting up to four different statuses. Or, abstractly speaking, each written form is on one of four cursive positions:

- *isolated* (abbreviated as *isol*): not joined above, not joined below
- *initial* (*init*): not joined above, joined below
- *medial* (*medi*): joined above, joined below
- *final* (*fina*): joined above, not joined below

Cursive positions are irrelevant to word boundaries, although they are usually consistent with word-wise positions in Hudum because cursive joining breaks inside a word are limited in the writing system.

3.3 Graphemes

Before examining the encoded phonetic letters, encoding-independent *graphemes* are defined in Table 3. All written forms in this specification are analyzed with and formally referred to as sequences of graphemes on certain cursive positions, for the sake of clarity and accuracy.

Graphemes are named after their historical origin letters, either Aramaic (beth, waw, etc.) or Tibetan (only zha, ha, tta, and ttha), and derived graphemes are distinguished with a dot-separated suffix. In this specification, graphemes are assigned single-letter transliterations that are always in SMALL-CAPS. Phonetic letters *ŋ* and *ł*, as well as certain written forms of phonetic letters, are sequences of graphemes, thus are not covered in the table.

Table 3. Graphemes

Grapheme		Positional forms: .isol, .init, .medi, .fina				Subgraphemic variants		Represented phonetic letters				Note
alephnun	A	𐴰	𐴰	𐴰	𐴰		• 𐴰	*a, ä*	*∅, a, ä*	*∅, a, ä, n*	*a, ä, n*	𐴰𐴰
—.splash	Á	𐴰						*a, ä*				
—.na	N		𐴰	𐴰	𐴰		•		*n*	*n*	*n*	*see* alephnun
beth	W		𐴰	𐴰	𐴰		• •		*w*	*w, e*	*w, e*	

9

Name	Code	Forms	Translit.			Notes
gimelheth	X		*x*	*x, g*	*x, g*	
—.ga	Ğ		*g*	*g*	*g*	
waw	U		*u, ü*	*u, ü*	*o, u, ö, ü, w*	*o, u, ö, ü, w*
—.o	O				*o, u*	
—.ü	Ü				*ö, ü*	
yodh	I		*i, j*	*i, j, y*	*i, y*	*i, y*
—.ya	Y		*y*	*y*		
kaph	G		*x, g*	*x, g*	*g*	
—.ka	K		*k*	*k*	*k*	Preferred in China / Preferred in Mongolia
lamedh	D		*d*	*t, d*	*d*	
mem	M		*m*	*m*	*m*	; *see* alephnun
samekhshim	S		*s*	*s*	*s*	
—.śa	Ś		*ś*	*ś*	*ś*	
pe	B		*b*	*b*	*b*	
—.pa	P		*p*	*p*	*p*	
—.fa	F		*f*	*f*	*f*	
sadhe	Ć		*ć*	*ć*	*ć*	
—.ja	J			*j*	*j*	
—.ca	C		*c*	*c*	*c*	
—.za	Z		*z*	*z*	*z*	
resh	R		*r*	*r*	*r*	
taw	T		*t, d*	*t*	*t*	
—.da	Đ			*d*	*d*	
lesh	L		*l*	*l*	*l*	*see* alephnun
zha	Ř		*ř*	*ř*	*ř*	*see* alephnun
ha	H		*ḥ*	*ḥ*	*ḥ*	*see* alephnun
tta	Ž		*ž*			
ttha	Č		*č*			

Positional forms. Positional forms of graphemes are denoted with an additional suffix. True isolated forms are absent for most graphemes, and many graphemes are not attested on all of the rest three positions either. An explicit cursive joining break is transliterated as "|", and an explicit joining as "-".

Regional and stylistic variants. The two regional variants of *k* are considered a single grapheme in this specification. Certain commonly used stylistic variants of graphemes exhibit significant structural differences:

- All graphemes that involve a crown (*titim*) on certain positional forms (i.e., alephnun, alephnun.na, mem, lesh, zha, and ha) have a historical preferred, non-crown variant.

- Graphemes mem and pe both have a final variant that was historically preferred. Note that the Todo writing system still prefers these two variants and consequently has disunified characters (U+184B ᴍᴏɴɢᴏʟɪᴀɴ ʟᴇᴛᴛᴇʀ ᴛᴏᴅᴏ ʙᴀ and U+184F ᴍᴏɴɢᴏʟɪᴀɴ ʟᴇᴛᴛᴇʀ ᴛᴏᴅᴏ ᴍᴀ).

