L2/20-XXX

Draft technical note:

Text representation and shaping specification of the Mongolian script

技术笔记草案: 蒙古文字的文本表记与变形规范

To: Unicode Technical Committee
From: Liang Hai / 梁海 <lianghai@gmail.com>
Date: 27 August 2020

Document history 文档历史

This is a revision of the following:

- L2/19-368, *Draft technical note: Text representation and shaping specification of the Mongolian script / 技术笔记草案: 蒙古文字的文本表记与变形规范*
- L2/19-130 (WG2 5104, MWG/3-N8R), *Towards a well-formed Mongolian specification that allows interoperable implementations*

The following are the major changes since L2/19-368:

- Highlighted significant additions and changes with yellow background.
- *Unfortunately*, as the author is bilingual only in Chinese and English, Cyrllic Mongolian translations will only be available for title and section headings once document content is stable. A Hudum Mongolian translation will also be available for title.
- SSBM for decomposing vowel letters’ marked medial forms.
- Removed default variants.
- Remove Phoenician names for written units.
- […]

And the major changes in L2/19-368 since L2/19-130:

- Rearranged sections.
  - Also added Chinese translations for title and section headings.
- Clarified design considerations.
- Revised character usage and shaping rules.
  - Unified U+183A KA into U+183B KHA.
  - Removed the special treatment of *ŋ* as not involving a bowed written unit.
  - Clarified that ZWJ and ZWNJ should be kept from average users.
  - Preliminarily introduced data files.
- Rephrased the four encoding principles.

  - Briefly introduced the most confusable phonetic letter groups.
  - Rewrote the section about particles (previously “enclitics”).
- Revised terminology.
  - Certain concepts have been substituted with more commonly used ones: grapheme → written unit, enclitic → particle.
  - Certain translated native terms have been substituted with transcriptions: splash → chachlag, pad → devsger. [Add cross references or add glossary, probably in Introduction.]
  - Merged consonant coda and vowel offglide into the traditional concept devsger.
  - Revised transliteration scheme: *ä* → *e* → *é*, *x* → *h* → *ħ*, X → H → Ħ.

Contents 目录

1. Introduction 引言

Two decades ago, Unicode/10646 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 (MGC/01-01, 2018), few complete specifications have ever been available for the community to discuss and evolve in order to eventually reach an agreement.

This specification 本规范

This *draft* specification demonstrates and proposes how to specify both comprehensive guidelines for text representation and coherent rules for text shaping. It mainly deals with the first three layers in the overall technical architecture of rendering texts in the Mongolian script:

- The Unicode Standard and ISO/IEC 10646, a synchronized pair of standards, specify identities of encoded characters and standardized variation sequences.
- The Unicode Standard and its various supplementary standards further 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.).
- The required complex shaping is implemented in fonts and text shaping engines, according to OpenType Layout (OTL) or other complex shaping models, such as AAT (Apple Advanced Typography) and Graphite.
- Hudum text is treated inline as horizontal and left-to-right, while text layout engines are responsible for setting lines vertically and arranging multiple lines with the preferred left-to-right order.

Although the architectural defects cannot be resolved without migrating to a radically different encoding model, a meticulous specification can help eliminate unwanted differences between vendor implementations.

Design considerations of this specification include:

- Systematic (with the Twelve Syllabaries as the foundation) and unambigous guidelines for text representation
- Moderate normalization of orthographical disagreements
- Minimal unnecessary confusability
- Explicitly modularized and ordered rules for text rendering, with well-defined fallback mechanisms
- Minimal dictionary-based shaping

How to read

[...]

Initial scope 初步范围

As a beginning, the scope has been restricted to the typical style of Hudum (ᠬᠤᠳᠤᠮ *hudum;* Худам *khudam;* “old” in Oirat Mongolian) writing system that is contemporarily used by the Mongolian language.

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.).

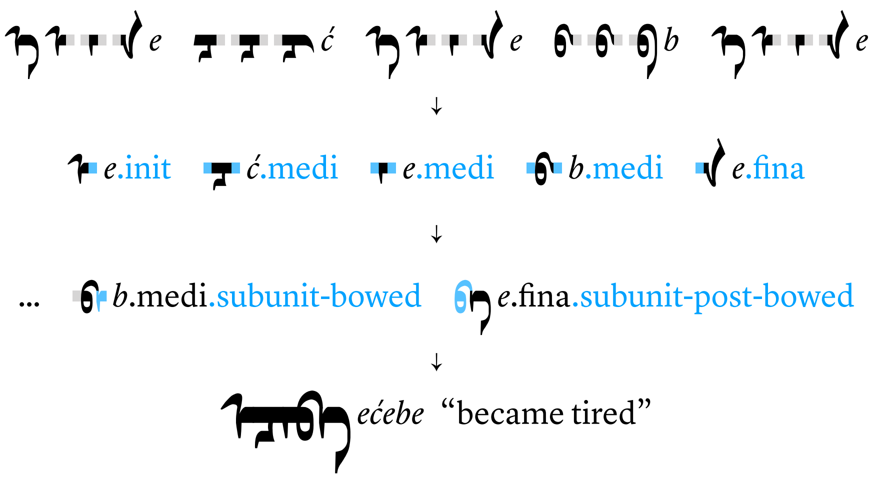
2. Principles 原则

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 actually 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, *Written units*. Figure 1 shows a simple word that 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:

1. **Phonetic letters are encoded as characters 语音字母编码为字符**
2. **Automatic shaping is orthographical 自动变形基于正字法**
3. **Particles are grammatically encoded 助词根据语法编码**
4. **Uncaptured forms are manually requested 漏网形受手动请求**

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 shaped* (Section 4, *Text rendering*).

2.1 Phonetic letters are encoded as characters 语音字母编码为字符

Hudum does not have a well-received system of typical letters (i.e., user-perceived primary units of writing) that is common to most writing systems. Instead, users are accustomed to identifying letters on a much more phonetic level, where letters are not reliably related to writing and are thus considered to be *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	*e*	*é*	*i*	*o*	*u*	*ö*	*ü*
ᠠ	ᠡ	ᠧ	ᠢ	ᠣ	ᠤ	ᠥ	ᠦ
a	ə	ə	i, ɪ	ɔ	ʊ	o	u

Transliteration
← *Representative written form*
Phoneme

n	*ŋ*	*b*	*p*	*h*	*g*	*m*	*l*	*s*	*ś*	*t*	*d*	*ć*	*j*	*y*	*r*	*w*
ᠨ	ᠩ	ᠪ	ᠫ	ᠬ	ᠭ	ᠮ	ᠯ	ᠰ	ᠱ	ᠲ	ᠳ	ᠴ	ᠵ	ᠶ	ᠷ	ᠸ
n	ŋ	p	p^h	x, χ	k, q	m	l	s	ʃ	t^h	t	tʃ^h	tʃ	j	r	w

f	*k*	*c*	*z*	*ħ*	*ř*	*ł*	*ž*	*č*
ᠹ	ᠺ	ᠼ	ᠽ	ᠾ	ᠿ	ᡀ	ᡁ	ᡂ
f	k^h	ts^h	ts	x	ɻ	ɬ	tʂ	tʂ^h

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 Mongolian phonology are also provided for reference. Hudum phonetic letters do not however cover all phonemes in the contemporary spoken language.
4. The seven native vowel letters are often referred to as *vowel one* to *vowel seven*, in order to distinguish in writing the two visually identical pairs, *o*/*u* and *ö*/*ü*.
5. Light gray highlighted ones are loanword letters, thus often excluded from the alphabet. *p* and *w* were introduced early, and today are often considered to be native letters. *ŋ* is often excluded, because it is the only consonant letter that cannot occur in syllable onset positions. *ŋ* and *ł* are often excluded together when they are considered to be letter sequences *ng* and *lh*.
6. Letters *é* and *ħ* are often considered to be disambiguating written forms of *e* and *h*, respectively, and their duplicated phonemes are highlighted in light gray.
7. Letters *ř*, *ž*, and *č* are often excluded when they are considered to be special written forms of *r*, *j*, and *ć* for transcribing Chinese syllables each in only one syllable: ᠿᠢ, ᡁᠢ, and ᡂᠢ. Actually *ř* has extended usage today for other Chinese r-initial syllables and loanword [ɻ~ʒ] sounds from other languages.

