

# [5 JUNE DRAFT 1842] An analysis model for the Hudum Mongolian writing system

Author: **LIANG Hai** (梁海) <lianghai@gmail.com>

Date: 5 June 2017

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), **Orlog** (ᠣᠷᠯᠣᠭ 欧日乐克), **Tengis** (ᠲᠡᠩᠭᠢᠰ 天格思),  
**ZHENG Weizhe** (郑维喆), et al.

## Background


Although it was introduced as early as in 3.0 (September 1999), the Unicode Mongolian encoding still doesn't work as expected.

The Mongolian script is a unified script covering a number of derivative scripts with conflicting behaviors, and it involves complicated contextual processes comparable to that of Indic scripts. However, for such a complicated script, a formal set of rules never exists in either the Unicode Standard or the OpenType specification. As a result, inconsistent vendor implementations create great difficulties for the user community to migrate to the Unicode encoding.

Recently, efforts have emerged to patch the encoding without paying enough attention to analyzing the orthography and formalizing contextual shaping rules, despite the fact that the encoding, the orthography, and the contextual shaping rules are highly entangled. Therefore, a coherent analysis model is presented here for concerned parties to better understand the script behavior and requirements.

**Writing systems.** Roughly speaking, the Unicode Mongolian block covers the following writing systems (or considered derivative scripts), with the languages they serve specified in parentheses:

- Hudum (Mongolian)
  - Hudum Ali Gali (Sanskrit–Tibetan)
- Todo (Oirat)
  - Todo Ali Gali (Sanskrit–Tibetan)
- Manchu (Manchu)
  - Manchu Ali Gali (Sanskrit–Tibetan)
  - Sibe (Sibe)

In this document we use the term “Hudum” (commonly written as  *xudum* or *худам xudam*, an exonym from Oirat) to refer to the writing system commonly simply called “Mongolian” or “traditional Mongolian”, thus the writing system is explicitly distinguished from the unified script.

It’s common to find a largely the same letter is disunified because in a certain context it behaves differently in different writing systems.

**Diachronic diversity.** Hudum has evolved significantly since its early stages (which are often referred to as “*Uyghur Mongolian*”), leading to great diachronic diversity in terms of orthography. The Manchu writing system also has an *old Manchu* stage (close to the Hudum back then) used before its orthography reform.

Instead of recklessly trying to encode historical texts with modern phonetic analysis and countless visual patches, it’s crucial to draw a line between the modern orthography and historical ones.

**Orthographic depth.** On the other hand, synchronically Hudum shows a complicated correspondence between the written language (graphemes) and the spoken language (phonemes), ie, a great orthographic depth. Recognizing a letter or character is not straightforward. A pair of letters generally considered separate can look exactly the same in many contexts, leading to that users have to master the language and the orthography before being able to input text. Some experts actually consider Hudum could’ve been encoded visually for a better model.



The idea of phonetic encoding in the current model often leads to scholars’ absurd attempts at forcing a character to be shaped as another so a grapheme can be encoded as whatever phoneme desired.

**Other difficulties.** Grammar and orthography debates, etc.

## Introduction

The focus of this analysis model is set on the Hudum writing system for the modern Mongolian language, with certain historical forms mentioned when necessary. This model itself is designed to include a minimal amount of details of current technical implementations.








However, 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 specify any mechanisms other than the Arabic cursive joining.

A scheme based on the *China–Mongolia Joint Proposal of Mongolian Transliteration* (draft, 2005) is used for transliterating, eg,  *xöxexota* (ie, Hohhot). Letter names consistent with the Unicode character names are used in technical presentations, eg,  <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 only visual aids, while their characters extracted or copied from the PDF are not reliable.


## Graphematics and glossary

We refer to the script's inherent behavior as “*graphematics*”, distinguished from orthography. The term “orthography” is commonly also used in Unicode–OpenType contexts to refer to the language-independent script processes between typography and spelling, because spelling is usually not Unicode's concern. But the language-involving spelling-orthography happens to be significant for the Mongolian encoding issues.







**Letter** — V Vowel letters (native *a, e, i, o/u, ö/ü*; loan *ě*) or C consonant letters (native *ʾ, n, ng, b, p, x/g, q, γ, m, l, s, š, t/d, č, j, y, r, w*; loan *f, k, ɣ, c, z, h, ž, lh, ž, ĉ*).