Subgraphemic variation involving bowed graphemes

Bowed-looking grapheme groups kaph and pe, namely, kaph, kaph.ka, pe, pe.pa, and pe.fa (on their initial and medial positions) cause certain graphemes (on their medial and final positions) to join perpendicularly to the stem. Attested perpendicularly joinable medial and final forms are marked in the Table 3 with either a bullet or the perpendicularly joined form (if significant structural change is observed).

Note the commonly referred "leftward tail" (or "feminine tail") of alephnun is considered merely the perpendicularly joined form of alephnun.fina. Grapheme waw lack a perpendicularly joined final form, as the forms it would orthographically take (waw.o.fina or, marginally, waw.ü.fina) have been disunified.

3.4 Encoded phonetic letters

Phonetic letters are encoded as abstract characters that may be used to represent any written forms of a given phonetic letter. Attested written forms are defined as positional forms of grapheme sequences, as shown in Table 4.

Yellow highlighted phonetic letters exhibit additional written forms (also highlighted in yellow) besides a single set of positional forms. For such complex characters, their default positional forms are specified to be the most reasonable forms (which are stray forms for consonant letters that are affected by syllable structure) when cursive positions are caused by medial forms of an abstract, ordinary letter.

Table 4. Encoded phonetic letters and their written forms

Phonetic letter & character	Written forms: default positional forms & additional forms								Note
a U+1820 A	ᠠ	ᠠ	ᠠ	ᠠ	AA	AA	A	A	• A.isol/init—*enclitics* • AA.medi—*compound words* • Á.isol—*the splash*
	ᠠ	ᠠ	ᠠ		A	A	AA		
	ᠠ				Á				

	Unicode	isol	init	medi	fina	Notes
ä	U+1821 E	A	A	A	A	• Á.isol—*the splash*
		Á				
i	U+1822 I	AI	AI	I	I	• I.isol/init—*enclitics*
		I	I	AI		• AI.medi—*compound words*
				II		• II.medi—*offglide*
o	U+1823 O	AO	AU	U	U	• AU.medi—*compound words*
				AU	O	• O.fina—*initial-body, post-bowed, disambiguating*
u	U+1824 U	AO	AU	U	U	• U.isol/init—*enclitics*
		U	U	AU	O	• AU.medi—*compound words*
						• O.fina—*initial-body, post-bowed*
ö	U+1825 OE	AÜ	AUI	U	U	
				AUI	Ü	• AUI.medi—*compound words*
				UI	O	• UI.medi, Ü.fina—*initial-body*
						• O.fina—*post-bowed*
ü	U+1826 UE	AÜ	AUI	U	U	• U.isol/init—*enclitics*
		U	U	AUI	Ü	• AUI.medi—*compound words*
		AU		UI	O	• UI.medi, Ü.fina—*initial-body, disambiguating*
						• AU.isol—*lexical*
						• O.fina—*post-bowed*
e	U+1827 EE	AW	AW	W	W	Cross-grapheme simple letter
n	U+1828 NA		N	N	N	• N.fina—*pre-splash*
				A	A	• A.medi—*coda*
ŋ	U+1829 ANG		—	AG	AG	.init unattested
b	U+182A BA		B	B	B	
p	U+182B PA		P	P	P	
x	U+182C QA		G	G	—	Default .fina unattested
			X	X	X	• G.init/medi—*feminine*
						• X.fina—*pre-splash*
g	U+182D GA		G	G	G	• G.init/medi/fina—*feminine*
			Ğ	Ğ	X	• X.medi—*coda*
				X	Ğ	• Ğ.fina—*pre-splash*
m	U+182E MA		M	M	M	
l	U+182F LA		L	L	L	
s	U+1830 SA		S	S	S	

		init	medi	fina				Notes
ś	U+1831 SHA				Ś	Ś	Ś	
t	U+1832 TA				T	T	T	• T.medi—*disambiguating*
						D		
d	U+1833 DA				T	D	D	• D.init/fina—*disambiguating*
					D	Đ	Đ	• Đ.medi—*coda*
ć	U+1834 CHA				Ć	Ć	Ć	
j	U+1835 JA				I	J	J	• I.isol—*pre-splash*
				I				
y	U+1836 YA			⚠	Y	Y	—	Default .fina unattested • I.init/medi—*enclitics*
					I	I	I	• I.fina—*pre-splash*
r	U+1837 RA				R	R	R	
w	U+1838 WA				W	W	W	• U.fina—*pre-splash*
							U	
f	U+1839 FA				F	F	F	
k	U+183A KA				K	K	K	Preferred in China
	U+183B KHA							Preferred in Mongolia
c	U+183C TSA				C	C	C	
z	U+183D ZA				Z	Z	Z	
b	U+183E HAA				H	H	H	
ř	U+183F ZRA				Ř	Ř	Ř	
ł	U+1840 LHA				LH	LH		.fina unattested
ž	U+1841 ZHI				Ž			.medi/fina unattested
č	U+1842 CHI				Č			.medi/fina unattested