Multi-to-one confusability. The system of phonetic letters is largely based on historical phonemes of the Mongolian written language reflected in the conservative orthography, instead of how the under-differentiated Hudum writing system actually works with its limited set of written units. 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.

Because modern pronunciations of words are actually distant from what spellings of Hudum literally suggest, identifying phonetic letters is a highly subjective process, and makes text representation fundamentally unreliable. Among others, the following groups of phonetic letters are especially prone to inconsistent identification:

- *a* ~ *e*
- *é* ~ *w*
- *i* ~ *j* ~ *y*
- *o* ~ *u*, *ö* ~ *ü*, and *o* ~ *u* ~ *ö* ~ *ü* ~ *w*
- *h* ~ *g* as in (*s*|*d*)*h* vs. (*s*|*d*)*g*, due to phonetic alternation: [χ] → [q] / [s, t] _
- *s* ~ *ś* as in *si* vs. *śi*, due to phonetic alternation: [s] → [ʃ] / _ [i]
- *t* ~ *d*

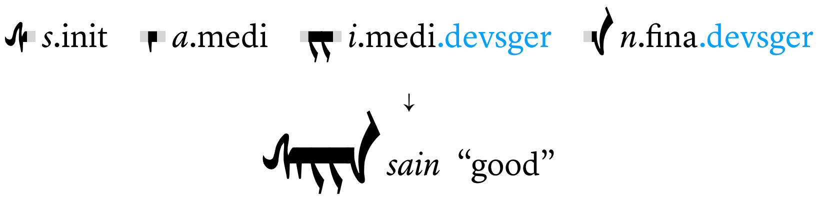
For the exact confusable forms of these phonetic letter groups, see Table 7, *Encoded phonetic letters and their written forms.*

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

2.2 Automatic shaping is orthographical 自动变形基于正字法

Typically, orthographical features of a writing system are directly reflected in text representation (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 4.2, *Orthographical shaping*.

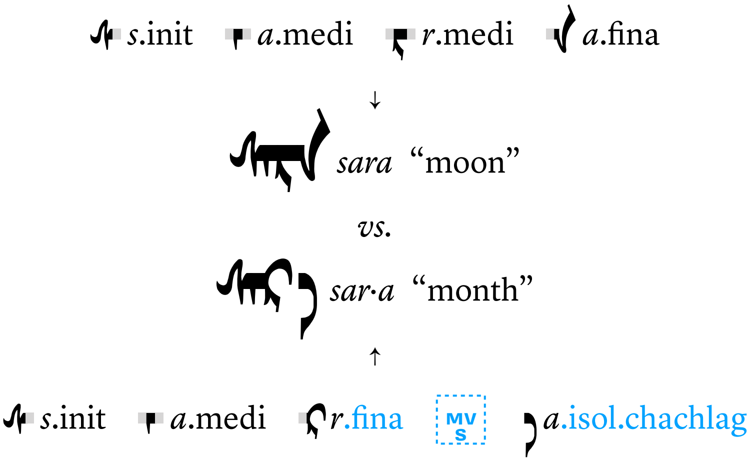
Figure 2. Orthographical shaping



Subjective and incomplete. The one-to-multi unpredictable nature of phonetic letters, combined with the orthographical flexibility of Hudum, 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.

Chachlag. The structure *chachlag* (ᠴᠠᠴᠤᠯᠭ᠎ᠠ *ćaćulg·a;* цацлага *tsatslaga;* "sprinkle"), a non-joining leftward tail form of *a* and *e* 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 chachlag is requested with a dedicated format control, see "Chachlag" in Section 4.2.

Figure 3. Chachlag

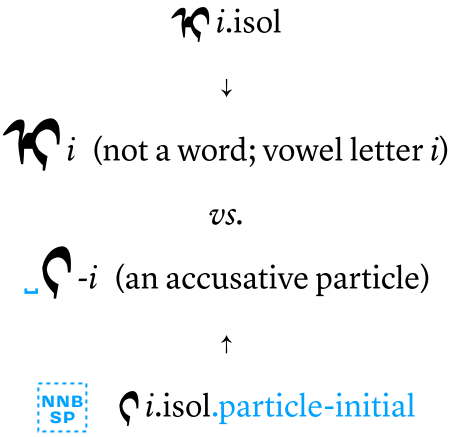


s.init *a*.medi *r*.medi *a*.fina
↓
sara "moon"
vs.
sar·a "month"
↑
s.init *a*.medi *r*.fina MVS *a*.isol.chachlag

2.3 Particles are grammatically encoded 助词根据语法编码

There is a group of special, frequently used *particles* that are written separately from their modified word (as postposition words), but are conventionally considered to be grammatically suffixes (and are thus considered to be part of the modified word). Many of them exhibit special orthographical features that make them distinct from ordinary words. But as not all of them are graphically distinct, the scope of such words relies on highly subjective grammatical categorization.

Figure 4. Particles



i.isol
↓
i (not a word; vowel letter *i*)
vs.
-*i* (an accusative particle)
↑
NNBSP *i*.isol.particle-initial

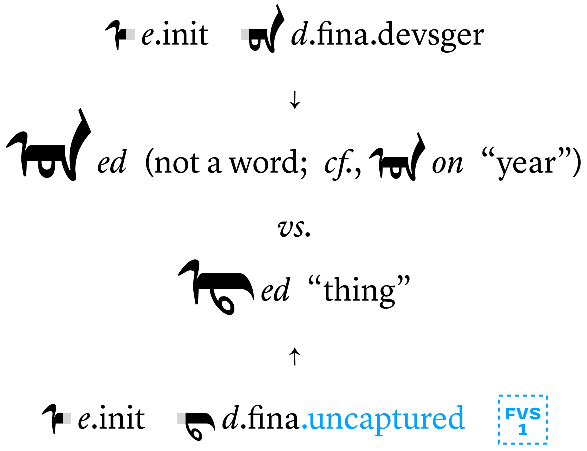
All of these particles are then all marked at the encoding level by a special whitespace that connects a particle to its preceding word (the modified word or another particle). This whitespace signals the particles' special shaping requirement (which cannot be orthographically predicted from underlying phonetic letters alone), and aids special text processing treatments (for their close relationship with the modified word). See Section 4.3, *Particles*.

2.4 Uncaptured forms are manually requested 漏网形受手动请求

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 Section 4.4, *Uncaptured forms*.

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.

