

[10 MAY DRAFT] A shaping model for the modern Hudum (Mongolian) writing system

Author:

- **Liang Hai** (梁海) <lianghai@gmail.com>

Major contributors (please cc when contacting the author):

- **Ma Xudong** (马旭东) <xudong.ma@gmail.com>
- **Shen Yilei** (沈逸磊) <917514182@qq.com>
- **Wang Yihua** (王奕桦) <caicaijuaner@qq.com>
- **Yan Shi** (严实) <biopolyhedron@gmail.com>

Other contributors:

Batur (ᠪᠠᠲᠤᠷ 巴特尔), **Bayindala** (ᠪᠠᠶᠢᠨᠳᠠᠯᠠ 巴音达拉), **Chimbai** (ᠴᠢᠮᠪᠠᠢ 青柏), **Hasutai** (ᠬᠠᠰᠤᠲᠠᠢ 哈苏台), **Li Shang** (李上), **Liu Chulong** (刘楚龙), **Ourileke** (ᠣᠣᠷᠢᠯᠡᠭᠡ 欧日乐克), **Tengis** (ᠲᠡᠩᠭᠢᠰ 天格思), **Zheng Weizhe** (郑维喆), et al.

Date: 10 May 2017

1. Introduction

The Unicode Mongolian encoding is being patched with random additions to resolve issues. To improve it properly, a coherent analysis model is proposed here for concerned parties to better understand behaviors of the script and requirements of the encoding.

Roughly speaking, the Unicode Mongolian script block covers the following writing systems or derivative scripts:

- Hudum (for the Mongolian language), from which derived:
 - Hudum Ali Gali (for Sanskrit–Tibetan)
 - Todo (for the Oirat language), from which derived:
 - Todo Ali Gali (for Sanskrit–Tibetan)

- Manchu (for the Manchu language), from which derived:
 - Manchu Ali Gali (for Sanskrit–Tibetan)
 - Sibe (for the Sibe language)

The writing system *Hudum* (ᠬᠣᠳᠤᠮᠤ *xudum*), commonly just referred to as the *Mongolian* writing system, is worded as *Hudum* to be clear that a subset of *the Mongolian script* is being talked. The scope of this model is set for the *Hudum* writing system of the *modern* Mongolian language.

This model is prepared with minimum details of current technical implementations or proposed changes, but to be readily understandable, it's still based on the current Unicode–OpenType Mongolian model. The *Current situation* sections try to outline current issues in the Unicode Standard and various vendor implementations (note the dated *Mongolian OpenType Specification* doesn't include any mechanisms other than the Arabic cursive joining model).

The *China–Mongolia Joint Proposal of Mongolian Transliteration* (draft, 2005) is used for transliterating, eg, ᠬᠣᠬᠬᠣᠲᠤ *xöxexota* (ie, Hohhot), while letter names consistent with the Unicode character names are used in technical presentations, eg, ᠬᠣᠬᠬᠣᠲᠤ <Q OE Q E Q O T A> or <Q+init OE+medi+firstVowel Q+medi E+medi Q+medi+masculine O+medi T+medi A+fin>. The Mongolian glyphs in this document are visual aids, while their characters extracted or copied from the PDF are not meaningful.

An outline of terminology:

- **Letter**
 - **Vowel**
 - Genders of vowel harmony:
 - **Feminine**: *e, ö, ü*.
 - **Neutral**: *i*
 - **Masculine**: *a, o, u*.
 - **Disjointed tail**: *{a}, {e}*.
 - Part of a syllable but not cursive-joining. Non-joining is unusual in the Mongolian script.
 - **Consonant**
 - **Gender-independent**: common ones.
 - **Gender-dependent**: *x, g*.
 - *g⁺am* (superscript "+" for masculine forms)
- **Syllable**
 - *c?V+c?* (an optional *onset* of a single consonant, a *nucleus* of one or more


vowels, an optional *coda* of a single consonant). An intervocalic consonant belongs to the next syllable.