Additional notes:

1. The grapheme sequence II.medi (in written forms of the offglide *i, öi/üi*, etc.) has a stylistic variant that looks similar to AI.medi, particularly in handwritten styles. The grapheme I.init of *y* also has a stylistic variant that looks similar to A.init, particularly in handwritten enclitic *yin* (which consequently looks similar to an ordinary *in*). Text representation should not be affected by such stylistic variations.

2. The two regional variants of *k* have disunified encodings, U+183A KA and U+183B KHA, and typically only one of them is used in a text according to the preferred regional style.

These characters are more abstract than what are commonly encoded for other cursive joining scripts, as not only do they involve positional forms for cursive joining, but some

of them also involve multiple possible written forms on a single cursive position and thus require examining additional variation patterns for determining a character. By comparison, cross-grapheme variations are also involved in the Arabic encoding but they are always constrained within the cursive joining mechanism.

Because characters are encoded on a phonetic basis, many written forms are shared across multiple characters (some character pairs even have fully overlapping sets of written forms). Therefore, identifying a word's underlying sequence of phonetic letters from its written form is highly subjective, and requires knowledge such as orthography, grammar, and ultimately a dictionary, which are far beyond the basic script behavior. Hudum text representation is thus even more complicated than that of Indic scripts, which also involve multiple variation mechanisms but are generally contextually predictable and self explanatory. Faced with the overly complicated variation mechanisms of Hudum, users often turn to piece together a written word graphically with phonetically arbitrary characters.

3.5 Additional variation patterns

A number of commonly recognized variation patterns are considered for the Hudum text presentation. Some patterns correspond to predictive shaping rules, while the rest are not executed in fonts.

Uncaptured variants. Forms not captured by the predictive shaping rules introduced below are requested with FVSes. See "The shaping step for uncaptured forms" in Section 4.2, *Hudum-specific shaping phase*.

Words, enclitics, stems, etc. A whitespace-separated word (a morphological word) may be either an ordinary word (which can be modified by one or more enclitics as a host word) or an enclitic. Enclitics are prosodically part of their host words, forming a single prosodic word. Ordinary words contain one or more word stems, and may receive one or more suffixes.

Graphemic variation after bowed graphemes

In addition to the subgraphemic variation introduced in Section 3.3, *Graphemes*, bowed graphemes also cause a following U.fina vowel form to change to O.fina. Letter *ŋ*, although is written with a bowed grapheme end, is not attested for interacting with a following vowel like a bowed consonant (*x, g, k, b, p, f*).

Syllabic variations

A written *syllable* in Hudum has the structure of C?V+C? (i.e., an optional *onset* consonant letter, one or more vowel letters forming the *nucleus*, and an optional *coda* consonant letter). The leading C?V part (a syllable excluding the trailing V*C?; i.e., the First Syllabary structure) has a special place in Hudum orthography patterns and is termed a *body*.

A consonant letter between two vowel letters belongs to the latter syllable, while a consonant letter excluded by the syllable structure is considered a *stray*. The stray forms (if distinct) are used as default positional forms.

Boundaries of written syllable always occur at morphological word boundaries (i.e., whitespaces) and between adjacent word stems (but not required between a word stem and a suffix). Note especially that syllable boundaries do not occur at cursive joining breaks inside a word, which are typically observed only before a splash.

Onset and coda forms of consonants. Certain consonants are written with different forms for their syllable onsets and codas roles. The analysis for onset–coda contrast is restricted to the three complex consonants included in the traditional set of so-called *pad* (🖎🖎 *däbisxär*) consonants (yellow highlighted in Table 5), namely, *n*, *g*, *d*, as well as the onset-only *x*, although there are a couple of more consonants (*t*, *ł*, *ž*, and *č*) exhibit distribution patterns affected by their syllabic roles. .... *Predictive*

Table 5. The traditional pads

Type & phonetic letters		Written forms				Note
Soft pads	*n*			A	A	
	m			M	M	
	l			L	L	
	ŋ			AG	AG	
Hard pads	*b*			B	B	
	g			X	X	Masculine and feminine forms.
				G	G	
	r			R	R	
	s			S	S	
	d			Đ	Đ	
Vocalic pads	*i*			II	I	Considered *y* or *yi*.medi / *y*.fina in alternative theories.
	u/ü			U	U	Considered *w* in alternative theories.