Figure 5. Manual overriding



3. Architecture 架构

This section introduces in summary how Hudum text is encoded and shaped. The following aspects are covered:

- Required characters
- Expected shaping mechanisms
- Cursive script shaping mechanism
- Hudum-specific shaping mechanism

For exact, Hudum-specific specification of text. representation and shaping, see Section 4, *Rules*.

3.1 Character set 字符集

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)**	`0020 SPACE`	
	Format controls (3)	`200C ZERO WIDTH NON-JOINER`	
		`200D ZERO WIDTH JOINER`	
		`202F NARROW NO-BREAK SPACE`	
	Misc	·()«»‹›0123456789 *etc.*	
Mongolian	**Punctuation (6)**	`1800 MONGOLIAN BIRGA`	BIRGA and FOUR DOTS have limited modern usage
		`..1805 MONGOLIAN FOUR DOTS`	
	Format controls (5)	`180A MONGOLIAN NIRUGU`	
		`..180E MONGOLIAN VOWEL SEPARATOR`	
	Digits (10)	`1810 MONGOLIAN DIGIT ZERO`	Limited modern usage
		`..1819 MONGOLIAN DIGIT NINE`	
	Phonetic letters (35)	`1807 MONGOLIAN SIBE SYLLABLE BOUNDARY MARKER`	U+1807 SSBM represents the syllable-initial consonant placeholder for medial forms; U+183A KA is excluded as its represented letterform is now unified into U+183B KHA
		`1820 MONGOLIAN LETTER A`	
		`..1839 MONGOLIAN LETTER FA`	
		`183B MONGOLIAN LETTER KHA`	
		`..1842 MONGOLIAN LETTER CHI`	

CJK	Misc	? ! — «» ‹› 0 1 2 3 4 5 6 7 8 9 *etc.*

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, *Written units* and Section 3.4, *Encoded phonetic letters*. The following subsections introduce other types of characters.

*3.1.0 *Notable changes*

- U+183A KA
- SSBM

Potential additional improvements:

- Unify h/g, t/d, then use existing characters to encode them graphically.
- Deprecate o and ö.

3.1.1 *Format controls* 格式控制符

[Clarification: including characters that have implication on shaping and formating of Hudum, not just Unicode Format_Control]

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, ZWNJ acts like an ordinary space (U+0020), while ZWJ and nirugu act like medial forms of an ordinary letter (defined in Section 2.1).

ZWNJ and ZWJ, however, should be kept from average users' access on common keyboard layouts, as day-to-day text does not need these tricky characters.

The visible character nirugu should be used for 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 chachlag variation. It is transcribed as "·" (a middle dot). See "Chachlag" in Section 3.5, *Additional variation patterns*. NNBSP is a Mongolian-specific whitespace and format control for marking and shaping particles, and note that it is also used as a general whitespace by other scripts. See "Particles" 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 Section 4.4, *Uncaptured forms*.

3.1.2 Other characters 其他字符

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 Shaping process 变形流程

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 3. Overview of shaping process

Shaping phase		*Shaping step*
IA. General		• **Basic character-to-glyph mapping**
IIA. Cursive script		• **Initiation of cursive positions**
III. Hudum-specific *Reduction of phonetic letters to written units*	*Phonetic:*	**1. Chachlag**
		2. Syllabic
		3. Particle
	Graphemic:	**4. Devsger *i***
		5. Post-bowed
	Uncaptured:	**6. FVS-selected**

IIb. Cursive script (continued) *Sub-written unit variations*	• **Variation involving bowed written units** • **Cleanup of format controls** • Optional treatments
Ib. General (continued) *Typography*	• **Vertical forms of punctuation marks** • Optional treatments

For details about the phase III, see Section 3.3, *Hudum-specific shaping phase.*

3.2.1 General shaping phases 通用变形阶段

These are the basic mechanisms in fonts, applicable for all scripts.

Basic character-to-glyph mapping (phase Ia) 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.

3.2.2 Cursive script shaping phases 连写文种变形阶段

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.

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 in one of the 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.

For the exact algorithm used here for initiating cursive positions, see “Arabic Cursive Joining” in Section 9.2, *The Unicode Standard, Version 12.0—Core Specification.*

Implementation. 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.

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

Graphemic variation after bowed written units. Before the sub–written unit variation introduced in Section 3.3, *Written units*, bowed written units first cause a following U.fina vowel form to change to O.fina.

3.2.3 Hudum-specific shaping phase 胡都木特有的变形阶段

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.

Forms not captured by the predictive conditions introduced below are requested with FVSes. See Section 4.4, *Uncaptured forms*.

Phonetic and written unit–level shaping steps. A number of commonly recognized variation patterns are considered for the Hudum text presentation. Some patterns correspond to predictive shaping rules formalized as the 11 conditions below, while the rest are not executed in fonts. In the phonetic and written unit–level shaping steps (1–5), certain letters are analyzed as subjects and a set of contextual rules determine if the subject letters are in one of the 11 conditions.

```
Chachlag
Marked
Onset
Devsger
Chachlag_Onset
Masculine_Onset
Masculine_Devsger
Feminine
Particle
Dictionary_Particle
Post_Bowed

In_Isolation
```

Contextual rules for each step and condition are specified in Section 4.2.1, *Orthographical shaping rules* (Table 6). The execution order of shaping steps is critical, as they have dependencies of previous ones, and a later decided condition overwrites an earlier one’s output. For a reference list of particles that are commonly recognized as NNBSP-applicable (step 3), see Appendix C. For predefined classes of phonetic letters that are used in contextual rules, see Table 7.

An alternative model for contextual rules. [Multi-pass analyses then single-pass substitution: advanced OTL rules (attested performance issue).]

[Compare with the current multi-pass substitutions: tightly coupled to rule order, maintenance of glyph classes of letters, ordinary OTL rules, arbitrary default positional forms and other intermediate forms, some conditions (e.g., onset) do not lead to forms different from the arbitrary default positional forms, arbitrary composition of conditions.]

4. Rules 规则

This section specifies how exactly Hudum text should be encoded, and how properly encoded text should be shaped. The section starts with a thorough introduction of the fundamental building blocks:

- Written units as a concrete low-level abstraction for describing written forms
- Phonetic letters and their written forms

Then in order to fill the gap between encoded phonetic letters and actual written units:

- Variation patterns considered for reducing phonetic letters to written forms

The exact text representation of a word is thus determined by:

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

Then tightly specified shaping rules are responsible for reproducing the written form of the word.

4.1 Written units and phonetic letters 书写单位和语音字母

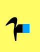
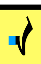
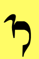
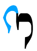
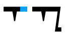
Instead of the confusingly encoded characters for phonetic letters, the written units of Hudum are first introduced for building a concrete analysis model.

4.1.1 Written units 书写单位

Before examining the encoded phonetic letters, encoding-independent *written units* are defined in Table 4. All written forms in this specification are analyzed with and formally referred to as sequences of written units on certain cursive positions, for the sake of clarity and accuracy.

Written units are assigned single-letter transliterations that are always in SMALL-CAPS, and are further marked by ‹...› when referred to inline. Phonetic letters *ŋ* and *ł*, as well as certain variants of other phonetic letters, are written as sequences of written units, thus do not appear in the table.