- **Cursive joining** — Almost all letters join to either side (dual-joining), therefore may have all positional forms: *isolate, initial, medial, and final*.
- **Disjointed tail** ( *čaculy{a}*, or a commonly seen transcription of the Khalkha pronunciation, *tsatslag*) — {*a*} and {*e*}, unusually non-joining (ie, always isolate) forms of *a* and *e*.
- It's important to understand a *disjointed tail* is different to a *left tail*  (a visual part of a subgrapheme variation of the *right tail* , eg, \* *ba* →  → ). Although often visually similar (eg, in the typeface style of *baiti*), they are not the same grapheme and show different graphematic behaviors. Note the *left tail*  is zero teeth while the *disjointed tail* has one tooth.
- From a phonetic point of view, seven native vowel letters (*a, e, i, o, u, ö, ü*) are generally recognized. But graphematically, the *U* letters (*u, ü*) are (nearly) always the same to their *O* counterparts (*o, ö*); and *A* letters (*a, e*) are also often the same, so do *O* letters (*o, ö*).

Therefore many consider that graphematically Hudum only have five (*a, e, i, o/u, ö/ü*) or even three native vowel letters (*a/e, i, o/ö/u/ü*).




- Certain consonant letters are not graphematically differentiated at all. Phonemes /x, g/ have separate structures (complementary distribution) for feminine and masculine genders (*x, q; g, γ*), however their feminine forms *x* and *g* are always the same, thus only three letters are recognized for /x, g/ phonemes: *x/g, q, γ*. Eg,  *xe/ge qa ya*.
- To disambiguate letters in loanwords, certain special forms are employed *ë, o, ü, t, d, k, h...*

**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. Eg, *i.ni, in.li*.

- Some consonant letters have special forms in a coda position. Some consonant letters natively don't appear in certain syllabic positions due to historical phonological limit of the Mongolian language.
- When an **I+medi** (a single *long tooth* ) follows another vowel in a syllable, it becomes two long teeth , except when the preceding vowel already ends with a long tooth:  
**I(+init|+medi)**   and **(OE|UE)(+init|+medi)+firstVowel)**  .
- The disjointed tail is a part of a syllable although not cursive-joining.

**Stray consonant** — A consonant *C* not belonging to any syllable therefore doesn't participate in any syllable-involving processes, leading to the default form always. Eg, *⟨g⟩.ram* (cf. *g<sup>+</sup>am, gem*).

**Word** — Words are separated with spaces. A word is either a **stem** sequence (one or more stems because a word may be a compound) or an **enclitic**.

- A sequence of **((syllable) | (stray consonant))+** (one or more instances of either syllables or stray consonants) forms either a stem or an enclitic.
- An enclitic is grammatically similar to suffixes but not joined to a stem. Eg,  *un*. Zero or more enclitics might follow a stem sequence, **(stem)+ (enclitic)\***, forming a *clitic group*. A special scope is significant in Hudum graphematics: **(stem) (enclitic)\***, ie, it (re)starts at the beginning of every stem and continues to the last enclitic.
- The disjointed tail only appears at the end of a word, and is changed to a regular vowel *a* or *e* when more letters is appended. Eg,  *qar{a}\_qorom* →  *qaraqorom*.

Stem-depending processes:

- **Aleph** — A prepending silent consonant letter ' required by vowel letters to start a stem. Eg,  $\aleph$  'a,  $\eta$  'e,  $\aleph$  'i,  $\eta$  'o.
- The vowel letter *e* shows exceptional interaction when prepended by an *aleph*: it reduces to *toothless* forms (nothing at all  $\emptyset$  or a piece of *stem* • when medial, and the *left tail*  $\eta$  when final). The consonant letter *h* behaves like a vowel in terms of the alternation of its *aleph*, although its non-word-beginning but stem-beginning form is unattested so far.
- Note the so-called *aleph* in this model is cognate with vowel letters *a* and *e* — all of them are descended from a historical *aleph*. Actually they still share the same set of graphemes. The common form of an *aleph* is a *cap*  $\aleph$  when initial, while common forms of vowel *a* and *e* are a *tooth*  $\aleph$  when medial and a *right tail*  $\aleph$  when final. The cap  $\aleph$ , the tooth  $\aleph$ , and the right tail  $\aleph$  are positional forms of the same underlying structure, but this alternation between a cap  $\aleph$  and a tooth  $\aleph$  is often overlooked. Eg,  $\aleph\aleph\aleph$  'odo,  $\aleph\aleph\aleph\aleph$  'altanodo;  $\aleph\aleph$  'ača,  $\aleph\aleph$  ača.
- Writing systems other than Hudum might require an *aleph* for hiatuses or diphthongs.
- **First vowel** — Certain vowel letters have special forms when they're the first vowel in a stem.
- **Vowel harmony** — The Mongolian language has a vowel harmony system of more than one features, among which the *gender* harmony is significant in Hudum.
  - Vowel genders: *e* (also  $\ddot{e}$ ) and  $\ddot{o}/\ddot{u}$  are *feminine*; *i* is *neuter*; *a* and *o/u* are *masculine*. Note how letter pairs of opposite genders (*a-e*, *o/u-ö/ü*) are often hard to tell apart. The gender harmony is also considered of *ATR* (advanced tongue root, with *feminine* being +*ATR*) or *tenseness* (with *masculine* being +*tense*).
  - Typically a stem is either masculine (having masculine and optional neuter vowels) or feminine (having feminine and optional neuter vowels; or only neuter vowels). Therefore a stem can be analyzed as feminine by default and becomes masculine when there're masculine vowels. Non-harmonious words (eg, loanwords) don't have stem-wide harmony but usually still follow a syllable-wide harmony.
  - Considering a vowel harmony context is feminine by default, and the stray forms of *x* and *g* are also their feminine forms, *q* and *γ* are analyzed to have marked masculine forms.
  - Enclitics, as native and grammatical structures, are always word-wide harmony. And the gender of an enclitic agrees with the preceding stem or enclitic.
  - The phonemes /x, g/ have feminine forms *x/g* and *g*, as well as masculine forms *q* and *γ*. Inside a syllable (note /x/ doesn't appear in a coda position):

- The feminine forms are used before or after a feminine vowel letter. Eg,  $x/g + (e \mid \ddot{o}/\ddot{u} \mid \ddot{e})$ ,  $(e \mid \ddot{o}/\ddot{u} \mid \ddot{e}) + g$ .
- The masculine forms are used before or after a masculine vowel letter. Eg,  $(q \mid \gamma) + (a \mid o/u)$ ,  $(a \mid o/u) + \gamma$ .
- The feminine forms are used before a neuter vowel letter *i*. Eg,  $x/g + i$ . The masculine forms are never used before *i* in modern Hudum.
- The feminine or masculine forms are used after a neuter vowel letter *i*, depending on the actual gender of the stem/enclitic. Eg,  $e \dots i + g \dots e$ ,  $a \dots i + \gamma \dots a$ . In other words, /ig/ is the only case that requires inter-syllable gender passing.

	isolate	initial	medial	final
<i>a</i>	( ɤ )	( ɹ )	ɹ	ɤ
<i>e</i>	( ɤ )	( ɹ )	ɹ ∅†	ɤ ɹ†
<i>{a}/{e}</i>	ɤ			
<i>i</i>	( ɿ )	( ɹ )	ɹ ɿ‡	ɿ
<i>o/u</i>	( ʊ )	( ʊ )	ʊ	ʊ ʊ*
<i>ö/ü</i>	( ʊ )	( ʊ )	ʊ ʊ*	ʊ ʊ*
<i>ë</i>			ɿ	ɿ

Notes: (...) — This positional form only appears in enclitics. \* First vowel. † Post-*aleph*. ‡ Postvocalic.

### Complication introduced by the Unicode–OpenType model

- The *aleph* is considered part of vowels therefore has too be shown contextually with additionally encoded context marks.
- Letters  $x/g$ ,  $q$ ,  $\gamma$  are encoded purely phonetically as two characters,  $x$  and  $g$ . Eg,  $\text{ḡ ḡ ḡ ḡ}$   $xe\ ge\ x^+a\ g^+a$ . Due to this unification, the shaping of  $x$  and  $g$  depends on vowel harmony.

## Shaping stages

**Input:** A sequence of Unicode characters.

**I. Automatic graphemes:** Grapheme 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 behavior, including cursive joining and some more contextual mechanisms. Eg, a fake word *antaba* for demonstrating, \*ᠠᠨᠲᠤᠪᠠ → \*ᠠᠨᠲᠤᠪᠠᠭᠤ.

See § 3. *Automatic grapheme features* for a detailed breakdown of this stage.

