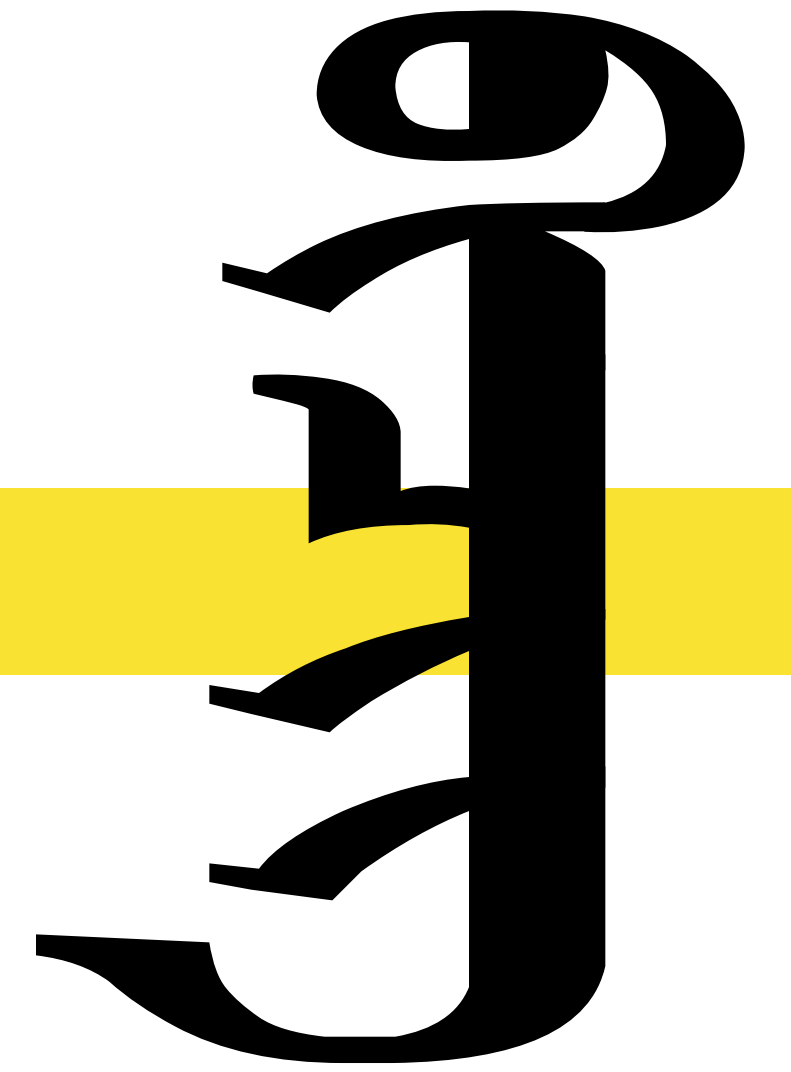




The Mongolian script:



What's going on?!

梁海 · Liang Hai · ल्यांग हाइ · ལྷཱུང་ རྩེ

lianghai@gmail.com

11 September 2018, IUC #42

* Get the latest revision from ↗ lianghai.github.io/mongolian *

Note

The views expressed by the speaker in this talk are his own and are NOT meant to reflect those of the Unicode Consortium or the Unicode Technical Committee.

Agenda

- I. A crash course on the script
- II.** The Unicode Mongolian, an encoding model from hell
- III. What exactly are not working?
- IV. Tough lessons learned
- V.** Ongoing efforts, and how to participate

• Part I •

A crash course on the script

What is the Mongolian script? Writing systems and features.

I. Crash course: *Origin*

Aramaic

Sogdian

Old Uyghur

Mongolian, initially “Uyghur Mongolian”