Table 4. Written units

Written unit		*Positional forms: .init, .medi, .fina, .isol*				*Sub-written unit variants*		*Represented phonetic letters*				*Note*
A							• 	ʼ/a/e	ʼ/a/e/n	a/e/n	a/e	
Á	A2										a/e	
N							•	n	n	n		see ‹A›

W								•	•	*w*	*w/é*	*w/é*		
H								•	•	*h*	*h/g*	*h/g*		
Ğ	Gh							•	•	*g*	*g*	*g*		
U										*u/ü*	*o/u/ö/ ü/w*	*o/u/ö/ ü/w*	*u/ü*	
O												*o/u*		
Ü	Ue											*ö/ü*		
I								•		*i/j/y*	*i/y*	*i/y*	*i/j*	
Y								•		*y*	*y*			
G								•		*h/g*	*h/g*	*g*		
K										*k*	*k*	*k*		
D								•		*d*	*t/d*	*d*		
M										*m*	*m*	*m*		; *see* ‹A›
S								•	•	*s*	*s*	*s*		
Ś	Sh							•	•	*ś*	*ś*	*ś*		
B										*b*	*b*	*b*		
P										*p*	*p*	*p*		
F										*f*	*f*	*f*		
Ć	Ch									*ć*	*ć*	*ć*		
J											*j*	*j*		
C										*c*	*c*	*c*		
Z										*z*	*z*	*z*		
R								•		*r*	*r*	*r*		
T										*t/d*	*t*	*t*		
Đ	D2										*d*	*d*		
L										*l*	*l*	*l*		*see* ‹A›
Ř	Rh									*ř*	*ř*	*ř*		*see* ‹A›
Ħ	H2									*ħ*	*ħ*	*ħ*		*see* ‹A›
Ž	Zr									*ž*				
Č	Cr									*č*				

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

Sub-written unit variation involving bowed written units. Bowed-looking written unit groups ‹G› and ‹B›, namely, ‹G›, ‹K›, ‹B›, ‹P›, and ‹F› (on their initial and medial positions) cause certain written units (on their medial and final positions) to join perpendicularly without a stem when following. Attested perpendicularly joinable medial and final forms are marked in the Table 4 with either a bullet or the perpendicularly joined form (if significant structural change is observed).

Note the commonly mentioned “leftward tail” (or “feminine tail”) of ‹A› is here considered merely the perpendicularly joined form of ‹A.fina›. Written unit ‹U› lacks a perpendicularly joined final form, as the forms it would orthographically take (‹O.fina› or, marginally, ‹Ü.fina›) have been disunified.

Stylistic variants. Certain commonly used stylistic variants of written units exhibit significant structural differences:

- All written units that involve a *titem* (ᠲᠢᠲᠢᠮ *titim;* титэм *titem;* “crown”) on certain positional forms (i.e., ‹A›, ‹N›, ‹M›, ‹L›, ‹Ž›, and ‹Ħ›) have a historical preferred, non-titem variant.
- Among the two variants of written unit ‹K›, the Ali Gali *kha* letterform (the default in the table) and the Manchu velar *k’a* letterform (which also appears to be similar to Ali Gali *ga* and Todo velar *ga*) are largely interchangeably used in China; while the Ali Gali *kha* letterform is strongly preferred in Mongolia.
- Written units ‹M› and ‹B› 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 MONGOLIAN LETTER TODO BA and U+184F MONGOLIAN LETTER TODO MA).

4.1.2 *Encoded phonetic letters* 编码的语音字母

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 written unit sequences, as shown in Table 5.

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. Red highlighted written forms are the ones only accessible with manual overriding.

Table 5. Encoded phonetic letters and their written forms

Phonetic letter & character		*Written forms: positional forms, including automatic contextual variants & manual variants*		*Note*
’	1807 SSBM	᠇	A	

a	`1820 A`	ᠠ	ᠠ			AA	AA			• AA.isol/init—*particles* • AA.medi—*compound words* • Á.isol—*chachlag*
		ᠠ	ᠠ	ᠠ	ᠠ	A	A	A	A	
		ᠠ				Á				
e	`1821 E`	ᠡ	ᠡ	ᠡ	ᠡ	A	A	A	A	• Á.isol—*chachlag*
		ᠡ				Á				
é	`1827 EE`	ᠧ	ᠧ			AW	AW			Cross-written unit simple letter
				ᠧ	ᠧ			W	W	
i	`1822 I`	ᠢ	ᠢ			AI	AI			• I.isol/init—*particles* • AI.medi—*compound words* • II.medi—*devsger*
		ᠢ	ᠢ	ᠢ	ᠢ	I	I	I	I	
				ᠢ				II		
o	`1823 O`	ᠣ	ᠣ			AO	AU			• AU.medi—*compound words* • O.fina—*initial-body, post-bowed, disambiguating*
				ᠣ	ᠣ			U	U	
					ᠣ				O	
u	`1824 U`	ᠤ	ᠤ			AO	AU			• U.isol/init—*particles* • AU.medi—*compound words* • O.fina—*initial-body, post-bowed*
		ᠤ	ᠤ	ᠤ	ᠤ	U	U	U	U	
					ᠤ				O	
ö	`1825 OE`	ᠥ	ᠥ			AÜ	AUI			• AUI.medi—*compound words* • UI.medi, Ü.fina—*initial-body* • O.fina—*post-bowed*
				ᠥ	ᠥ			U	U	
				ᠥ	ᠥ			UI	Ü	
					ᠥ				O	
ü	`1826 UE`	ᠦ	ᠦ			AÜ	AUI			• U.isol/init—*particles* • AUI.medi—*compound words* • UI.medi, Ü.fina—*initial-body, disambiguating* • AU.isol—*lexical* • O.fina—*post-bowed*
		ᠦ	ᠦ	ᠦ	ᠦ	U	U	U	U	
				ᠦ	ᠦ			UI	Ü	
					ᠦ				O	
		ᠦ				AU				
n	`1828 NA`		ᠨ	ᠨ	ᠨ		N	N	N	• N.fina—*pre-chachlag* • A.medi—*devsger*
				ᠨ	ᠨ			A	A	
ŋ	`1829 ANG`			ᠩ	ᠩ			AG	AG	.init unattested
b	`182A BA`		ᠪ	ᠪ	ᠪ		B	B	B	

p	`182B PA`						P	P	P	
h	`182C QA`						H	H	H	Default .fina unattested • G.init/medi—*feminine* • H.fina—*pre-chachlag*
							G	G		
g	`182D GA`						Ğ	Ğ	Ğ	• G.init/medi/fina—*feminine* • H.medi—*devsger* • Ğ.fina—*pre-chachlag*
							G	G	G	
								H	H	
m	`182E MA`						M	M	M	
l	`182F LA`						L	L	L	
s	`1830 SA`						S	S	S	
ś	`1831 SHA`						Ś	Ś	Ś	
t	`1832 TA`						T	T	T	• T.medi—*disambiguating*
								D		
d	`1833 DA`						T			• D.init/fina—*disambiguating* • Đ.medi—*devsger*
							D	D	D	
								Đ	Đ	
ć	`1834 CHA`						Ć	Ć	Ć	
j	`1835 JA`							J	J	• I.isol—*pre-chachlag*
						I	I			
y	`1836 YA`						Y	Y		Default .fina unattested • I.init/medi—*particles* • I.fina—*pre-chachlag*
							I	I	I	
r	`1837 RA`						R	R	R	
w	`1838 WA`						W	W	W	• U.fina—*pre-chachlag*
									U	
f	`1839 FA`						F	F	F	
k	`183B KHA`						K	K	K	
c	`183C TSA`						C	C	C	
z	`183D ZA`						Z	Z	Z	
ħ	`183E HAA`						Ħ	Ħ	Ħ	
ř	`183F ZRA`						Ř	Ř	Ř	

ł	`1840 LHA`		LĦ LĦ	.fina unattested
ž	`1841 ZHI`		Ž	.medi/fina unattested
č	`1842 CHI`		Č	.medi/fina unattested

Notes:

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

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-written unit variations are also involved in various Arabic-script writing systems, but they are always constrained within the cursive joining mechanism in the Unicode Arabic encoding.