Offglide form of *i*. A vowel may be followed by an *offglide* vowel *i*, *u*, or *ü*, forming a diphthong. Offglides are not analyzed as coda forms of *y* and *w*. A medial offglide *i* (red highlighted in Table 5) takes its special offglide form ɪɪ unless the preceding vowel's written form already ends with a grapheme ɪ. .... *Predictive*

Variations in stem-led scopes

A special *stem-led scope* is a word stem together with its following zero or more suffixes and enclitics. Certain variation patterns appears to be affected by boundaries of such scopes, however these scopes and boundaries are obscure to native users and are thus not fully utilized in encoding.

Onset placeholder. The syllable onset is required at the beginning of a stem-led scope, and a historical consonant letter aleph is used as the onset placeholder when an initial consonant is absent. This onset placeholder is generally considered by native users to be a part of the nucleus vowel letter's written form:

- It is a part of the default isolated and initial forms.

- It is considered to be a part of uncaptured forms on medial positions, which typically occur at beginning of non-first stems in a compound word. Such a form is unattested for the loanword vowel letter *e*.

- It is unattested on final positions although theoretically possible if the last stem of a compound word is a single vowel letter.

Initial-body forms of rounded vowels. Vowel letters *o*, *u*, *ö*, and *ü* are disambiguated when they are in the initial (i.e., not preceded by any other letter) body (CʔV) of a stem-led scope:

- The default isolated and initial forms of *o*, *u*, *ö*, and *ü* are already initial-body forms.

- For the first stem in a word, initial-body medial and final forms (initial-body medial forms are only applicable to *ö* and *ü*) are predictable when one of these vowels follows an initial consonant. .... *Predictive*

- For the non-first stem in a compound word, medial and final initial-body forms are uncaptured forms.

Note that loanwords exhibit two major types of violation: Chinese loanwords tend to write a final u sound as ᴜ.fina even when it is in the initial body, and as ᴜ̈.fina after a bowed consonant; o-type sounds are generally normalized to other vowels. Other loanwords, where o sounds and u sounds are contrasted, tend to always write o as ᴜ.medi/ᴏ.fina and u as ᴜɪ.medi/ᴜ̈.fina.

Gender-specific forms of x *and* g

Consonant letters *x* and *g* both have contrasted masculine and feminine classes of written forms, which have a distribution related to vowel genders (i.e., vowel harmony classes). Vowels *a*, *o*, and *u* are masculine; *ä*, *e*, *ö*, and *ü* are feminine; *i* is neuter.

In principle, a stem-led scope is internally gender harmonious, and thus should have a determinate gender value. However, acquiring a stem-led scope's actual harmony status is non-trivial as a stem can easily have mixed genders (e.g., loanwords), thus the gender-specific variation patterns of *x* and *g* are broken down into smaller patterns:

- A stray *g* takes the feminine form, which is handled by the default positional forms.

- An onset *x* or *g* agrees with its following vowel's gender (masculine or feminine); the neuter vowel *i* is treated like a feminine vowel for the onset. .... *Predictive*

- A coda *g* agrees with its preceding vowel's gender (masculine or feminine). A coda *g* that follows *i* takes the feminine form by default, and takes the masculine form if it remotely follows a masculine vowel and there is no feminine vowels in between. Note that this specification only considers single-directional, forward gender propagation for *ig*. .... *Predictive*

- Other situations are considered uncaptured forms.

The splash

This non-joining grapheme of *a* and *ä* is usually observed after *n*, masculine *x*, masculine *g*, *m*, *l*, *y*, *r*, *w*, and only occasionally after *j*. The format control MVS is used to break cursive joining between the preceding consonant and *a/ä*, as well as to request the special forms required for certain consonants:

- Letters *n*, *x*, and *g* take what appear to be their special, onset final forms. ....*Predictive with MVS*
- Letters *y*, *w*, and *j* take their under-differentiated forms (i.e., yodh, waw, and yodh, respectively). ....*Predictive with MVS*
- Writing in the form of splash is orthographically mandatory for *a* to appear after a word-medial *x/g*.
- Note that *x* and *y* normally do not have final forms as they are not used as codas.