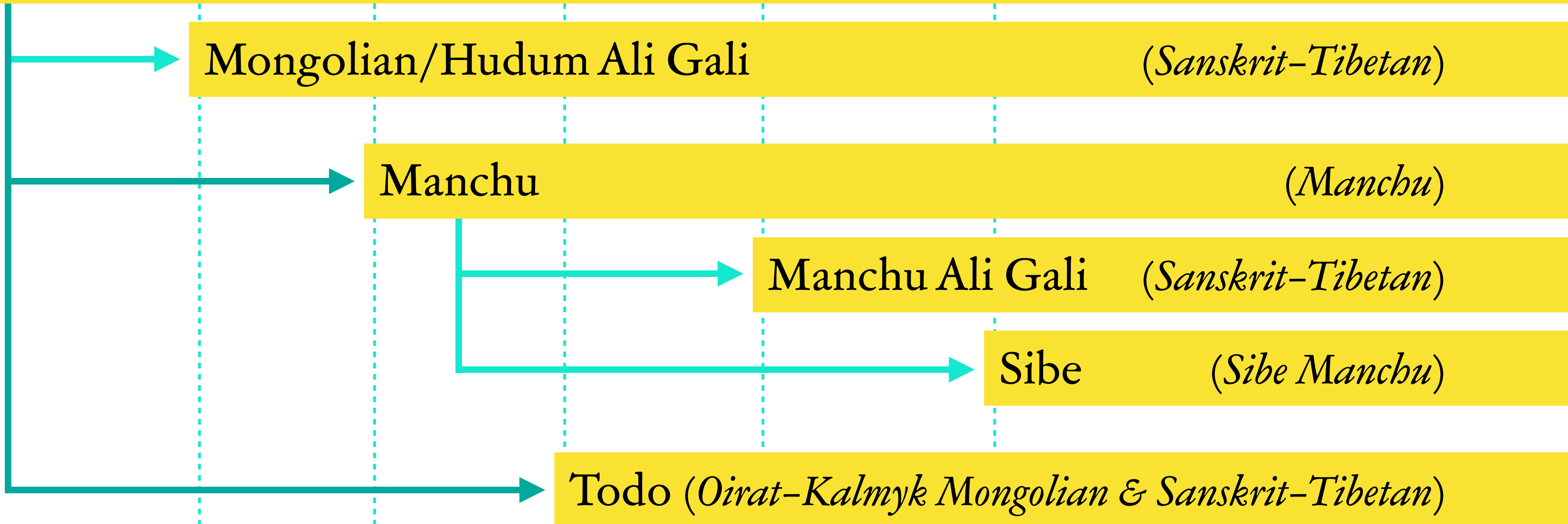
early 13th century

I. Crash course: *Writing systems & languages*

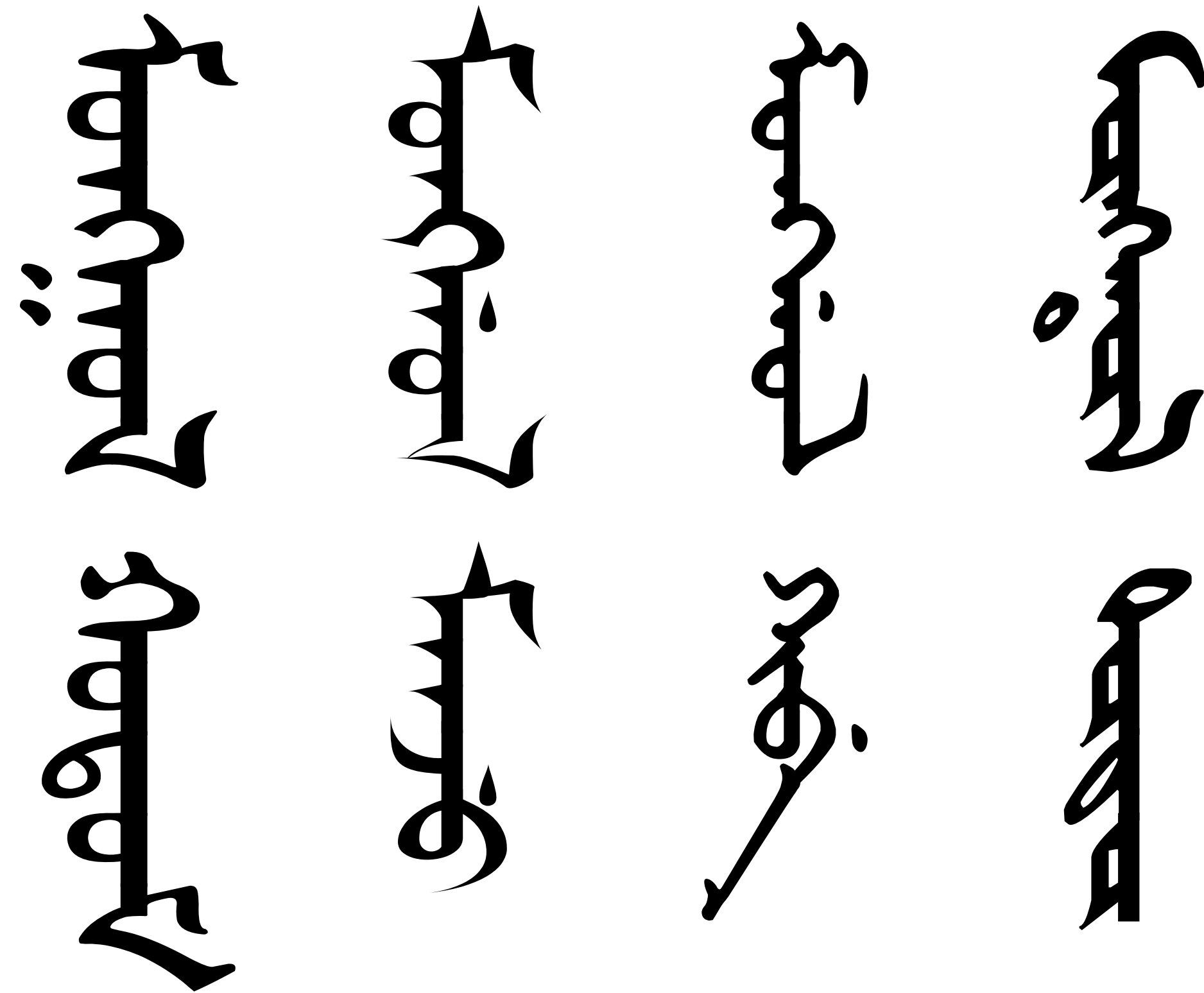
Mongolian/Hudum

(Mongolian)

..... early 13th late 16th . early 17th . mid-17th ... mid-18th mid-20th ..



I. Crash course: *Writing systems & languages* [cont.]



[→] Hudum, Manchu, Sibe, and Todo, in some typical styles:

mongol budum | ... *manju* | ... *sibe* | ... *todo*

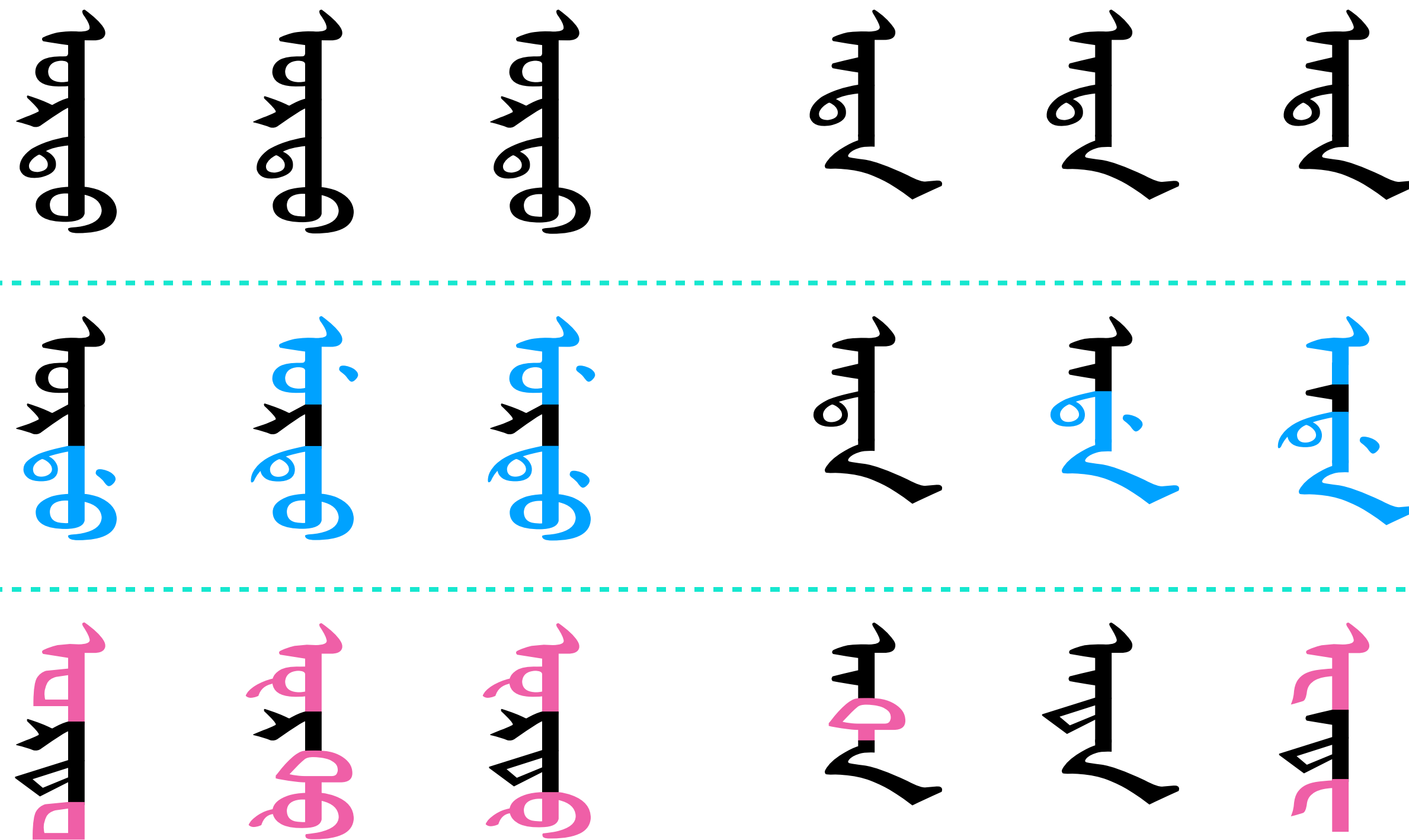
I. Crash course: *Writing systems & languages* [cont.]

Three rows of vertical script characters, each row containing six characters. The first row shows three variations of a character with a circular base and a vertical stem, followed by three variations of a character with a hook-like base and a vertical stem. The second row shows three variations of the first character with a dot above the stem, and three variations of the second character with a dot above the stem. The third row shows three variations of the first character with a horizontal bar at the top, and three variations of the second character with a horizontal bar at the top. Two horizontal dashed lines separate the rows.

[↓] Hudum, Manchu–Sibe, and Todo, normalized to the same style:

ordo urtu urdu | ata ada ende

I. Crash course: *Writing systems & languages* [cont.]



[↓] Hudum, **Manchu–Sibe**, and **Todo**, normalized to the same style:

ordo urtu urdu | ata ada ende

I. Crash course: *Writing systems & languages* [cont.]

Writing systems:

- **Hudum** and Hudum Ali Gali
- **Manchu–Sibe** and Manchu Ali Gali
- **Todo**

Served languages:

- Mongolian, incl. Oirat–Kalmyk
- Manchu–Sibe
- Sanskrit–Tibetan


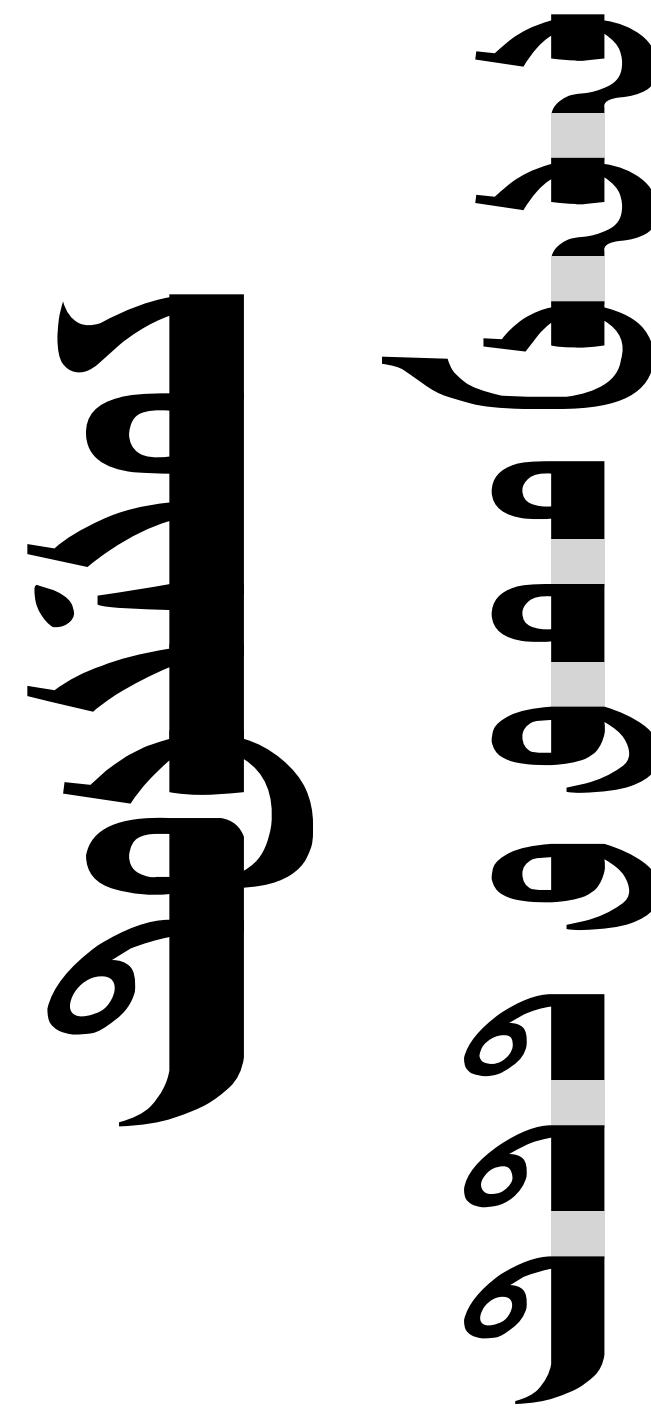
Also, note some marginal cases:

Manchu–Sibe for *Daur*, Hudum for *Evenki*, and **Vagindra** for *Buryat Mongolian*.