- *i.ni, in.li*
- **Stray consonant**
 - A consonant not belonging to any syllable.
 - *<g>.ram, g⁺am, gem*
- **Stem**
 - *[syllable] | [stray consonant]]⁺* (one or more syllables or stray consonants).
 - *#in, #i.nin, #<g>.ram*
- **Word**
 - *[stem]⁺* (one or more stems). A word is bracketed by spaces.
 - *#in#in, #in_#in, #i.nin*
 - The disjointed tail only appears at the end of a word:
 - *#xa.r{a}_#xo.rom → #xa.ra#xo.rom*
 - **Host**
 - A regular word.
 - **Enclitic**
 - Grammatically similar to suffixes but not joined to the host. Inside an enclitic the concept stem is not considered.
 - *^un*
- **Clitic group**
 - *[host] [enclitic]** (a host followed by zero or more enclitics)
 - *#al ^un*
- **Segment** (a coined term for this model)
 - *[stem] [enclitic]** (in other words, such a scope (re)starts at the beginning of every stem and continues to the end of the clitic group.)
 - Note: *[clitic group] = [host] [enclitic]* = [stem]⁺ [enclitic]**
 - Some orthographical behaviors happen inside a *segment*, instead of a word or a complete clitic group:
 - *The aleph theory*: Vowels don't start a stem, therefore a silent consonant "aleph" is appended to stem-beginning vowels.
 - First vowel: *#ba¹.tu#mö'ng.xe ^üü*
 - Vowel harmony: *#xö.xe#x⁺o.ta ^uu*

The aleph theory. Alternatively, *+isol* and *+init* forms of vowels can be analyzed as *+fina* and *+medi* forms, respectively, prepended by a so-called *aleph*. Below is an excerpt from UTR 2 Proposals for Sinhala, Mongolian, and Tibetan:

* Aleph. Is it a rule of Mongolian spelling that every word must

begin with a consonant? If so, as in semitic alphabets having this rule, there should be a "silent consonant" (aleph) to start words that phonetically begin with vowels. It could be said that the form of such an aleph is seen in our chart as the "cap" on the heads of the initial vowel forms U+1062 -> U+1068.

Thus the graphemes in  (Example 3.1.1) can be decomposed to <Aleph I+fin> <Aleph I+medi I+medi I+fin>, leading to an analysis that vowels usually only have +medi and +fin forms, while their +isol and +init forms only exist in special cases when the aleph is not prepended.

Note the term *aleph* in the theory specially refers to the "silent consonant" part in vowel forms, although some other parts of vowels also originated from a historical *aleph*, which actually led to them sharing the same set of graphemes. The common form of an aleph is a *cap* ʁ when +init, while common forms of an aleph-less vowel *a* (historically also an *aleph*) are a *tooth* ʀ when +medi and a *right tail* ʁ when +fin. The cap ʁ, the tooth ʀ, and the right tail ʁ are positional forms of the same underlying structure, but this alternation between a cap ʁ and a tooth ʀ is often overlooked.

The vowel *e* shows exceptional interaction when prepended by an aleph: it reduces to *toothless* forms (nothing at all or a piece of *stem* ʁ when +medi, and the *left tail* ʁ when +fin).

Although practically this aleph theory can't be the base of a new Mongolian encoding model, it's very helpful for better understanding the behaviors of vowels when they start a *word stem* (where aleph is present, not only at the beginning of a word, see § 3.4. *Multi-stem words*) or an *enclitic* (where aleph is absent, although an enclitic resembles a regular word, see § 3.8. *Clarifying enclitics*).

2. Shaping phases

Input: A sequence of Unicode characters.

I. Automatic graphemes: Graphemic variations that are algorithmically predictable from a limited context of both letters and certain formatting characters. These are part of the Mongolian script's inherent orthographical behaviors, including cursive joining and some more contextual mechanisms. Eg, a fake word *antaba* for demonstrating, *ᠠᠨᠲᠤᠪᠠ → *ᠠᠨᠲᠤᠪᠠᠨᠲᠤᠪᠠ.

See § 3. *Automatic graphemic features* for a detailed breakdown of this phase.

II. Manual graphemes: Other graphemic variations. These are typically lexical variations that are *not* algorithmically predictable from a limited context, therefore need to be requested manually in encoding. Eg, *antaba* outcome from the phase I $\ast\text{𐤀𐤎𐤁𐤁}$ → $\ast\text{𐤀𐤎𐤁𐤁}$, with a disambiguated *t*.