Enclitics

The special whitespace between an enclitic and its preceding word (either the modified word or a preceding enclitic) is sometimes considered an internal gap in a grammatical word, and thus is sometimes preferred to be non-line-breaking, non-word-breaking, and narrower than an ordinary space. Typical enclitics exhibit at least one of the following variations:

- Absence of onset placeholder and initial-body variation (because the beginning of an enclitic it is not the beginning of a stem). ....*Predictive with NNBSP*
- An initial *d* takes its disambiguating lamedh form. ....*Predictive with NNBSP*
- The under-differentiated, historical form of *y*. ....*Predictive with NNBSP*
- Letters *x* and *g* show gender harmony with the preceding word. The gender features are only observable on these two letters because the first feature has neutralized gender-distinguishing features of all vowel letters.

NNBSP is used both for representing this whitespace and for requesting the special variations shown in enclitics. As the set of enclitics and usage of NNBSP are decided grammatical, an enclitic may or may out exhibit special variations but is still encoded with a preceding NNBSP. See Appendix A for a reference list and comparison.

4 Text rendering

This section specifies how Hudum text should be displayed, once properly encoded according to specification of the previous section.

4.1 Minimal shaping process

The shaping process of Hudum is based on the well-implemented technology foundation for general scripts and cursive scripts, while an additional phase of Hudum-specific shaping steps is inserted into the ordinary shaping process required by cursive scripts. The minimal shaping process consists of a number of steps as shown in the table below.

Table 6. Minimal shaping process

Shaping phase	Shaping step	
Iᴀ. General	• **Basic character-to-glyph mapping**	
IIᴀ. Cursive script	• **Initiation of positional forms**	
III. Hudum-specific *Reduction of phonetic letters to graphemes*	*Phonetic:*	**1. Onset and coda**
		2. Gender-specific
		3. MVS-involving
		4. NNBSP-involving
	Graphemic:	**5. Offglide**
		6. Graphemic post-bowed
	Uncaptured:	**7. FVS-selected**
IIв. Cursive script (continued) *Subgraphemic variations*	• **Variation involving bowed graphemes**	
	• **Cleanup of format controls**	
	• Optional treatments	
Iв. General (continued) *Typography*	• **Vertical forms of punctuation marks**	
	• Optional treatments	

See Section 4.2, *Hudum-specific shaping phase* for details about the phase III.

General shaping phases

These are the basic mechanisms in fonts, applicable for all scripts. Basic character-to-glyph mapping (phase Iᴀ) is typically controlled by the TrueType/OpenType table "cmap". The Unicode representative glyphs may be used here as the default glyph mappings for phonetic letters, however these representative glyphs are not actually kept

in the final rendering in typical implementations. Vertical forms of punctuation marks (phase IB) are critical to proper typesetting of Hudum text, but are not part of the complex shaping between letters and format controls.

Note that the two disunified regional variants of *k* are character-level variants, and are not to be interchangeable in rendering (e.g., a font designed for users in China must not render U+183B KHA with the look of U+183A KA, even with a tagged locale of "China"). Fonts are recommended to support both to cover users' various preferences.

Cursive script shaping phases

On top of the general shaping mechanisms, complex scripts require additional shaping phases to be inserted after the basic character-to-glyph mapping and before typographical treatments. In particular, cursive scripts all undergo the cursive joining mechanism.

The originally mapped glyphs from the last phase are converted to default positional forms in phase IIA. Although these default positional forms are not necessarily kept till the end of shaping process, the cursive positions are immutable once initiated, for the sake of simplicity in later shaping phases.

For the exact algorithm used here for initiating positional forms, see "Arabic Cursive Joining" in Section 9.2, *The Unicode Standard, Version 12.0—Core Specification*.

Unattested default positional forms, especially default isolated forms of consonants, are recommended to be explicitly marked invalid in rendering. It is recommended to implement bowed-grapheme variation with contextual glyph variants, although many fonts use ligatures.

4.2 Hudum-specific shaping phase

The phase III consists a series of steps for Hudum-specific shaping requirements, and inside each step there may be more than one set of non-overlapping rules, each for a different group of letters.

Phonetic and graphemic shaping steps

In the phonetic and graphemic shaping steps (1–6), certain letters are analyzed as subjects and a set of contextual rules determine if the subject letters are in one of the following 13 conditions:

```
onset
onset-masculine
onset-feminine
coda
coda-masculine
coda-feminine
pre-splash
splash
enclitic-initial
```

```
enclitic-lexical
initial-body
offglide
graphemic-post-bowed
```

Contextual rules for each step and condition are specified in the table below. Letter categories are specified in Table 8. The execution order of shaping steps is critical, as they have dependencies of previous ones, and a later decided condition overwrites an earlier one. For a reference list of enclitics that are commonly recognized as NNBSP-applicable (step 4), see Appendix A.