I. Crash course: *General features*

Inherited from Aramaic ~ Sogdian:

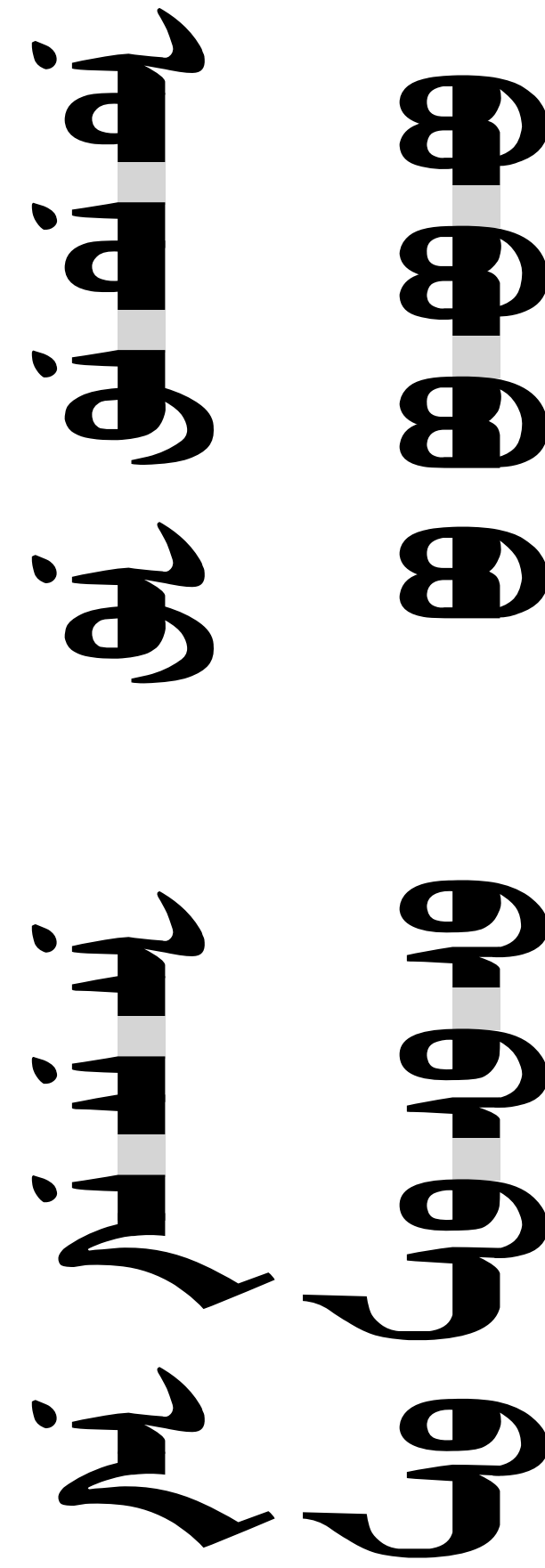
- **Cursive**
 - Largely dual-joining.
 - cf. Arabic
- Bowed consonants



I. Crash course: *General features* [cont.]

Inherited from Aramaic ~ Sogdian:

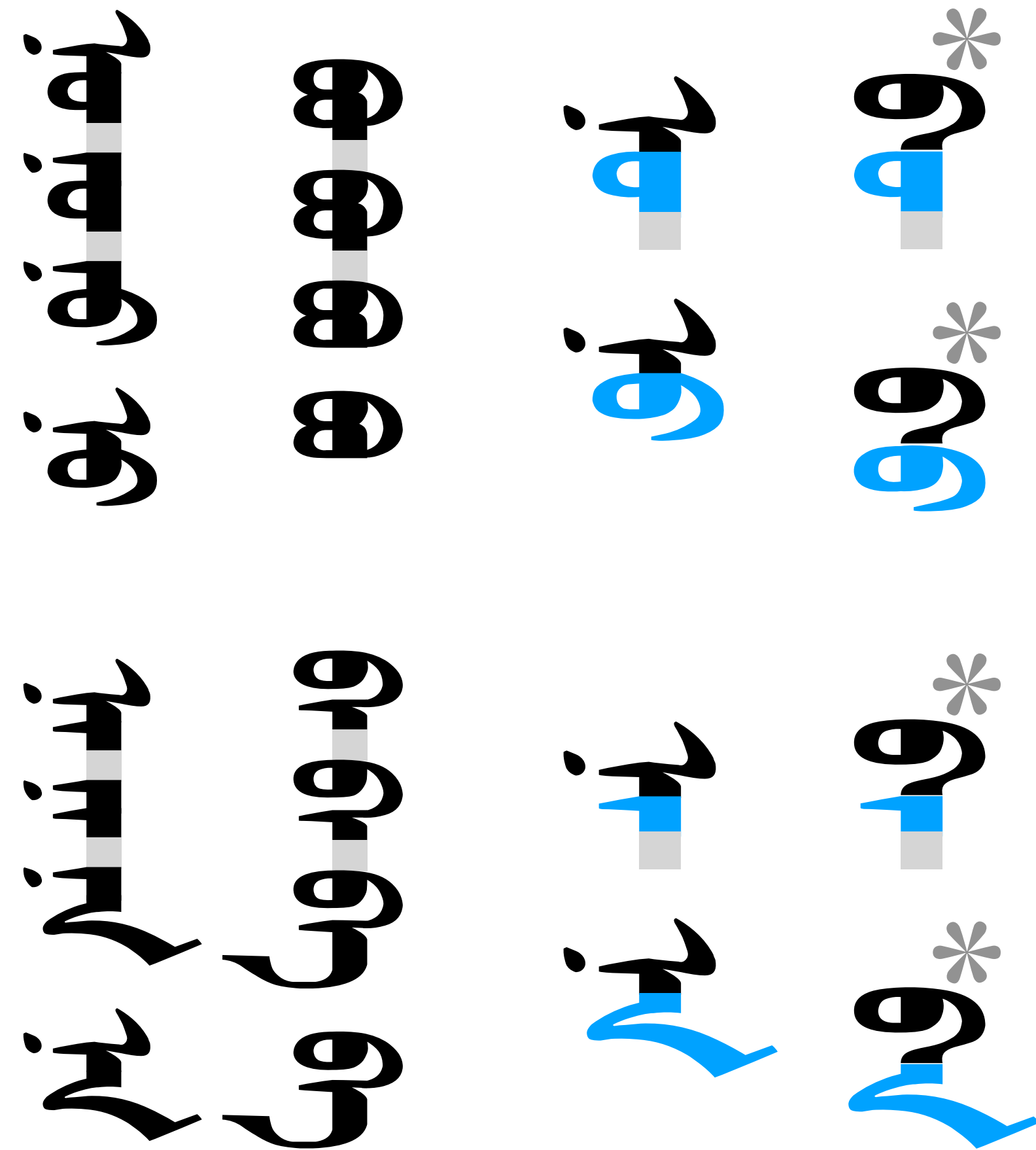
- Cursive
- Bowed consonants



I. Crash course: *General features* [cont.]

Inherited from Aramaic ~ Sogdian:

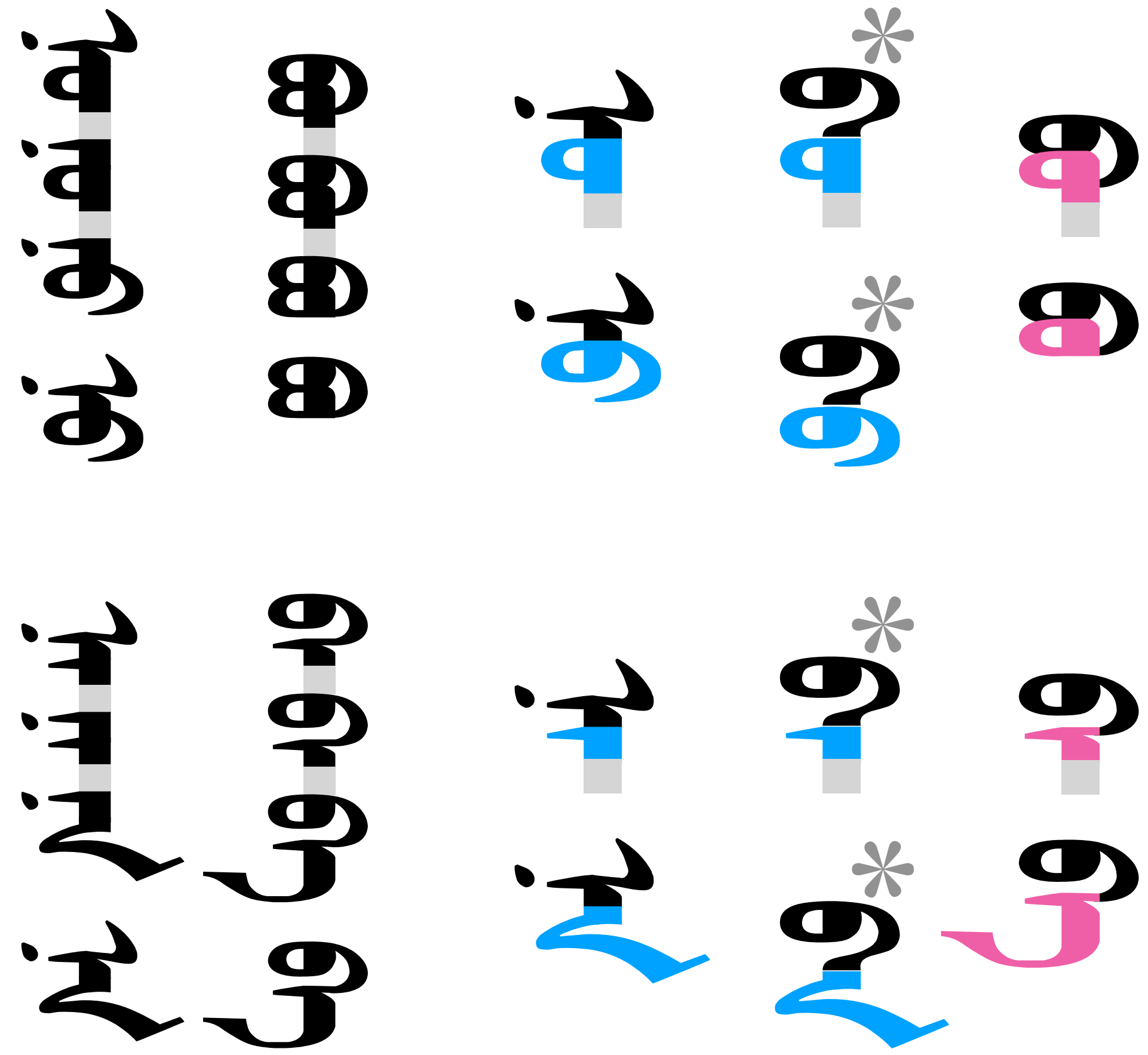
- Cursive
- Bowed consonants



I. Crash course: *General features* [cont.]

Inherited from Aramaic ~ Sogdian:

- Cursive
- Bowed consonants
 - [fun fact] A bare left tail is not a grapheme (unless in Ali Gali usages), while *tooth + left tail* as a whole is a contextual allograph of positional allographs *tooth* and *right tail*.



I. Crash course: *General features* [cont.]

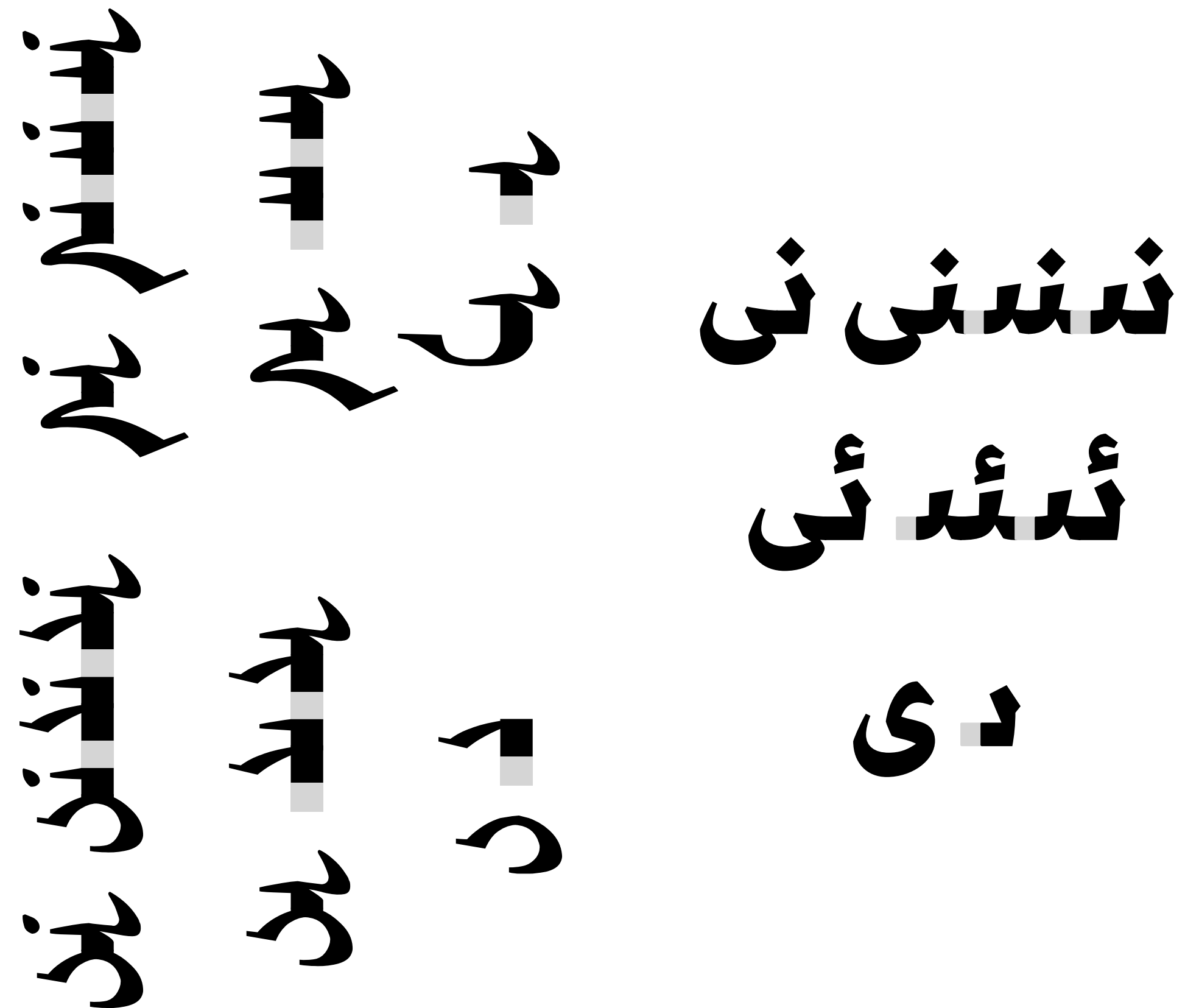
Inherited from Sogdian ~ Old Uyghur:

- **Vertical writing** ↓•→
 - Originated from ←•↓ being rotated 90° counterclockwise.
 - ↓•→-*in-narrow-column* or →•↓ as fallback in horizontal writing.
- **True alphabet**
 - (+ *dual-joining* =) Consonant letters seldom appear on their own, but are usually written in internally joined syllables.

I. Crash course: *General features* [cont.]

Invented during Old Uyghur ~ Mongolian:

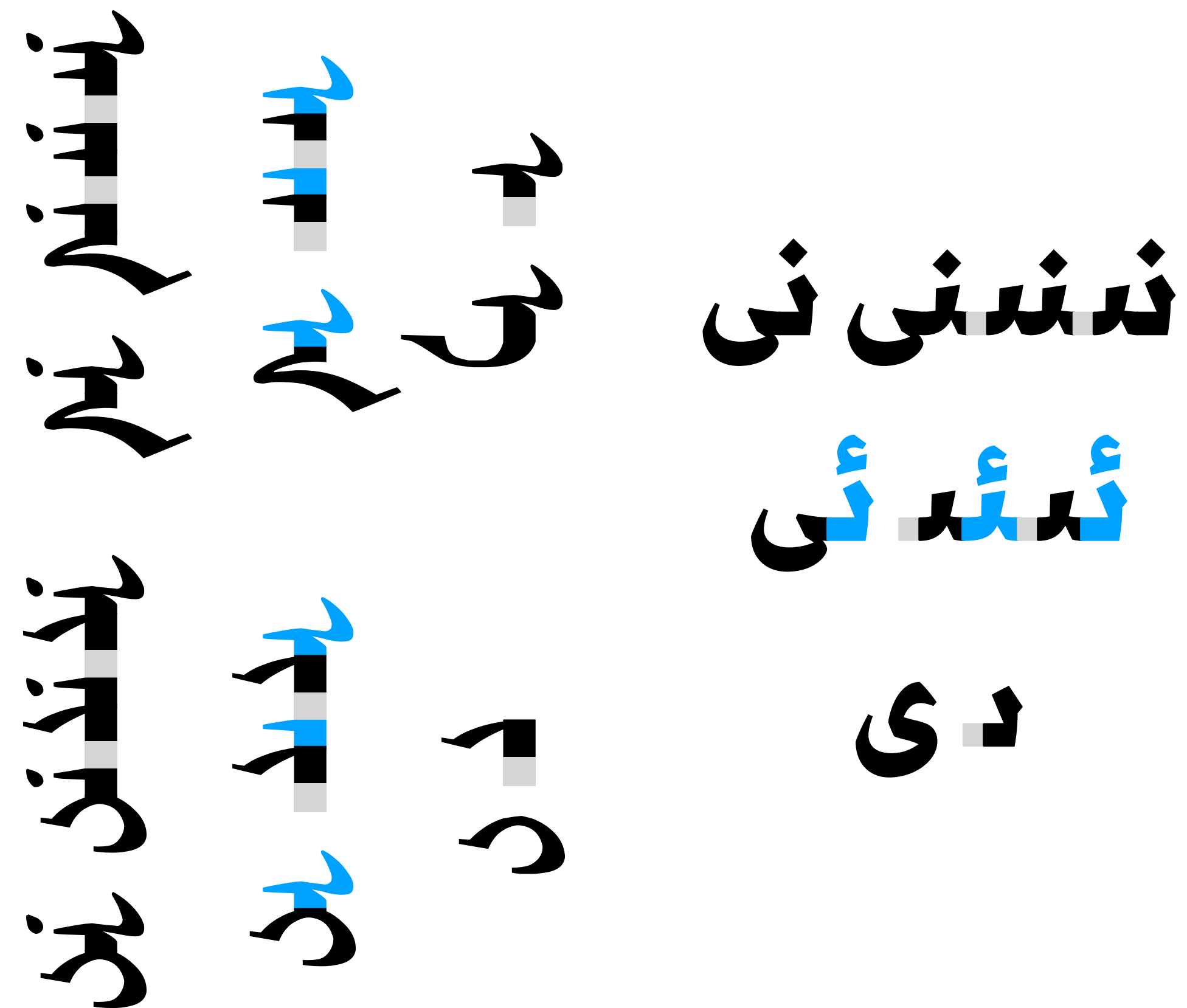
- Syllable onset placeholder (*aleph*)
 - cf. Uyghur ئ
- Syllable coda forms (*n, g, d...*)
- *Vowel harmony class*–specific consonants
- Phonetic letters/syllables



I. Crash course: *General features* [cont.]

Invented during Old Uyghur ~ Mongolian:

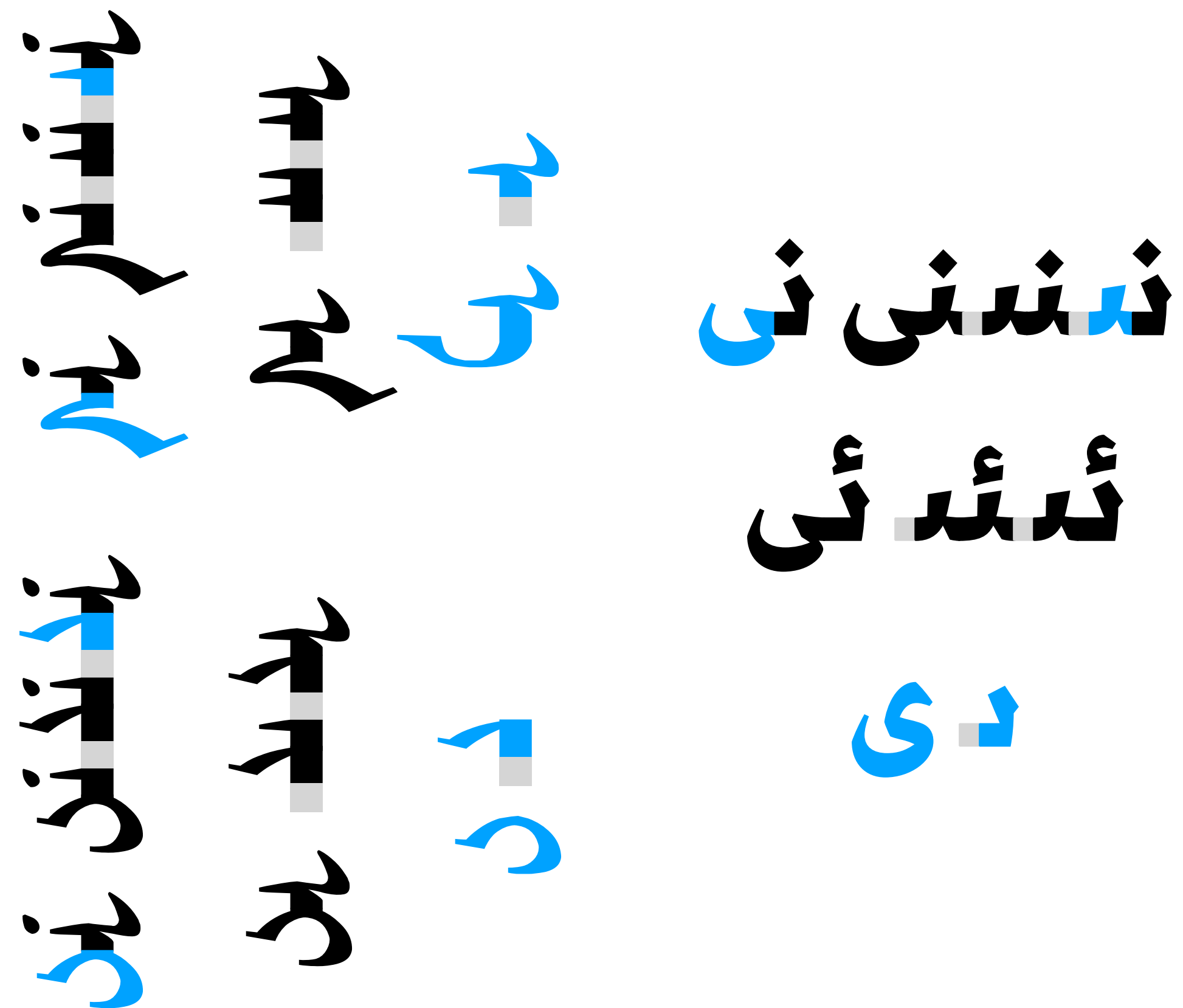
- Syllable onset placeholder (*aleph*)
 - cf. Uyghur ئ
 - [fun fact] The *crown* and the *tooth* are positional allographs to each other.
- Syllable coda forms (*n, g, d...*)
- *Vowel harmony class*–specific consonants
- Phonetic letters/syllables



I. Crash course: *General features* [cont.]

Invented during Old Uyghur ~ Mongolian:

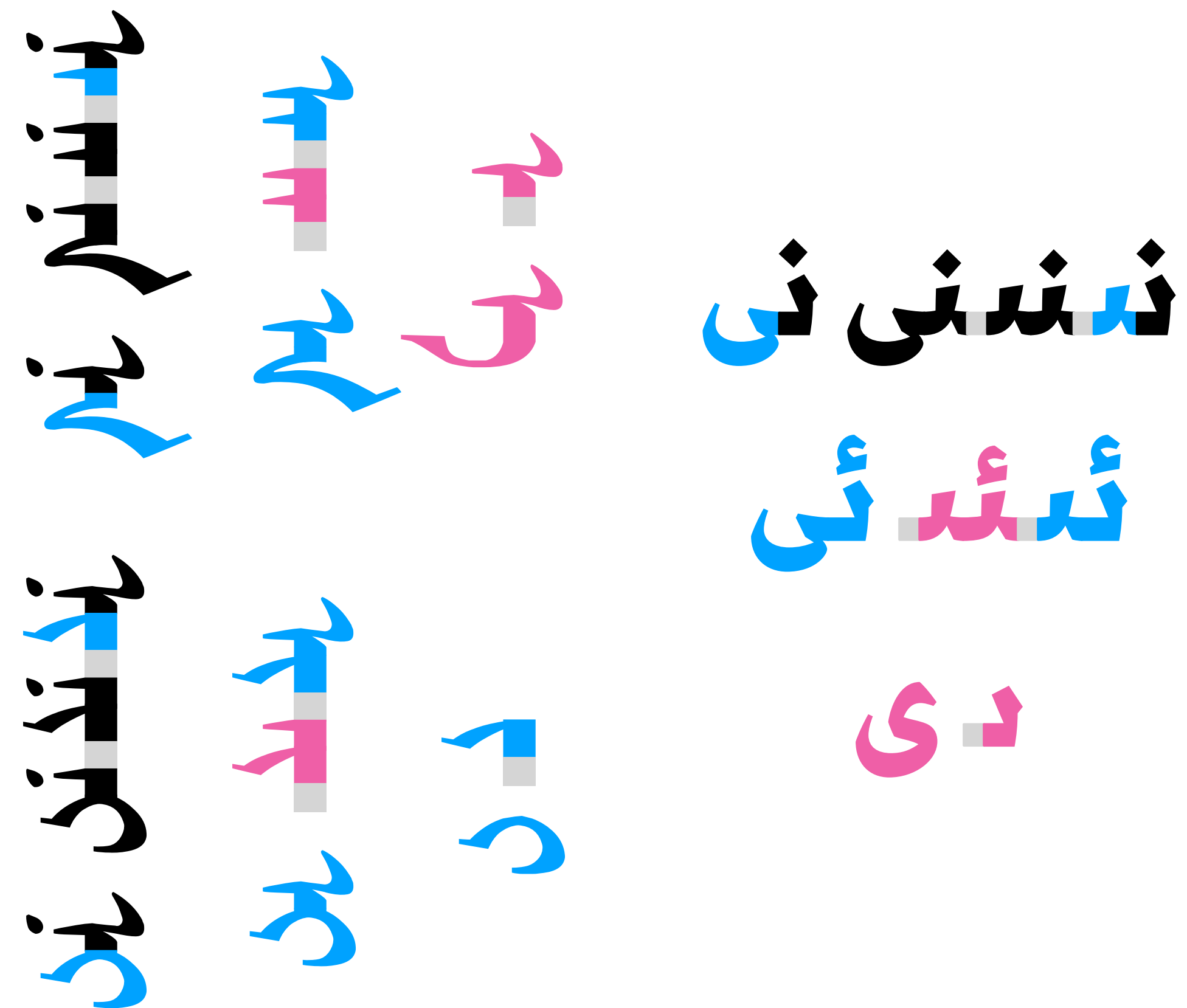
- **Syllable onset placeholder** (*aleph*)
 - cf. Uyghur ئ
 - [fun fact] The *crown* and the *tooth* are positional allographs to each other.
- Syllable coda forms (*n, g, d...*)
- *Vowel harmony class*–specific consonants
- Phonetic letters/syllables



I. Crash course: *General features* [cont.]

Invented during Old Uyghur ~ Mongolian:

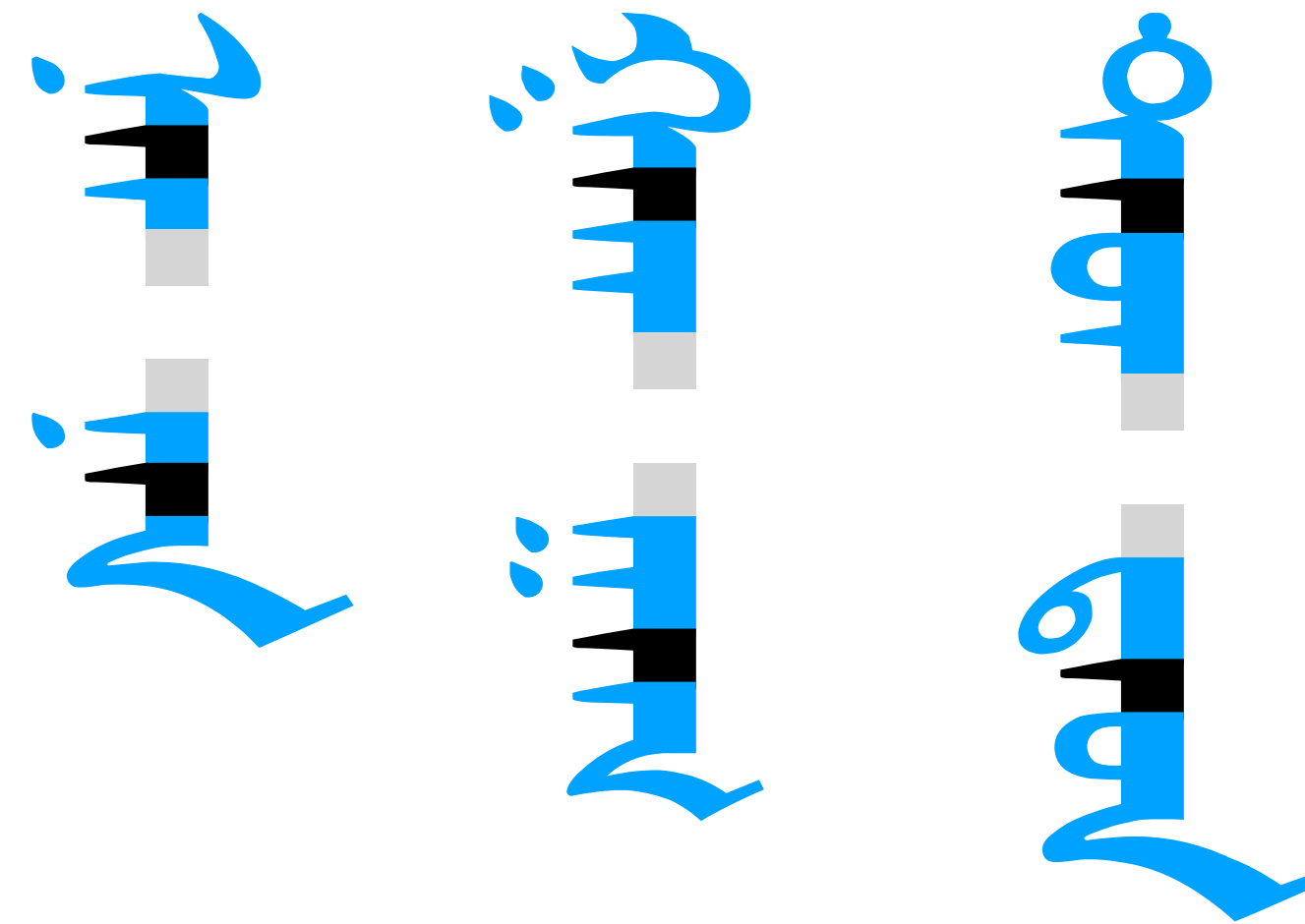
- **Syllable onset placeholder** (*aleph*)
 - cf. Uyghur ئ
 - [fun fact] The *crown* and the *tooth* are positional allographs to each other.
- Syllable coda forms (*n, g, d...*)
- *Vowel harmony class*–specific consonants
- Phonetic letters/syllables



I. Crash course: *General features* [cont.]

Invented during Old Uyghur ~ Mongolian:

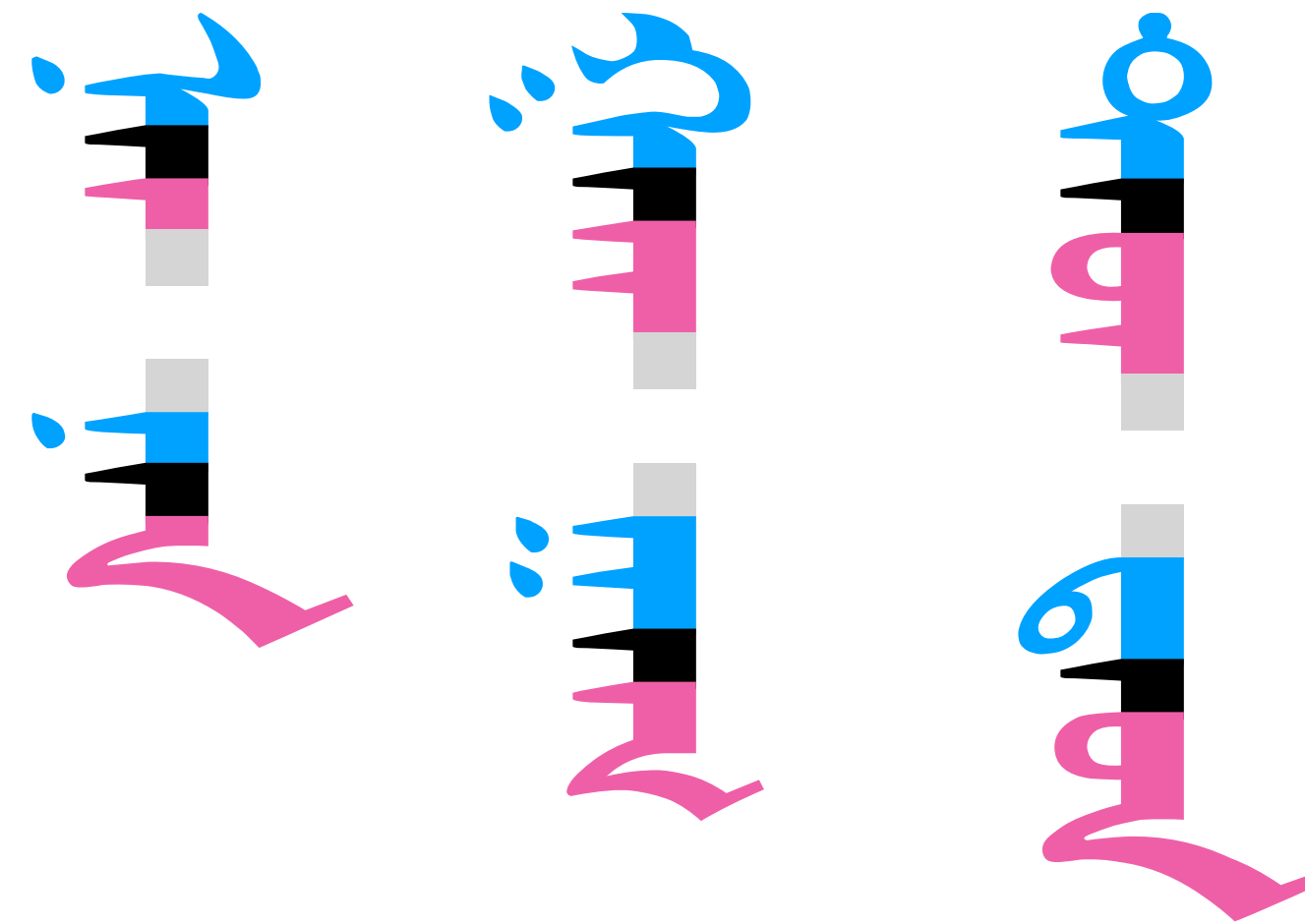
- Syllable onset placeholder (*aleph*)
- **Syllable coda forms** (*n, g, d...*)
- *Vowel harmony class*–specific consonants
- Phonetic letters/syllables