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.

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

4.2 Orthographical shaping 正字法变形

4.2.1 Orthographical shaping rules

Table 6. Hudum shaping: phonetic and written unit–level

Shaping step	*Subject letters*	*Rules*	*Resulted condition*
1. **Chachlag**	*a/e*	**if** follows MVS:	**`Chachlag`**
	o/u/ö/ü	**if** follows an initial *consonant*:	**`Marked`**
	n/j/y/w	**if** precedes MVS that precedes *a/e*.isol:	**`Chachlag_Onset`**
	h/g	**if** precedes MVS that precedes *a*.isol:	**`Chachlag_Onset`**

2. Syllabic	*n/d*	**if** precedes a *vowel*:	`Onset`
		else if follows a *vowel*:	`Devsger`
	h/g	**if** precedes a *masculine vowel*:	`Masculine_Onset`
		else if precedes a *feminine vowel* or *neuter vowel*:	`Feminine`
	g	**else if** follows a *masculine vowel*:	`Masculine_Devsger`
		else if follows a *feminine vowel*:	`Feminine`
		else if remotely follows a *masculine vowel* without a blocking *feminine vowel*:	`Masculine_Devsger`
		else:	`Feminine`
3. Particle	*a/i/u/ü/d*	**if** follows NNBSP:	`Particle`
	u/ü	**if** follows an initial *consonant* that follows NNBSP:	`Particle`
	y	**if** is in a word *yin/yi/iyar/iyer/iyan/iyen* that follows NNBSP:	`Dictionary_Particle`
4. Devsger *i*	*i*	**if** follows a *vowel* written form that does not end with a written unit ɪ:	`Devsger`
5. Post-bowed	*o/u/ö/ü*	**if** is in the written form of ᴜ and follows a bowed written unit ɢ/ᴋ/ʙ/ᴘ/ғ:	`Post_Bowed`

Table 7. Phonetic letter classes

Class	*Members*
masculine vowel	*a o u*
feminine vowel	*e é ö ü*
neuter vowel	*i*
(vowel)	class *masculine vowel* + class *feminine vowel* + class *neuter vowel*
consonant	*n ŋ b p h g m l s ś t d ć j y r w f k c z ħ ř ł ž č ʼ*

When a specific positional written form of a letter is decided by a font to be in one of the conditions above, it is converted to the specified conditional form in Table 8. Unspecified positions in the table do not affect the subject written form.

Table 8. Conditional forms of phonetic letters

Phonetic letter & character	*Condition*	*Conditional forms*						*Note*
	`Chachlag`	ᠠ				Á		

a 1820 A	Particle	ᠠ	ᠠ			A	A			
e 1821 E	Chachlag	ᠡ				Á				
i 1822 I	Particle	ᠢ	ᠢ			I	I			
	Devsger			ᠢ				II		
o 1823 O	Marked				ᠣ				O	
	Post_Bowed				ᠣ				O	
u 1824 U	Particle	ᠤ	ᠤ	ᠤ	ᠤ	U	U	U	U	
	Marked				ᠤ				O	
	Post_Bowed				ᠤ				O	
ö 1825 OE	Marked			ᠥ	ᠥ			UI	Ü	
	Post_Bowed				ᠥ				O	
ü 1826 UE	Particle	ᠦ	ᠦ	ᠦ	ᠦ	U	U	U	U	
	Marked			ᠦ	ᠦ			UI	Ü	
	Post_Bowed				ᠦ				O	
n 1828 NA	Onset		ᠨ	ᠨ			N	N		Onset and pre-chachlag forms match default positional forms
	Devsger			ᠨ	ᠨ			A	A	
	Chachlag_Onset				ᠨ				N	
h 182C QA	Masculine_Onset		ᠬ	ᠬ			H	H		Onset + feminine forms match default positional forms
	Feminine		ᠬ	ᠬ			G	G		
	Chachlag_Onset				ᠬ				H	
g 182D GA	Masculine_Onset		ᠭ	ᠭ			Ğ	Ğ		Onset/devsger + feminine forms match default positional forms
	Masculine_Devsger			ᠭ	ᠭ			H	H	
	Feminine		ᠭ	ᠭ	ᠭ		G	G	G	
	Chachlag_Onset				ᠭ				Ğ	
d 1833 DA	Onset		ᠳ	ᠳ			T	D		Onset forms match default positional forms
	Devsger			ᠳ	ᠳ			Đ	Đ	
	Particle		ᠳ				D			
j 1835 JA	Chachlag_Onset	ᠵ				I				.fina?
y 1836 YA	Chachlag_Onset				ᠶ				I	.isol?

y 1836 YA	Dictionary_Particle						I	I		
w 1838 WA	Chachlag_Onset								U	.isol?

4.2.2 *Data files*

Phonetic letter classes can be recorded in a data file as part of the Unicode Character Database (UCD). For example in the following format:

```
# Mongolian_Category = Vowel_Masculine

1820       ; Vowel_Masculine #      MONGOLIAN LETTER A
1823..1824 ; Vowel_Masculine #  [2] MONGOLIAN LETTER O..MONGOLIAN LETTER U

# Mongolian_Category = Vowel_Feminine

1821       ; Vowel_Feminine  #      MONGOLIAN LETTER E
1825..1827 ; Vowel_Feminine  #  [3] MONGOLIAN LETTER OE..MONGOLIAN LETTER EE

# Mongolian_Category = Vowel_Neuter

1822       ; Vowel_Neuter    #      MONGOLIAN LETTER I

# Mongolian_Category = Consonant

1828..1839 ; Consonant       # [18] MONGOLIAN LETTER NA..MONGOLIAN LETTER FA
183B..1842 ; Consonant       #  [9] MONGOLIAN LETTER KHA..MONGOLIAN LETTER CHI
```

The crucial information provided in Table 8, *Conditional forms of phonetic letters* (Section 4.2, *Orthographical shaping*) can also be recorded in a data file, with a different format:

[Use the same non-ASCII transliteration scheme in the example.]

```
# Mongolian_Forms_Default

1820; A A ; A A ; A   ; A   # MONGOLIAN LETTER A
1821; A   ; A   ; A   ; A   # MONGOLIAN LETTER E
1822; A I ; A I ; I   ; I   # MONGOLIAN LETTER I
...
1828;     ; N   ; N   ; N   # MONGOLIAN LETTER NA
...

# Mongolian_Forms_Chachlag

1820; Á   ;     ;     ;     # MONGOLIAN LETTER A
1821; Á   ;     ;     ;     # MONGOLIAN LETTER E

# Mongolian_Forms_Devsger

1822;     ;     ; I I ;     # MONGOLIAN LETTER I
1828;     ;     ; A   ; A   # MONGOLIAN LETTER NA
1833;     ;     ; Đ   ; Đ   # MONGOLIAN LETTER DA

...
```

4.3 Particles 助词

The special whitespace between a particle and its preceding word (either the modified word or a preceding particle) 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 particles exhibit at least one of the following variations:

- Absence of onset placeholder and initial-body variation (because the beginning of a particle 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 *h* 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 particles. As the set of particles and usage of NNBSP are decided grammatical, a particle may or may not exhibit special variations but is still encoded with a preceding NNBSP. See Appendix C for a reference list and comparison.