The FVS mechanism (Free Variation Selectors) is the current solution: a special formatting character is appended to a letter to request a variation, which is independent from and does not affect the context. FVSes are assigned to positional forms instead of letters, leading to different functions of an FVS on a letter's positional forms.

Historical forms are currently handled by this mechanism.

III. Mandatory subgraphemes: Orthographical transformations, irrelevant to graphemic variations (cf. Arabic 𐤀 <LAM ALEF>). Eg, *antaba* outcome from the phase II $\ast\text{𐤀𐤎𐤁𐤁}$ → 𐤀𐤎𐤁𐤁 , now finally orthographically correct.

Limited sets of *mandatory ligatures* are defined in existing specs of contextual rules, however analyzing them as *contextual variants* would be more flexible and less implementation-dependent.

IV. Other subgraphemes: Typographical transformations, etc.

Output: A sequence of OpenType glyphs.

3. Automatic graphemic features

In the shaping phase I (see § 2. *Shaping phases*), an abstract character might be transformed into *automatic graphemes* by *automatic graphemic features*. These features are algorithmically extracted from certain contexts, and are designed to reflect linguistic *markedness*.

Sections	Automatic graphemic features
3.1. Word stems*	beginningOfNonFirstStem
3.2. Vowel count*	firstVowel
3.3. Enclitics*	beginningOfEnclitic
...	...

3.4. Syllabification*	coda
3.5. Vowel harmony*	masculine
3.6. Diphthongs*	postvocalic
3.7. The disjointed tail	preDisjointedTail
3.8. Arabic cursive joining	isol init medi fina

Features are explained beflow. The asterisked sections are about the issues not well defined in the current standard and are causing significant issues.

3.1. Word stems

- Context: [stem]
- Feature: beginningOfNonFirstStem
- Example 3.1.1: **ከሰ** *#in#in*
 - cf. **ከ** **ከ** *#in_#in*, **ከ** **ከ** *#i.nin*
- Example 3.1.2: **ከሰ** *#al.tan#o.do*
 - cf. **ከሰ** **ከሰ** *#al.tan_#o.do*, **ከሰ** **ከሰ** *#al.ta.no.do*
- Example 3.1.3: **ከሰ** *#bu.man#er.de.ni*
 - cf. **ከሰ** **ከሰ** *#bu.man_#er.de.ni*, **ከሰ** **ከሰ** *#bu.ma.ner.de.ni*
- Example 3.1.4 **ከሰ** *#ba¹.tu#mō'ng.xe*
 - cf. **ከሰ** **ከሰ** *#ba¹.tu_#mō'ng.xe*, **ከሰ** **ከሰ** *#ba¹.tu.mōng.xe*
- Example 3.1.5 **ከሰ** *#xa.ra#xo.rom*
 - cf. **ከሰ** **ከሰ** *#xa.r{a}_#xo.rom*
- Note: "#" marks the beginning of a stem; "_" marks a space between stems.

Sometimes stem boundaries are visible due to aleph-appending, syllable-breaking, and vowel-count-resetting. Note although the effects are highly predictable from the stem boundaries, since most stem boundaries are not visible, and the effects of a stem boundary is obscure to native users, stem boundaries are not generally considered suitable to be encoded directly.

With the aleph theory, it's clear that a vowel is prepended by an aleph not only when starting a word but also when starting every stem, whatever position (+isol, +init, or +medi, and even theoretically +fina) the vowel is in, leading to both +init and +medi forms of an aleph. Note the common Aleph+init is a *cap* ᐃ while an Aleph+medi is a *tooth* ᐃ.

- See Example 3.1.1 and 3.1.2 for how both a preceding consonant and a following vowel can be affected by a stem boundary, because the the preceding consonant becomes +coda and the vowel begins a stem.
- See Example 3.1.3 for how the stem-beginning E+medi doesn't seem to have a

- See Example 3.1.4 for how a stem boundary also affects the vowel count.

Current situation. The Unicode Standard employs FVSes (see § 2. *Shaping phases*, phase III. Mandatory subgraphemes) to request aleph-prepended `+medi` forms of vowels (theoretical aleph-prepended `+fina` forms are not included). Since FVSes don't have contextual effects, the effects of a stem boundary (instead of the stem boundary itself) are encoded with one or more manual FVSes applied to the letters around the stem boundary.