I. Crash course: *General features* [cont.]

Invented during Old Uyghur ~ Mongolian:

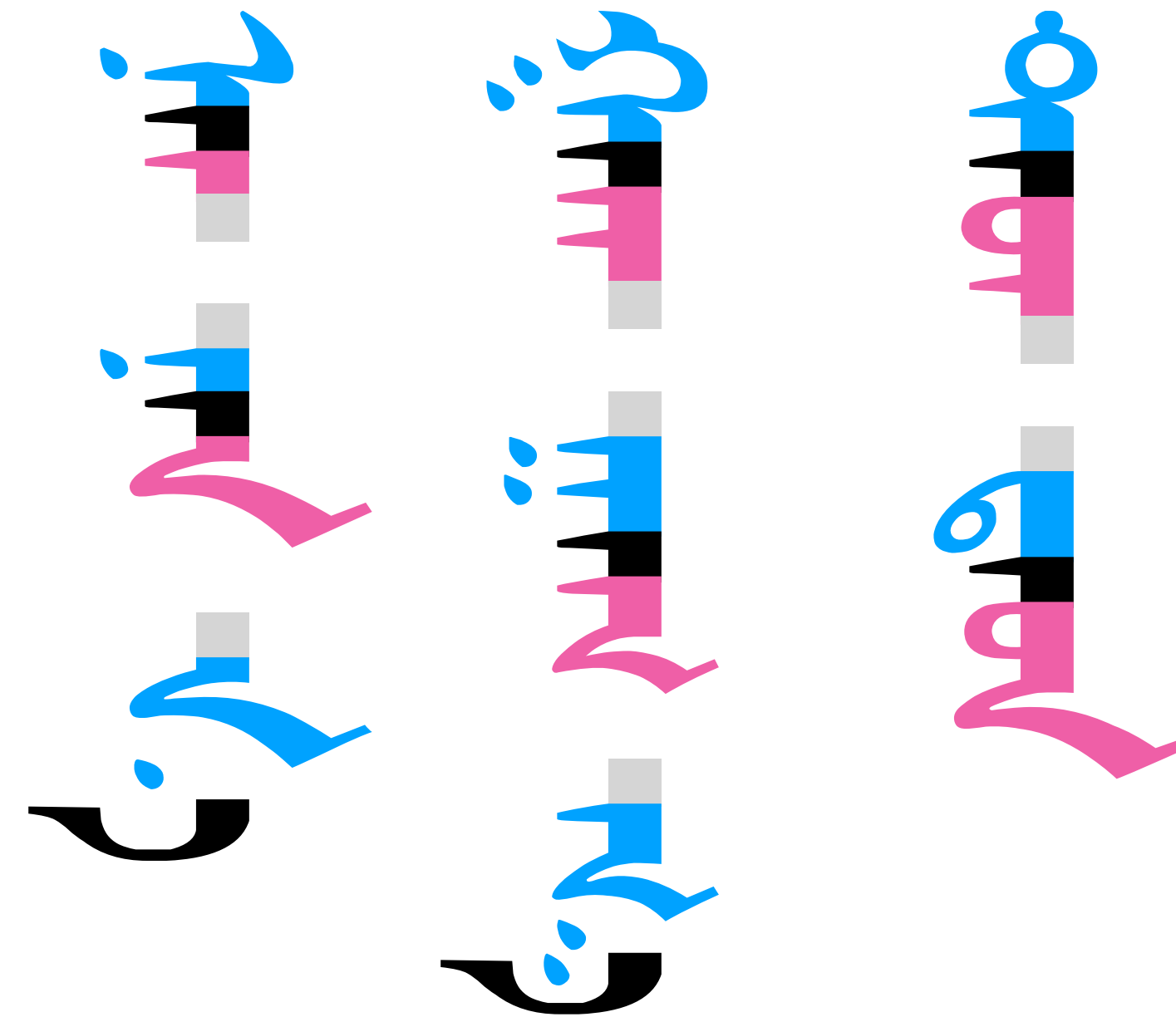
- Syllable onset placeholder (*aleph*)
- **Syllable coda forms** (*n, g, d...*)
- *Vowel harmony class*–specific consonants
- Phonetic letters/syllables



I. Crash course: *General features* [cont.]

Invented during Old Uyghur ~ Mongolian:

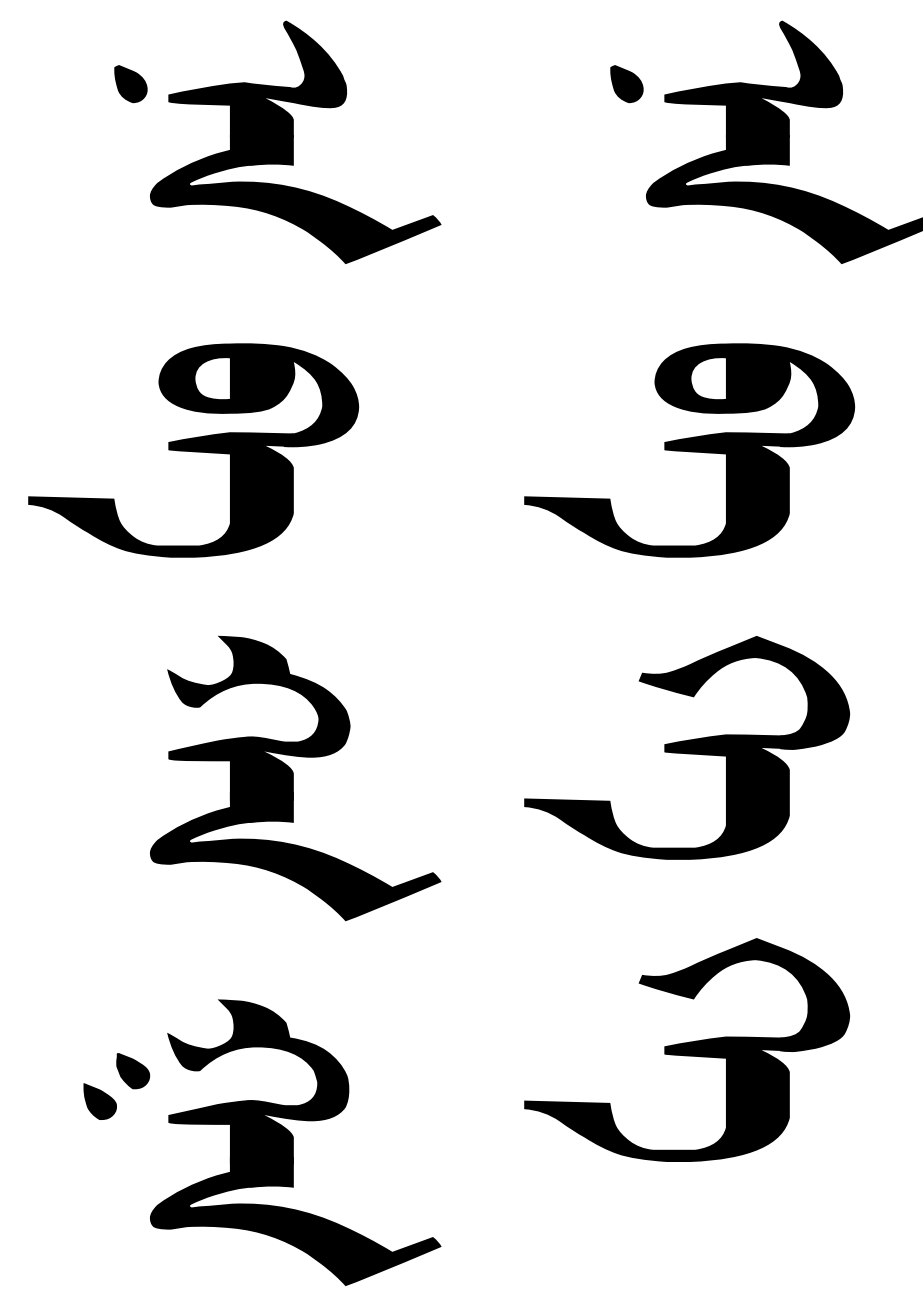
- Syllable onset placeholder (*aleph*)
- **Syllable coda forms** (*n, g, d...*)
- *Vowel harmony class*–specific consonants
- Phonetic letters/syllables