Table 7. Hudum shaping conditions: phonetic and graphemic

Shaping step	Subject letters	Rules	Resulted condition
1. Onset and coda	*n/g/d*	**if** precedes a vowel:	`onset`
		else if follows a vowel:	`coda`
	x	**if** precedes a vowel:	`onset`
2. Gender-specific	onset *x/g*	**if** precedes a masculine vowel:	`onset-masculine`
		else:	`onset-feminine`
	coda *g*	**if** follows a masculine vowel:	`coda-masculine`
		else if follows a feminine vowel:	`coda-feminine`
		else if remotely follows a masculine vowel without a blocking feminine vowel:	`coda-masculine`
		else:	`coda-feminine`
3. MVS-involving	*n/j/y/w*	**if** precedes MVS:	`pre-splash`
	x/g	**if** precedes MVS that does not precede *ä*:	`pre-splash`
	a/ä	**if** follows MVS:	`splash`
4. NNBSP-involving	*a/i/u/ü/d*	**if** follows NNBSP:	`enclitic-initial`
	o/u/ö/ü	**if** follows an initial consonant that does not follow NNBSP:	`initial-body`
	y	**if** is in a word *yin/yi/iyar/iyär/iyan/iyän* that follows NNBSP:	`enclitic-lexical`
5. Offglide	*i*	**if** follows a vowel written form that does not end with grapheme ɪ:	`offglide`
6. Graphemic post-bowed	*o/u/ö/ü*	**if** is in the written form of ᴜ and follows a bowed consonant written form that ends with bowed grapheme ɢ/ᴋ/ʙ/ᴘ/ꜰ:	`graphemic-post-bowed`

Table 8. Letter categories

Letter category		Letters	Note
vowel	masculine vowel	a, o, u	
	feminine vowel	ä, e, ö, ü	
	neuter vowel	i	
consonant	bowed consonant	x, g, k, b, p, f	· ŋ is not included. · x and g exhibit the bowed grapheme conditionally.
	ordinary consonant	ZWJ and nirugu, as well as any letter not categorized as a vowel or a bowed consonant.	

When a specific positional written form of a letter is decided to be in a certain condition, it is converted to the specified conditional form in Table 9. Unspecified positions in the table do not affect the subject written form.

Data file. These conditional forms can be easily recorded in a data file for the Unicode Character Database (UCD).

Table 9. Conditional forms of phonetic letters

Phonetic letter & character	Condition	Conditional forms				Note
a U+1820 A	splash	⟨glyph⟩		Á		
	enclitic-initial	⟨glyph⟩ ⟨glyph⟩		A	A	
ä U+1821 E	splash	⟨glyph⟩		Á		
i U+1822 I	enclitic-initial	⟨glyph⟩ ⟨glyph⟩		I	I	
	offglide		⟨glyph⟩		II	
o U+1823 O	initial-body		⟨glyph⟩		O	
	graphemic-post-bowed		⟨glyph⟩		O	
u U+1824 U	enclitic-initial	⟨glyph⟩ ⟨glyph⟩		U	U	
	initial-body		⟨glyph⟩		O	
	graphemic-post-bowed		⟨glyph⟩		O	
ö U+1825 OE	initial-body		⟨glyph⟩ ⟨glyph⟩	UI	Ü	
	graphemic-post-bowed		⟨glyph⟩		O	
ü U+1826 UE	enclitic-initial	⟨glyph⟩ ⟨glyph⟩		U	U	
	initial-body		⟨glyph⟩ ⟨glyph⟩	UI	Ü	

		graphemic-post-bowed				◌				O	
n	U+1828 NA	onset		◌	◌			N	N		Onset and pre-splash forms match default positional forms
		coda			◌	◌			A	A	
		pre-splash				◌				N	
x	U+182C QA	onset-masculine		◌	◌			X	X		Onset + feminine forms match default positional forms
		onset-feminine		◌	◌			G	G		
		pre-splash				◌				X	
g	U+182D GA	onset-masculine		◌	◌			Ğ	Ğ		Onset/coda + feminine forms match default positional forms
		onset-feminine		◌	◌			G	G		
		coda-masculine			◌	◌			X	X	
		coda-feminine			◌	◌			G	G	
		pre-splash				◌				Ğ	
d	U+1833 DA	onset		◌	◌			T	D		Onset forms match default positional forms
		coda			◌	◌			Đ	Đ	
		enclitic-initial		◌				D			
j	U+1835 JA	pre-splash	◌				I				.fina?
y	U+1836 YA	pre-splash			◌				I		.isol?
		enclitic-lexical		◌	◌			I	I		
w	U+1838 WA	pre-splash				◌				U	.isol?