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.*]

4.4 Uncaptured forms 漏网形

Proposed Draft Unicode Technical Report #54, *Unicode Mongolian 12.1 Baseline*.

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 参考文献

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.

Standards 标准

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.

MGC/01-01 (2018, 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 12.1 (2019). Mountain View: The Unicode Consortium. http://unicode.org/versions/Unicode12.1.0/.

UTC Document Registry UTC 文档记录处

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

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

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

L2/19-211 (Proposed Draft Unicode Technical Report #54), *Unicode Mongolian 12.1 Baseline*. Ken Whistler.

L2/19-247 (WG2 N5054), *Recommendations from WG 2 meeting 68*.

Other documents 其他文档

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.

6. Acknowledgements 致谢

The author would like to thank Shen Yilei / 沈逸磊 for his always enlightening feedback. Wang Yihua / 王奕桦 contributed to the analysis of letter *k*'s variants.

Liang Jinbao / 梁金宝 and Narantsogt Baatarkhuu / Баатархүүгийн Наранцогт provided the title's Hudum and Cyrillic Mongolian translations, respectively.

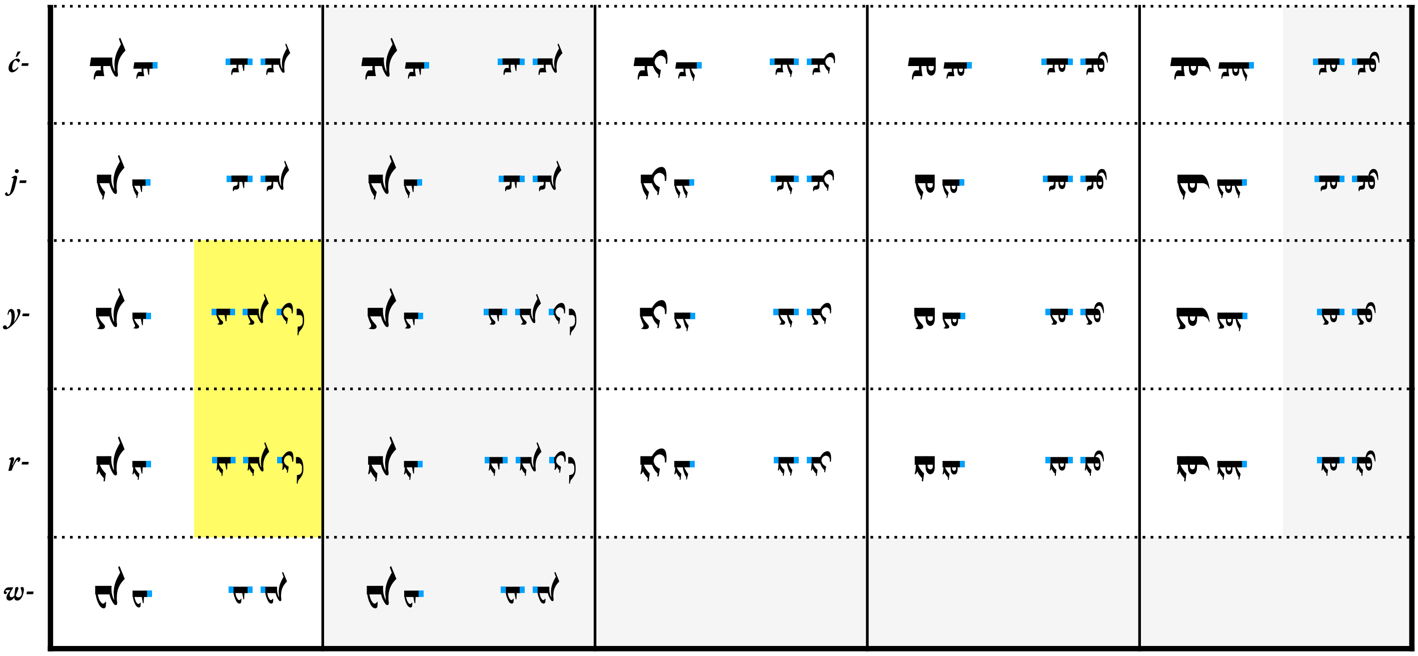
The major font used for Hudum examples is Menk Vran Tig (version 1.02), one of the fonts freely released by Menksoft (http://font.menksoft.com). Menk Garqag Tig (1.03) and Menk Scnin Tig (1.02) are also used.

A. The Twelve Syllabaries 十二字头

A.1 The First Syllabary 第一字头

Written forms of syllables: word-wise beginning *(isolated or initial) vs.* continuing *(medial or final)*