3.2. Vowel count

- Some vowels have special forms when they're the first vowel in a *segment*.

Current situation. Not defined in the Unicode Standard. Vendor implementations are inconsistent.

7

- Context: [enclitic]
- Feature: beginningOfEnclitic
- Example 3.3.1: *אֶל אֶל #al ^un → אֶל אֶל
 - <A+init L+fina space U+init+beginningOfEnclitic N+fina>
- Example 3.3.2: *אֶל אֶל #ül ^üü → אֶל אֶל (cf. אֶל אֶל #ül ügei)
 - <UE+init+firstVowel L+fina space UE+init+beginningOfEnclitic UE+fina>
- Example 3.3.3: *אֶל אֶל #an ^a → אֶל אֶל
 - <A+init N+fina+coda space A+isol+beginningOfEnclitic>
- Example 3.3.4: *אֶל אֶל #an {a} → אֶל אֶל (cf. *אֶל אֶל #an {a} → אֶל אֶל)
 - <A+init N+fina+coda space [disjointed tail]> (cf. <A+init N+fina+preDisjointedTail [disjointed tail]>)

With thorough analysis and proper categorizing, [enclitic] shaping is highly predictable and local to the first letter as a combination of:

- Context [stem]: Not being beginningOfStem, therefore no prepended aleph.
- Context [vowel count]: Not being firstVowel. (Since a host has at least one vowel, vowels in enclitics are always considered non-firstVowel.)
- D.init's disambiguated form is not clearly related to other mechanisms but is predictable in [enclitic].

The only two truly irregular groups, ^iy... and ^yi..., show lexical variations from their historical forms, therefore are not actually part of predictable [enclitic] behaviors. But it's still arguable if they can be handled as special cases of the [enclitic] context for practical reasons.

Although *clitic* and *enclitic* are grammatical terms, it should be noted they're used in this model to denote grammar-related orthographical behaviors, instead of pure grammatical categorization (which is always controversial). In existing specs, based on certain grammatical analysis, אֶל ügei is commonly categorized as NNBS- applicable while אֶל ^üü not, leading to irregular effects of NNBS. But by examining orthographical behaviors, it's clear that אֶל ügei is not an [enclitic] while אֶל ^üü is.

Locality. Possibly local and intrasyllabic; currently implemented case by case.

An enclitic can have arbitrary length, but only the first letter shows enclitic-specific behaviors, therefore the feature is localized to whether a +isol or +init letter is in an enclitic, thus beginningOfEnclitic.

Current situation. Due to the superficial analysis of "suffixes" on which the current standard is based, U+202F NARROW NO-BREAK SPACE (NNBSP) is employed to both provide the white space and suggest special shaping behaviors of "suffixes" that are not quite predictable, resulting in case-by-case, non-local contextual rules.

3.4. Syllabification

- Context: [syllable]
- Feature: coda
- Example 3.4.1: *ᠠᠨᠯᠢ in.li → ᠠᠨᠯᠢ (cf. ᠠᠨᠯᠢ i.ni)
- Example 3.4.2: ᠭᠠᠮ <g>.ram (cf. *ᠭᠠᠮ gam → ᠭᠠᠮ)
- Note: "." marks a syllable boundary; "<...>" marks a stray consonant.

Some consonants have special forms in a coda position.

Locality. Possibly local to a limited number of letters if the syllabification algorithm is well defined; the scope of "intrasyllabic" itself.

Current situation. Neither the Unicode Standard nor current font specs have a well defined syllabification algorithm, leading to inconsistent behaviors of fonts.

3.5. Vowel harmony

- Context: [vowel harmony]
- Feature: masculine
- Example 3.5.1: *ᠭᠠᠯᠠᠯᠠᠭ g⁺al → ᠭᠠᠯᠠᠯᠠᠭ (cf. gelᠠᠯ)
- Example 3.5.2: ᠬᠣᠭᠡᠭᠡᠭᠡ xöxex⁺ota (cf. *ᠬᠣᠭᠡᠭᠡᠭᠡ)
 - <Q+init OE+medi+firstVowel Q+medi E+medi Q+medi+masculine O+medi T+medi A+fina>

The Mongolian language has [vowel harmony] behaviors, with three genders of vowels — masculine: *a, o, u*; feminine: *e, ö, ü*; neutral: *i*.