The shaping step for uncaptured forms

The step 7 does not involve contextual effects, as an FVS only affects the base character it is applied to. FVSes are only used to request desired written forms that are not captured by all the previous predictive shaping rules in steps 1–6. [*To be elaborated in a later revision.*]

5 References

Bao Yuzhu/宝玉柱 and Menghebaoyin/孟和宝音. 2011. 现代蒙古语正蓝旗土语音系研究. 北京: 民族出版社.

Mongolian Research Institute, School of Mongolian Studies, Inner Mongolia University / 内蒙古大学蒙古学研究院蒙古语文研究所. 1999. 蒙汉词典. 呼和浩特: 内蒙古大学出版社.

Nicholas Poppe. 1954. *Grammar of Written Mongolian*. Wiesbaden: Harrassowitz Verlag.

Qinggeertai/清格尔泰. 1991. 蒙古语语法. 呼和浩特: 内蒙古人民出版社.

Quejingzhabu/确精扎布. 2000. 蒙古文编码. 呼和浩特: 内蒙古大学出版社.

Rita Kullmann and D. Tserenpil. 1996. *Mongolian Grammar*. Hong Kong: Jensco Ltd.

Documents

Myatav Erdenechimeg, Richard Moore, and Yumbayar Namsrai. 1999. UNU/IIST Report No. 170, *Traditional Mongolian Script in the ISO/IEC 10646 and Unicode Standards*. Macau: UNU/IIST. Accessed from http://babelstone.co.uk/Mongolian/Report170.pdf, http://babelstone.co.uk/Mongolian/Report170A.pdf, and http://babelstone.co.uk/Mongolian/Report170B.pdf.

Badral Sanlig and Munkh-Uchral Enkhtur. L2/18-293, *Solution for NNBSP Issues*. UTC Document Registry.

Greg Eck, Andrew West, Badral Sanlig, Siqinbilige, and Ou Rileke. L2/17-036, *Encode Mongolian Suffix Connector (U+180F) To Replace Narrow Non-Breaking Space (U+202F)*. UTC Document Registry.

Shen Yilei/沈逸磊. L2/17-332, *Positional Mismatches in Mongolian Encoding*. UTC Document Registry.

Standards

GB/T 25914-2010, 信息技术 传统蒙古文名义字符、变形显现字符和控制字符使用规则 / *Information technology—Traditional Mongolian nominal characters, presentation characters and use rules of controling characters*. 北京: 中华人民共和国国家质量监督检验检疫总局 and 中国国家标准化管理委员会. Accessed from http://www.gb688.cn/bzgk/gb/newGbInfo?hcno=62808E0BCB8246A287CFD9CF795ECF94.

Liang Jinbao/梁金宝. 2018. MGC/01-01 (version 1.0.2), 信息技术 传统蒙古文名义字符到变形显现 字符的转换补充规则 / *Information technology—The Transferring Rules of Traditional Mongolian Nominal Form to Variant Form*. 呼和浩特: 内蒙古自治区民族事务委员会 and 内蒙古大学. http://nmgmzw.gov.cn/nmmwh/gsgg/201808/5938899e00fc43aebd189acaa5c6f9e4.shtml.

MNS 4932: 2000, *Монголжин бичгийн кодыг хэрэглэх дүрэм / Use of Mongolian Character Encoding*. Улаанбаатар: Стандартчилал, хэмжилзүйн үндэсний төв. Accessed on 7 September 2018 from http://estandard.gov.mn/index.php?module=standart&cmd=standart_desc&sid=7813.

The Unicode Standard, Version 12.0. Mountain View: The Unicode Consortium. http://unicode.org/versions/Unicode12.0.0/.

Appendix A

Enclitic lists in the following resources are compared in Table 10:

- *The Users' Convention*, abbreviated as "UC" in the table, published in the form of UNU/IIST Report No. 170 (Myatav Erdenechimeg et al. 1999) and MNS 4932: 2000.
- GB/T 25914–2010, "GB/T".
- L2/17-036 (Greg Eck et al. 2017), "17-036".
- L2/18-293 (Badral Sanlig and Munkh-Uchral Enkhtur 2018), "18-293". The listed enclitics in L2/18-293 are marked with white bullets in the table, because their proposed text representations are not explicitly given, and the document has inconsistencies in its Latin transliteration column.