**II. Manual graphemes:** Other grapheme 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 stage I \*ᠠᠨᠲᠤᠪᠠᠭᠤ → \*ᠠᠨᠲᠤᠪᠠᠭᠤᠲᠤ, 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 grapheme variations (cf. Arabic ﻻ <LAM ALEF>). Eg, *antaba* outcome from the stage II \*ᠠᠨᠲᠤᠪᠠᠭᠤᠲᠤ → ᠠᠨᠲᠤᠪᠠᠭᠤᠲᠤ, 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 grapheme features

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

Sections	Automatic grapheme 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 below. 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: ጥጥ *#in#in*
  - cf. ጥጥ *#in\_#in*, ጥጥ *#i.nin*
- Example 3.1.2: ለጥጥ *#al.tan#o.do*
  - cf. ለጥጥ *#al.tan\_#o.do*, ለጥጥ *#al.ta.no.do*
- Example 3.1.3: ህጥጥ *#bu.man#er.de.ni*
  - cf. ህጥጥ *#bu.man\_#er.de.ni*, ህጥጥ *#bu.ma.ner.de.ni*
- Example 3.1.4: ህጥጥ *#ba<sup>1</sup>.tu#mō<sup>1</sup>ng.xe*
  - cf. ህጥጥ *#ba<sup>1</sup>.tu\_#mō<sup>1</sup>ng.xe*, ህጥጥ *#ba<sup>1</sup>.tu.mōng.xe*
- 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* ጥ while an Aleph+medi is a *tooth* ጥ.

- 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 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 prepended aleph, but the N+medi+coda clearly shows a stem boundary.
- See Example 3.1.4 for how a stem boundary also affects the vowel count.

**Locality.** Possibly local if stem boundaries are marked in encoding; not intrasyllabic because requiring information outside of a syllable; currently not achieved with automatic grapheme features due to lack of a stem boundary mark.



**Current situation.** The Unicode Standard employs FVSes (see § 2. *Shaping stages*, stage 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.

If stem boundaries themselves (instead of their effects) are marked in encoding, although the rule of prepending aleph is actually simple (beginning of stem), in the Unicode Standard, the majority of cases (beginning of word, therefore also beginning of stem) is already achieved with the +isol and +init forms (see § 3.1. *Arabic cursive joining*), then the concerning cases here (not beginning of word, but beginning of stem) have to be handled separately, thus `beginningOfNonFirstStem`.

### 3.2. Vowel count

- Context: `[vowel count]`
- Feature: `firstVowel`
- Example 3.2.1: \*`ٲٲ ٲٲ ٲٲ ٲٲ nu¹_nu¹nu_nü¹_nü¹nü` → `ٲٲ ٲٲ ٲٲ ٲٲ`
  - `<N+init U+fina+firstVowel> <N+init U+medi N+medi U+fina> <N+init UE+fina+firstVowel> <N+init UE+medi+firstVowel N+medi UE+fina>`
  - U+medi doesn't have a special +firstVowel form.
- Note: "1" marks a first vowel.

**Locality.** Not local, but possibly derivable from a limited number of letters since only need to know `firstVowel`; not intrasyllabic because requiring information outside of a syllable.

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

### 3.3. Enclitics

- Context: `[enclitic]`
- Feature: `beginningOfEnclitic`
- Example 3.3.1: \*`ٲٲ ٲٲ #al ^un` → `ٲٲ ٲٲ`
  - `<A+init L+fina space U+init+beginningOfEnclitic N+fina>`
- Example 3.3.2: \*`ٲٲ ٲٲ #ül ^üü` → `ٲٲ ٲٲ` (cf. `ٲٲ ٲٲ #ül ügei`)
  - `<UE+init+firstVowel L+fina space UE+init+beginningOfEnclitic UE+fina>`
- Example 3.3.3: \*`ٲٲ ٲٲ #an ^a` → `ٲٲ ٲٲ`
  - `<A+init N+fina+coda space A+isol+beginningOfEnclitic>`
- Example 3.3.4: \*`ٲٲ ٲٲ #an ^{a}` → `ٲٲ ٲٲ` (cf. \*`ٲٲ ٲٲ #an{a}` → `ٲٲ ٲٲ`)
  - `<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 *stem sequence* 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, <sup>^</sup>iy... and <sup>^</sup>yi..., show lexical variations from their historical forms, therefore are not actually part of predictable [enclitic] behavior. 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 behavior, instead of pure grammatical categorization (which is always controversial). In existing specs, based on certain grammatical analysis, ᠢᠭᠡᠢ ügei is commonly categorized as NNBS applicable while ᠠᠭᠢᠢ <sup>^</sup>üü not, leading to irregular effects of NNBS. But by examining orthographical behavior, it's clear that ᠢᠭᠡᠢ ügei is not an [enclitic] while ᠠᠭᠢᠢ <sup>^</sup>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 behavior, 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 (NNBS) is employed to both provide the white space and suggest special shaping behavior 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: \*ᠢᠨᠯᠢ in.li → ᠢᠨᠯᠢ (cf. ᠢᠨᠢ i.ni)
- Example 3.4.2: ᠭᠠᠷᠠᠮ ᠭᠠᠷᠠᠮ ᠭᠠᠷᠠᠮ (cf. \*ᠭᠠᠷᠠᠮ ᠭᠠᠷᠠᠮ → ᠭᠠᠷᠠᠮ)
- Note: "." marks a syllable boundary; "<..." marks a stray consonant.