	-a		*-e*		*-i*		*-o/u*		*-ö/ü*	
Ø-	ᠠ ᠠ᠊	—	ᠡ ᠡ᠊	—	ᠢ ᠢ᠊	—	ᠣ ᠣ᠊	—	ᠥ ᠥ᠊	—
n-	ᠨᠠ ᠨᠠ᠊	᠊ᠨᠠ᠊ ᠊ᠨᠠ ᠊ᠨ᠎ᠠ	ᠨᠡ ᠨᠡ᠊	᠊ᠨᠡ᠊ ᠊ᠨᠡ ᠊ᠨ᠎ᠡ	ᠨᠢ ᠨᠢ᠊	᠊ᠨᠢ᠊ ᠊ᠨᠢ	ᠨᠣ ᠨᠣ᠊	᠊ᠨᠣ᠊ ᠊ᠨᠣ	ᠨᠥ ᠨᠥ᠊	᠊ᠨᠥ᠊ ᠊ᠨᠥ
b-	ᠪᠠ ᠪᠠ᠊	᠊ᠪᠠ᠊ ᠊ᠪᠠ	ᠪᠡ ᠪᠡ᠊	᠊ᠪᠡ᠊ ᠊ᠪᠡ	ᠪᠢ ᠪᠢ᠊	᠊ᠪᠢ᠊ ᠊ᠪᠢ	ᠪᠣ ᠪᠣ᠊	᠊ᠪᠣ᠊ ᠊ᠪᠣ	ᠪᠥ ᠪᠥ᠊	᠊ᠪᠥ᠊ ᠊ᠪᠥ
p-	ᠫᠠ ᠫᠠ᠊	᠊ᠫᠠ᠊ ᠊ᠫᠠ	ᠫᠡ ᠫᠡ᠊	᠊ᠫᠡ᠊ ᠊ᠫᠡ	ᠫᠢ ᠫᠢ᠊	᠊ᠫᠢ᠊ ᠊ᠫᠢ	ᠫᠣ ᠫᠣ᠊	᠊ᠫᠣ᠊ ᠊ᠫᠣ	ᠫᠥ ᠫᠥ᠊	᠊ᠫᠥ᠊ ᠊ᠫᠥ
h-	ᠬᠠ ᠬᠠ᠊	᠊ᠬᠠ᠊ ᠊ᠬᠠ	ᠬᠡ ᠬᠡ᠊	᠊ᠬᠡ᠊ ᠊ᠬᠡ	ᠬᠢ ᠬᠢ᠊	᠊ᠬᠢ᠊ ᠊ᠬᠢ	ᠬᠣ ᠬᠣ᠊	᠊ᠬᠣ᠊ ᠊ᠬᠣ	ᠬᠥ ᠬᠥ᠊	᠊ᠬᠥ᠊ ᠊ᠬᠥ
g-	ᠭᠠ ᠭᠠ᠊	᠊ᠭᠠ᠊ ᠊ᠭᠠ	ᠭᠡ ᠭᠡ᠊	᠊ᠭᠡ᠊ ᠊ᠭᠡ	ᠭᠢ ᠭᠢ᠊	᠊ᠭᠢ᠊ ᠊ᠭᠢ	ᠭᠣ ᠭᠣ᠊	᠊ᠭᠣ᠊ ᠊ᠭᠣ	ᠭᠥ ᠭᠥ᠊	᠊ᠭᠥ᠊ ᠊ᠭᠥ
m-	ᠮᠠ ᠮᠠ᠊	᠊ᠮᠠ᠊ ᠊ᠮᠠ ᠊ᠮ᠎ᠠ	ᠮᠡ ᠮᠡ᠊	᠊ᠮᠡ᠊ ᠊ᠮᠡ ᠊ᠮ᠎ᠡ	ᠮᠢ ᠮᠢ᠊	᠊ᠮᠢ᠊ ᠊ᠮᠢ	ᠮᠣ ᠮᠣ᠊	᠊ᠮᠣ᠊ ᠊ᠮᠣ	ᠮᠥ ᠮᠥ᠊	᠊ᠮᠥ᠊ ᠊ᠮᠥ
l-	ᠯᠠ ᠯᠠ᠊	᠊ᠯᠠ᠊ ᠊ᠯᠠ ᠊ᠯ᠎ᠠ	ᠯᠡ ᠯᠡ᠊	᠊ᠯᠡ᠊ ᠊ᠯᠡ ᠊ᠯ᠎ᠡ	ᠯᠢ ᠯᠢ᠊	᠊ᠯᠢ᠊ ᠊ᠯᠢ	ᠯᠣ ᠯᠣ᠊	᠊ᠯᠣ᠊ ᠊ᠯᠣ	ᠯᠥ ᠯᠥ᠊	᠊ᠯᠥ᠊ ᠊ᠯᠥ
s-	ᠰᠠ ᠰᠠ᠊	᠊ᠰᠠ᠊ ᠊ᠰᠠ	ᠰᠡ ᠰᠡ᠊	᠊ᠰᠡ᠊ ᠊ᠰᠡ	ᠰᠢ ᠰᠢ᠊	᠊ᠰᠢ᠊ ᠊ᠰᠢ	ᠰᠣ ᠰᠣ᠊	᠊ᠰᠣ᠊ ᠊ᠰᠣ	ᠰᠥ ᠰᠥ᠊	᠊ᠰᠥ᠊ ᠊ᠰᠥ
ś-	ᠱᠠ ᠱᠠ᠊	᠊ᠱᠠ᠊ ᠊ᠱᠠ	ᠱᠡ ᠱᠡ᠊	᠊ᠱᠡ᠊ ᠊ᠱᠡ	ᠱᠢ ᠱᠢ᠊	᠊ᠱᠢ᠊ ᠊ᠱᠢ	ᠱᠣ ᠱᠣ᠊	᠊ᠱᠣ᠊ ᠊ᠱᠣ	ᠱᠥ ᠱᠥ᠊	᠊ᠱᠥ᠊ ᠊ᠱᠥ
t-	ᠲᠠ ᠲᠠ᠊	᠊ᠲᠠ᠊ ᠊ᠲᠠ	ᠲᠡ ᠲᠡ᠊	᠊ᠲᠡ᠊ ᠊ᠲᠡ	ᠲᠢ ᠲᠢ᠊	᠊ᠲᠢ᠊ ᠊ᠲᠢ	ᠲᠣ ᠲᠣ᠊	᠊ᠲᠣ᠊ ᠊ᠲᠣ	ᠲᠥ ᠲᠥ᠊	᠊ᠲᠥ᠊ ᠊ᠲᠥ
d-	ᠳᠠ ᠲᠠ᠊ ᠳᠠ᠊	᠊ᠳᠠ᠊ ᠊ᠳᠠ	ᠳᠡ ᠲᠡ᠊ ᠳᠡ᠊	᠊ᠳᠡ᠊ ᠊ᠳᠡ	ᠳᠢ ᠲᠢ᠊ ᠳᠢ᠊	᠊ᠳᠢ᠊ ᠊ᠳᠢ	ᠳᠣ ᠲᠣ᠊ ᠳᠣ᠊	᠊ᠳᠣ᠊ ᠊ᠳᠣ	ᠳᠥ ᠲᠥ᠊ ᠳᠥ᠊	᠊ᠳᠥ᠊ ᠊ᠳᠥ



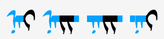
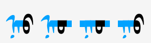
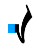
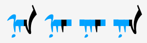
Notes:

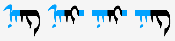
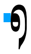
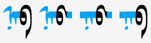
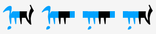
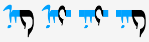
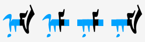
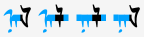
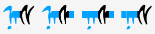
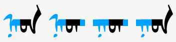
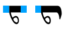
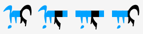
1. The Twelve Syllabaries only exhibit an incomplete, core set of orthographical rules.
2. Only the 16 native onsets, seven native vowels, and 11 native devsgers are enumerated in the basic (C)V(C) syllable structure, making up the Twelve Syllabaries. This means not only loanword syllables, even native syllables as common as *sain* are out of scope.
3. Only the enumerated syllables' *marked* beginning forms and *unmarked* continuing forms are included, however they also have *unmarked* beginning forms (in particles) and *marked* continuing forms (in non-initial stems of compound words). Also, various marginal written forms are not included.
4. Cells with light gray background contain only repeated written forms, which introduce great ambiguity into the orthography.
5. [...]

A.2 Patterns exhibited in the First Syllabary 第一字头展现的模式

[...]

A.3 The other eleven syllabaries 其余十一字头

Syllabary	*Devsger*	*Written forms: medial & final*		*Examples shown with na/ne-*	*Note*
2nd	*-i*				Medial form appends only a single ɪ.medi after an existing ɪ.medi.
3rd	*-u/ü*				
4th	*-n*				

5th	*-ŋ*				
6th	*-b*				
7th	*-g*				Masculine form for nucleus -*a*/*i*/*o*/*u*-.
					Feminine form for nucleus -*e*/*i*/*ö*/*ü*-.
8th	*-m*				
9th	*-l*				
10th	*-s*				
11th	*-d*				Alternative forms D.medi/fina  are used for disambiguating.
12th	*-r*				

B. Orthography background 正字法背景

Words, particles, stems, etc. A whitespace-separated word (a morphological word) may be either an ordinary word (which can be modified by one or more particles as a host word) or a particle. Particles 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.

Syllabic variations

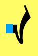
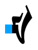
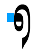
A written *syllable* in Hudum has the structure of C?V+C? (i.e., an optional *onset* consonant letter, one or more vowel letters forming the *nucleus*, and an optional *coda* consonant letter). The leading C?V part (a syllable excluding the trailing V*C?; i.e., the First Syllabary structure) has a special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

Devsger-specific forms. Traditionally, one of the eleven appendable phonetic letters from *the Second* to *Twelfth Syllabaries* (i.e., a postvocalic phonetic letter in a syllable) is recognized as a *devsger* (ᠳᠡᠪᠢᠰᠭᠡᠷ *debisger;* дэвсгэр *devsger;* “pad”). Consonant *n*, *g*, and *d* as well as vowel *i* have devsger-specific forms.