Yellow and red highlighted are character sequences that would rely on NNBSP for achieving their special written forms. Red ones, in particular, involve lexical variations that are not predictable even if recognized as enclitics.

Table 10. Enclitics commonly recognized as NNBSP-applicable

Written form		Text representation following an NNBSP		UC	GB/T	17-036	18-293	Note
	IIN	*yin*		•	•	•	○	
	UA	*un*	*ün*	•	•	•	○	Genitive
	U	*u*	*ü*	•	•	•	○	
	DU	*du*	*dü*	•	•	•	○	
	TU	*tu*	*tü*	•	•	•	○	
	DUR	*dur*	*dür*	•	•	•	○	Dative
	TUR	*tur*	*tür*	•	•	•	○	
	Á	*·a*	*·ä*		*a, ä*	*a, ä*	○	
	II	*yi*		•	•	•	○	Accusative
	I	*i*		•	•	•	○	
	AĆA	*aća*	*äćä*	•	•	•	○	Ablative
	IIAR	*iyar*	*iyär*	•	•	•	○	Instrumental
	BAR	*bar*	*bär*	•	•	•	○	
	TAI	*tai*	*täi*	•	•	•	○	
	LUĞ|Á	*lug·a*		•	•	•	○	Comitative
	LUGA		*lügä*	•	•	•	○	

Glyph	Translit	Reading 1	Reading 2					Category
	BAA	*ban*	*bän*	•	•	•	○	Reflexive
	IIAA	*iyan*	*iyän*		•	•	○	
	NAIIXAA	*naixan*						Reflexive genitive
	NAIIGAA		*näixän*					
	UBAA	*uban*	*übän*				○	
	DAĞAA	*dagan*		rendering only	•	•	○	Reflexive dative
	DAGAA		*dägän*	encoding only	•	•	○	
	TAĞAA	*tagan*			•	•	○	
	TAGAA		*tägän*	•	•	•	○	
	DURIYAA	*duriyan*	*düriyän*				○	
	YUĞAA	*yugan*			•	•	○	Reflexive accusative
	YUGAA		*yügän*		•	•	○	
	AĆAĞAA	*aćagan*			•	•	○	Reflexive ablative
	AĆAGAA		*äćägän*		•	•	○	
	TAIIĞAA	*taigan*				*tayigan*	○	Reflexive comitative
	TAIIGAA		*täigän*			*täyigän*	○	
	AURUĞU	*urugu*				•/not	○	Directive
	MINI		*mini*				○	
	ĆINI		*ćini*				○	
	MANI	*mani*	*mäni*				○	
	TANI	*tani*	*täni*				○	
	AANU	*anu*					○	Possessive
	AINU		*inü*				○	
	NI		*ni*				○	
	XINI		*xini*					
	DUNI	*duni*	*düni*					Possessive dative
	TUNI	*tuni*	*tüni*					
	XI		*xi*			•		Empty noun
	XIA		*xin*			•		

𐰲	DAXI	*daxi*	*däxi*				•		Empty noun dative?
𐰲	TAXI	*taxi*	*täxi*						
𐰺	NUĞUÐ	*nugud*					•	○	Plural
𐰺	NUGUÐ		*nügüd*				•	○	
𐰢	UÐ	*ud*	*üd*	•	•	•	○		
𐰣	NAR	*nar*	*när*	*nar*	•	•	○		
𐰺	DUĞAR	*dugar*					•		Ordinal
𐰺	DUGAR		*dügär*				•		
𐰶	SIX	*sig*							Like
𐰶	SIG		*sig*						
𐰴	ĆIX	*ćig*							Even
𐰴	ĆIG		*ćig*						
𐰲	ĆU	*ću*	*ćü*						
𐰞	LA	*la*	*lä*						Just
𐰑	DAX	*dag*				•/not			?
𐰑	DAG		*däg*			•/not			
𐰃	TA	*ta*	*tä*						?
𐰽	SAA	*san*	*sän*						?
𐰀	A	*a*	*ä*						Exclamatory
𐰆	UU	*uu*	*üü*			not			Interrogative
𐰑	DA	*da*	*dä*						Modal particle?
𐰺	AUIGAI		*ügäi*		•	•/not	○		Negative

* EOF *