Typically a [vowel harmony] context is either masculine (having masculine and optional neutral vowels) or feminine (having feminine and optional neutral vowels; or only neutral vowels). Therefore a [vowel harmony] context can be analyzed as feminine by default and becomes masculine when there're masculine vowels. A [vowel harmony] context restarts at the beginning of every stem and continue to the end of the clitic group. The gender of a [vowel harmony] becomes complicated in non-harmonious loanwords.

Due to unification of consonants that have a complementary distribution between masculine and feminine contexts, the shaping of *x* and *g* depends on the [vowel harmony] context. Considering a [vowel harmony] context is feminine by default, and the stray forms of *x* and *g* are also their feminine forms, *x* and *g* are analyzed to have marked masculine forms.

Locality. Not local and not intrasyllabic; possibly defined to local and intrasyllabic, leave the other cases to manual graphemic control.

The reference algorithm below is similar to what is mentioned in the Core Spec but this is explicitly restricted to a local context (without vowel harmony at a distance):

```
if followed by (A|O|U):
    +masculine
else if followed by (E|I|OE|UE):
    pass # feminine
else if preceded by (A|O|U):
    +masculine
else:
    pass # feminine
```

3.6. Diphthongs

- Context: [diphthong]
- Feature: postvocalic
- Example 3.6.1: *𐰽𐰺𐰸 *sain* → 𐰽𐰺𐰸
- cf. 𐰽𐰺 *sai*, 𐰽𐰺𐰸 *sin*
- Example 3.6.2: *𐰽𐰺𐰸𐰺𐰸 *nü'i.nüin* → 𐰽𐰺𐰸𐰺𐰸
- cf. 𐰽𐰺𐰸 *nü'i.nün*, 𐰽𐰺𐰸𐰺𐰸 *nu'i.nuin*, 𐰽𐰺𐰸 *nu'i.nun*

When an I+medi (a single *long tooth* 𐰽) follows another vowel in a [syllable], it becomes two long teeth 𐰽𐰽, except when the preceding vowel already ends with a long tooth: I(+init|+medi) 𐰽𐰽 and (OE|UE)(+init|+medi)+firstVowel) 𐰽𐰽.

Locality. Local and intrasyllabic to a vowel *i* and the preceding vowel.

Current situation. Not defined in the Unicode Standard. Vendor implementations are inconsistent on which of <I>, <Y>, or <Y I> should be used for this structure, because of controversial analyses of the underlying spelling.



3.7. The disjointed tail

- Context: [disjointed tail]
- Feature: preDisjointedTail
- Example 3.7.1: $*i\ell\{a\} \rightarrow i\gamma$ (cf. $i\ell a$)
 - `<I+init L+fina [disjointed tail]>`
- Example 3.7.2: $*i\ell\ell in\{a\} \rightarrow i\ell\gamma$ (cf. $i\ell ina$)
 - `<I+init N+fina+preDisjointedTail [disjointed tail]>`
- Example 3.7.3: $*i\ell\ell iw\{a\} \rightarrow i\ell\gamma$ (cf. $i\ell iwa$)
 - `<I+init W+fina+preDisjointedTail [disjointed tail]>`
- Note: "{...}" marks a disjointed tail.

The *disjointed tail* (ᠴᠠᠴᠢᠯᠭᠠᠨ, *čačulḡ_a*, or a commonly seen transcription of the Khalkha pronunciation, *tsatslag*) is a special, non-joining form of vowels *a* and *e*.

It only appears at the end of a sequence of letters, and is changed to a regular vowel *a* or *e* when another sequence of letters (eg, a suffix or a word stem) is joined after (see Example 3.4.5).

Although visually the gap before a disjointed tail is not necessarily smaller than a word space, it's seldom analyzed as a whitespace character, because of the clear interaction across the gap. Some consonants have special forms when preceding a disjointed tail, ie, `+preDisjointedTail`. The *dot* or *double-dot* logically belonging to those forms might be visually placed on the disjointed tail (see Example 3.7.2). Some `+preDisjointedTail` forms are related to *onset* forms (see § 3.3. *Syllabification*) because a disjointed tail extends a syllable boundary, while the other are related to historical undifferentiation of writing certain consonants and vowels.