**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 rendering from fonts.

### 3.5. Vowel harmony

- Context: [vowel harmony]
- Feature: masculine
- Example 3.5.1: \*ᠭᠠᠯᠠᠯ ᠭᠠᠯᠠᠯ → ᠭᠠᠯᠠᠯ (cf. ᠭᠡᠯᠠᠯ ᠭᠡᠯᠠᠯ)
- Example 3.5.2: ᠭᠠᠯᠠᠯᠠᠯ ᠭᠠᠯᠠᠯᠠᠯ ᠭᠠᠯᠠᠯᠠᠯ (cf. \*ᠭᠠᠯᠠᠯᠠᠯ ᠭᠠᠯᠠᠯᠠᠯ)

- <Q+init OE+medi+firstVowel Q+medi E+medi Q+medi+masculine O+medi T+medi A+fina>

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

The reference algorithm below is similar to the rules mentioned in the Core Spec but this is explicitly restricted to a local context (without vowel harmony at a distance). Note this algorithm doesn't require syllabification, and it leaves masculine *ig*<sup>+</sup> to be manually requested.

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

If a line break happens inside a word, there's no way to acquire gender info from the last line?

### 3.6. Diphthongs

- Context: [diphthong]
- Feature: postvocalic
- Example 3.6.1: \*ሳይን *sain* → ሳይን
- cf. ሳይ *sai*, ሳይን *sin*
- Example 3.6.2: \*ኒውኒን *nü<sup>1</sup>i.nüin* → ኒውኒን
- cf. ኒውኒን *nü<sup>1</sup>.nün*, ኒውኒን *nu<sup>1</sup>i.nuin*, ኒውኒን *nu<sup>1</sup>.nun*

**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: \*ሳይን *il{a}* → ሳይን (cf. ሳይ *ila*)
- <I+init L+fina [disjointed tail]>
- Example 3.7.2: \*ሳይን *in{a}* → ሳይን (cf. ሳይ *ina*)
- <I+init N+fina+preDisjointedTail [disjointed tail]>
- Example 3.7.3: \*ሳይን *iw{a}* → ሳይን (cf. ሳይ *iwa*)
- <I+init W+fina+preDisjointedTail [disjointed tail]>

- Note: "{...}" marks a disjointed tail.

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.

**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: \* $\text{ك ك ك ك } i \text{ } iii \rightarrow \text{ك ك ك ك}$ 
  - `<I+isol> <I+init I+medi I+fina>`
- Example 3.8.2: \* $\text{ا } u \rightarrow \text{ا}$ 
  - `<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 **L** 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. Asterisks denote additional automatic or manual graphemes not covered by the features in the table head. See notes for details.
4. Gray cells are invalid.
5. Superscript forms are historical forms.
6. The aleph parts in vowel forms are highlighted in cyan.
7. **E+isol** and **E+init** often shows a longer stem. What about **E+medi+beginningOfNonFirstStem**?
8. **I+medi †+postvocalic**, see § 3.4. *Underlying letters in diphthongs*.
9. **UE+isol**, \*manual: A special form only used in some Mandarin transcriptions.
10. **L**: A typical simple consonant for comparison.
11. **N**: A typical complex consonant that has **+coda** and **+preDisjointedTail** forms.
12. **Q**, **G**, (**+masculine**): **Q** doesn't have **+coda** usage.
13. **T**, **D**, \*manual: **T** and **D** have disambiguated forms. **T+medi+coda** and **T+fina** are used only in loanwords.
14. **Y+init**, **Y+medi**, \*manual: **Y** has lexical variations from its historical form. See § 3.8. *Clarifying enclitics* for the cases in **[enclitic]**.
15. **Y+medi**, **Y+fina**, **W+medi**, **W+fina**: The postvocalic **I/U**-like structures are encoded as vowels.
16. **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**.