Table 9. The eleven devsgers

Type & phonetic letters		*Written forms*				*Note*
Soft devsgers	***n***			A	A	
	m			M	M	
	l			L	L	
	ŋ			AG	AG	
	b			B	B	

Hard devsgers	*g*			H	H	Masculine and feminine forms.
				G	G	
	r			R	R	
	s			S	S	
	d			Ð	Ð	
Vocalic devsgers	*i*			II	I	Considered *y* or *yi*.medi / *y*.fina in alternative theories.
	u/ü			U	U	Considered *w* in alternative theories.

A medial devsger *i* takes its devsger-specific form II unless the preceding vowel's written form already ends with a written unit I. The two vocalic devsgers are not analyzed as *y* and *w*. .... *Predictive*

Phonetic Letters *h* and y are not eligible for a devsger position. .... *Predictive*

Variations in stem domains

A special *stem domain* is a word stem together with its following zero or more suffixes and particles. 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 domain, 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 (C?V) of a stem domain:

- 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 U.fina even when it is in the initial body, and as Ü.fina after a bowed written unit; 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 U.medi/O.fina and u as UI.medi/Ü.fina.

Gender-specific forms of h and g

Consonant letters *h* 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; *e*, *é*, *ö*, and *ü* are feminine; *i* is neuter.

In principle, a stem domain is internally gender harmonious, and thus should have a determinate gender value. However, acquiring a stem domain'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 *h* 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 *h* 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 devsger *g* agrees with its preceding vowel's gender (masculine or feminine). A devsger *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.

Chachlag

This non-joining written unit of *a* and *e* is usually observed after *n*, masculine *h*, 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*/*e*, as well as to request the special forms required for certain consonants:

- Letters *n*, *h*, 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 chachlag is orthographically mandatory for *a* to appear after a word-medial *h/g*.
- Note that *h* and *y* normally do not have final forms as they are not used as devsgers.

C. Generally recognized particles 普遍承认的助词

Particle lists provided in the following documents and standards have been normalized and compared in Table 10:

- *The Users' Convention*, abbreviated as "UC" in the table, published as UNU/IIST Report No. 170 (Myatav Erdenechimeg et al. 1999) and MNS 4932: 2000.
- GB/T 25914–2010, "GB/T".
- MGC/01-01 (2018, version 1.0.2), "EAC".
  - Note that only the words that require special shaping are listed in the standard.
- L2/17-036 (Greg Eck et al. 2017), "17-036".
- L2/18-293 (Badral Sanlig and Munkh-Uchral Enkhtur 2018), "18-293".
  - Its listed particles 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 particles.

Table 10. Particles commonly recognized as NNBSP-applicable

Written form		*Text representation following an NNBSP*		*UC*	*GB/T*	*EAC*	*17-036*	*18-293*	*Note*
ᠶᠢᠨ	IIN	*yin*		•	•	•	•	○	Genitive
ᠤᠨ	UA	*un*	*ün*	•	•	•	•	○	
ᠤ	U	*u*	*ü*	•	•	•	•	○	
ᠳᠤ	DU	*du*	*dü*	•	•	•	•	○	Dative
ᠲᠤ	TU	*tu*	*tü*	•	•	•	•	○	
ᠳᠤᠷ	DUR	*dur*	*dür*	•	•	•	•	○	
ᠲᠤᠷ	TUR	*tur*	*tür*	•	•	*tür*	•	○	
ᠠ	Á	·*a*	·*e*		*a, e*	*a, e*	*a, e*	○	
ᠶᠢ	II	*yi*		•	•	•	•	○	Accusative
ᠢ	I	*i*		•	•	•	•	○	
ᠠᠴᠠ	AĆA	*aća*	*eće*	•	•	*aća*	•	○	Ablative

	IIAR	*iyar*	*iyer*	•	•	•	•	○	Instrumental
	BAR	*bar*	*ber*	•	•		•	○	
	TAI	*tai*	*tei*	•	•		•	○	Comitative
	LUĞ|Á	*lug·a*		•	•		•	○	
	LUGA		*lüge*	•	•	•	•	○	
	BAA	*ban*	*ben*	•	•		•	○	Reflexive
	IIAA	*iyan*	*iyen*		•	•	•	○	
	NAIIHAA	*naihan*							Reflexive genitive
	NAIIGAA		*neihen*						
	UBAA	*uban*	*üben*					○	
	DAĞAA	*dagan*		rendering only	•	•	•	○	Reflexive dative
	DAGAA		*degen*	encoding only	•	•	•	○	
	TAĞAA	*tagan*			•		•	○	
	TAGAA		*tegen*	•	•		•	○	
	DURIYAA	*duriyan*	*düriyen*					○	
	YUĞAA	*yugan*			•		•	○	Reflexive accusative
	YUGAA		*yügen*		•	•	•	○	
	AĆAĞAA	*aćagan*			•	•	•	○	Reflexive ablative
	AĆAGAA		*ećegen*		•		•	○	
	TAIIĞAA	*taigan*					*tayigan*	○	Reflexive comitative
	TAIIGAA		*teigen*				*teyigen*	○	
	AURUĞU	*urugu*					•/not	○	Directive
	MINI	*mini*						○	Possessive
	ĆINI	*ćini*						○	
	MANI	*mani*	*meni*					○	
	TANI	*tani*	*teni*					○	
	AANU	*anu*						○	
	AINU		*inü*					○	
	NI	*ni*						○	

	HINI	*hini*							
	DUNI	*duni*	*düni*			•			Possessive dative
	TUNI	*tuni*	*tüni*			*tüni*			
	HI	*hi*					•		Empty noun
	HIA	*hin*					•		
	DAHI	*dahi*	*dehi*			•	•		Empty noun dative?
	TAHI	*tahi*	*tehi*						
	NUĞUĐ	*nugud*					•	○	Plural
	NUGUĐ		*nügüd*			•	•	○	
	UĐ	*ud*	*üd*	•	•	•	•	○	
	NAR	*nar*	*ner*	*nar*	•		•	○	
	DUĞAR	*dugar*				•	•		Ordinal
	DUGAR		*düger*			•	•		
	SIH	*sig*							Like
	SIG		*sig*						
	ĆIH	*ćig*							Even
	ĆIG		*ćig*						
	ĆU	*ću*	*ćü*			•			
	LA	*la*	*le*						Just
	DAH	*dag*				•	•/not		?
	DAG		*deg*			•	•/not		
	TA	*ta*	*te*						?
	SAA	*san*	*sen*						?
	A	*a*	*e*						Exclamatory
	UU	*uu*	*üü*			not	not		Interrogative
	DA	*da*	*de*			•			Modal particle?
	AUIGAI		*ügei*		•		•/not	○	Negative
	NUGAA		*nügen*			•	•	○	
	YUM		*yüm*			•			

ᠶᠦᠮᠰᠡᠨ	YUMSAA		*yümsen*	•		
ᠬᠦ	GO		*hü*	•		
ᠪᠦᠦ	BUU		*büü*	not		

⁎ EOF ⁎