It's important to understand a *disjointed tail* is different to a *left tail* ് (a visual part of a subgraphemic variation of the *right tail* ്, eg, * *ba* →  *ba*). Although often visually similar (in *Baiti*, etc), they are not the same grapheme and show different orthographical behaviors.

Locality. Local and intrasyllabic to a disjointed tail and the preceding consonant.

Current situation. The Unicode Standard employs a formatting character, U+180E MONGOLIAN VOWEL SEPARATOR (MVS), to form the disjointed tail. An MVS inserted before a normal *a* or *e* separates it from the preceding letter while marks it special, ie, `<MVS (A|E)> → <(A|E)+isol+disjointedTail> → [disjointed tail]` (currently the `+isol` feature is neither well defined nor well implemented, leading

to some marginal issues). To be less dependent on the current standard and implementations, the internal structure of how a disjointed tail is encoded is not exposed in this model.

3.8. Arabic cursive joining

- Context: [joining letter]
- Feature: `isol | init | medi | fina`
- Example 3.8.1: *`ك ك ك ك / iii` → `ك ك ك`
 - `<I+isol> <I+init I+medi I+fina>`
- Example 3.8.2: *`و و` → `و`
 - `<U+isol>`

Many letters are affected only by this context and are considered *simple*, while *complex* letters are affected by more contexts and are the major source of complications and issues. See the rest of this section and § 4. *Graphemes of complex letters* to understand complex letters.

Vowels are conventionally considered to have all four positional forms. Consonants typically don't have true `+isol` forms while `+init` forms are actually used when isolated. Also, the code chart representative glyphs are often different to `+isol` forms (see Example 3.1.2).

Locality. Local to adjacent letters; not intrasyllabic. Bounded by spaces, but "what is a relevant space" is also a question.

Current situation. The Unicode Standard has positional forms analyzed with a special model that is neither self-consistent nor consistent to the Arabic cursive joining model and the OpenType implementation. There is at least a proposal requesting changes of standardized variation names.

4. Graphemes of complex letters

Table attached on the last page.

Notes:

1. All vowels are listed in the table. Only *complex* consonants are listed, except for the *simple* consonant `ل` here for comparison.
2. Four major columns are for the four positional forms. For every positional form, if the column is split, the wider left column shows the default form, and the

narrower right column shows certain marked form.

3. Parentheses denotes additional automatic or manual graphemes not covered by the features in the table head. See notes for details.
4. Gray cells are invalid. Gray forms are selected historical forms worth mentioning here.
5. The aleph parts in vowel forms are highlighted with magenta.
6. Note the `+medi+beginningOfNonFirstStem` and (theoretical) `+fina+beginningOfNonFirstStem` forms of vowels have been broken down (the aleph parts removed) to plain `+medi` and `+fina` forms.
7. `E+medi`, `E+fina`, (`postAleph`): Reduced to toothless forms. Also see `E+isol` and `E+init` 's aleph-prepended and aleph-less forms to see `E` 's special interaction with aleph, while other vowels are not graphemically affected by the presence of aleph.
8. `E+isol` and `E+init` often shows a longer stem. What about `E+medi+beginningOfNonFirstStem`?
9. `I+medi` (`+postvocalic`), see § 3.4. *Underlying letters in diphthongs*.
10. `UE+isol`, (`manual`): A special form only used in some Mandarin transcriptions.
11. `L`: A typical simple consonant for comparison.
12. `N`: A typical complex consonant that has `+coda` and `+preDisjointedTail` forms.
13. `Q`, `G`, (`+masculine`): `Q` doesn't have `+coda` usage.
14. `T`, `D`, (`manual`): `T` and `D` have disambiguated forms. `T+medi+coda` and `T+fina` are used only in loanwords.
15. `Y+init`, `Y+medi`, (`manual`): `Y` has lexical variations from its historical form. See § 3.8. *Clarifying enclitics* for the cases in `[enclitic]`.
16. `Y+medi`, `Y+fina`, `W+medi`, `W+fina`: The postvocalic `I/U`-alike structures are encoded as vowels.
17. `J+isol+preDisjointedTail`: Consonants typically don't have true `+isol` forms. This `J+isol` is only used in an `beginningOfEnclitic` and `preDisjointedTail` situation, and is predictable without `beginningOfEnclitic`.