I. Crash course: *General features* [cont.]

Invented during Old Uyghur ~ Mongolian:

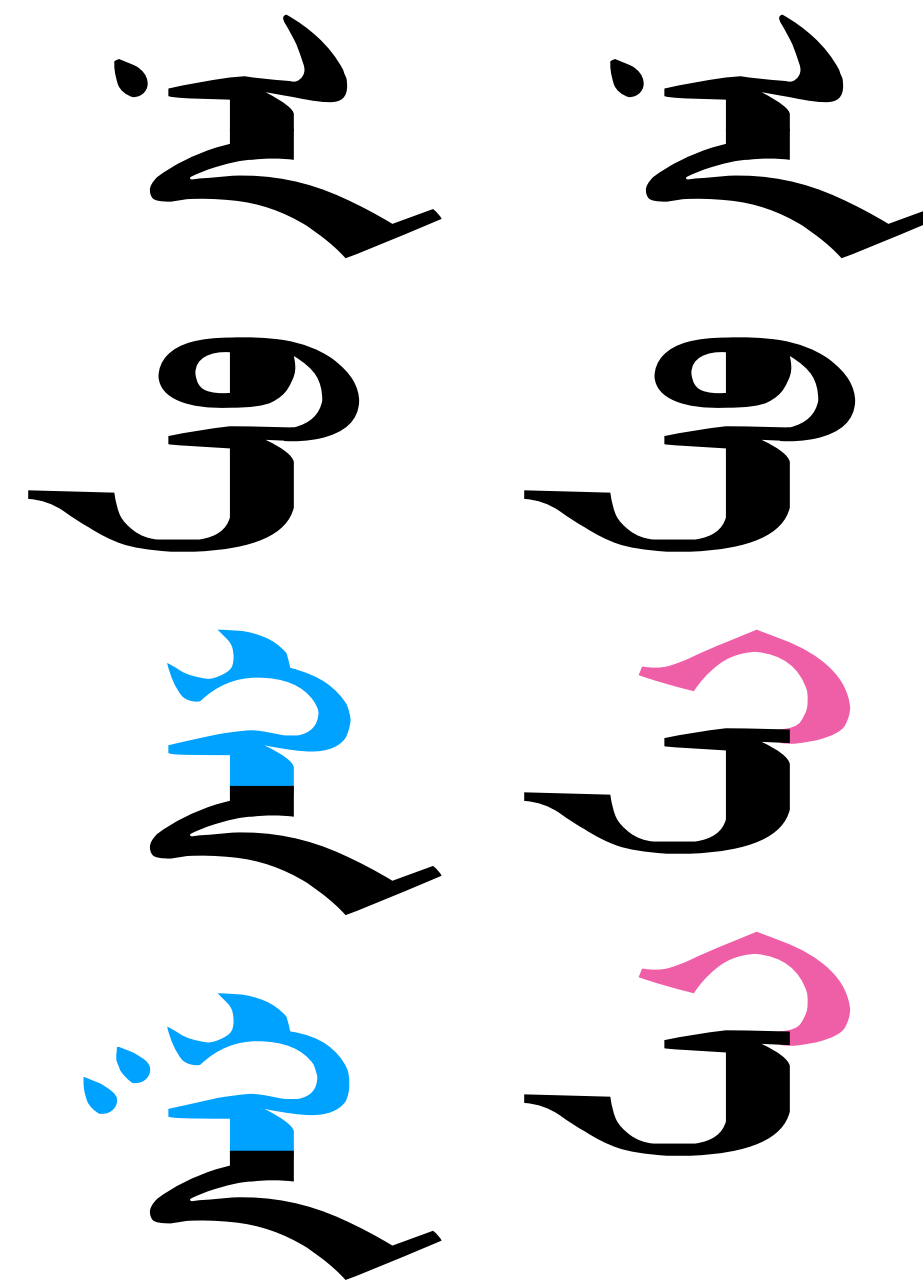
- Syllable onset placeholder (*aleph*)
- Syllable coda forms (*n, g, d...*)
- **Vowel harmony class-specific consonants**
 - Complementary distribution of two guttural series: *gimel* and *kaph*.
- Phonetic letters/syllables



I. Crash course: *General features* [cont.]

Invented during Old Uyghur ~ Mongolian:

- Syllable onset placeholder (*aleph*)
- Syllable coda forms (*n, g, d...*)
- **Vowel harmony class-specific consonants**
 - Complementary distribution of two guttural series: *gimel* and *kaph*.
- Phonetic letters/syllables



I. Crash course: *General features* [cont.]

Invented during Old Uyghur ~ Mongolian:

- Syllable onset placeholder (*aleph*)
- Syllable coda forms (*n, g, d...*)
- *Vowel harmony class*–specific consonants
- **Phonetic letters/syllables**
 - Reanalyzed letters on the basis of *phonemes* instead of *graphemes*.

	Graphs ...	Letters ...	Phones
<i>Spanish</i>		G L ...	P
<i>English</i>		G L P
<i>Arabic</i>		G ... L P
<i>Tibetan</i>	G L P
<i>Mongolian</i>	G L*	... P

رَسْمِيٌّ

رَسْمِيَّةٌ

رَسْمِيَّةَانِ

رَسْمِيَّةَانِ

رَسْمِيَّةَانِ

رَسْمِيَّةَانِ

رَسْمِيَّةَانِ

↓ V V_ nV nV_nV_nV
 → a e i o u ö ü

اَلصَّبْرُ صَبْرٌ

اَلصَّبْرُ صَبْرٌ

اَلصَّبْرُ صَبْرٌ

اَلصَّبْرُ صَبْرٌ

اَلصَّبْرُ صَبْرٌ

اَلصَّبْرُ صَبْرٌ

اَلصَّبْرُ صَبْرٌ

↓ V V_ nV nV_nV_nV
→ a e i o u ö ü

I. Crash course: *Writing system analyses*

- **Hudum** and **Todo** are analyzable as either *phonetic syllables* or *phonetic letters*.
 - Hudum is largely *unpredictable* (one-to-multi, involving grammatical/lexical information) and highly *confusable* (multi-to-one).
 - Todo is highly predictable and minimally confusable.
- **Manchu–Sibe** is analyzable as *phonetic syllables*.
 - Also highly predictable and minimally confusable.
 - Can be *very weird* if must be analyzed as *phonetic letters*.

ᠣᠷᠳᠣ	ᠤᠷᠲᠤ	ᠤᠷᠳᠤ	ᠠᠲᠠ	ᠠᠳᠠ	ᠠᠳᠠ
ᠣᠷᠳᠤᠨ	ᠣᠷᠳᠤᠨ	ᠣᠷᠳᠤᠨ	ᠠᠲᠠ	ᠠᠳᠠᠨ	ᠠᠳᠠᠨ
ᠣᠷᠳᠤᠨ	ᠣᠷᠳᠤᠨ	ᠣᠷᠳᠤᠨ	ᠠᠲᠠ	ᠠᠳᠠ	ᠠᠳᠠ

[↓] Hudum, Manchu–Sibe, and Todo:

ordo urtu urdu | ata ada ende

ᠣᠷᠳᠤ ᠤᠷᠲᠤ ᠤᠷᠳᠤ ᠠᠲᠤ ᠠᠳᠤ ᠡᠨᠳᠡ

ᠣᠷᠳᠤ ᠤᠷᠲᠤ ᠤᠷᠳᠤ ᠠᠲᠤ ᠠᠳᠤ ᠡᠨᠳᠡ

ᠣᠷᠳᠤ ᠤᠷᠲᠤ ᠤᠷᠳᠤ ᠠᠲᠤ ᠠᠳᠤ ᠡᠨᠳᠡ

[↓] Hudum, Manchu–Sibe, and Todo:

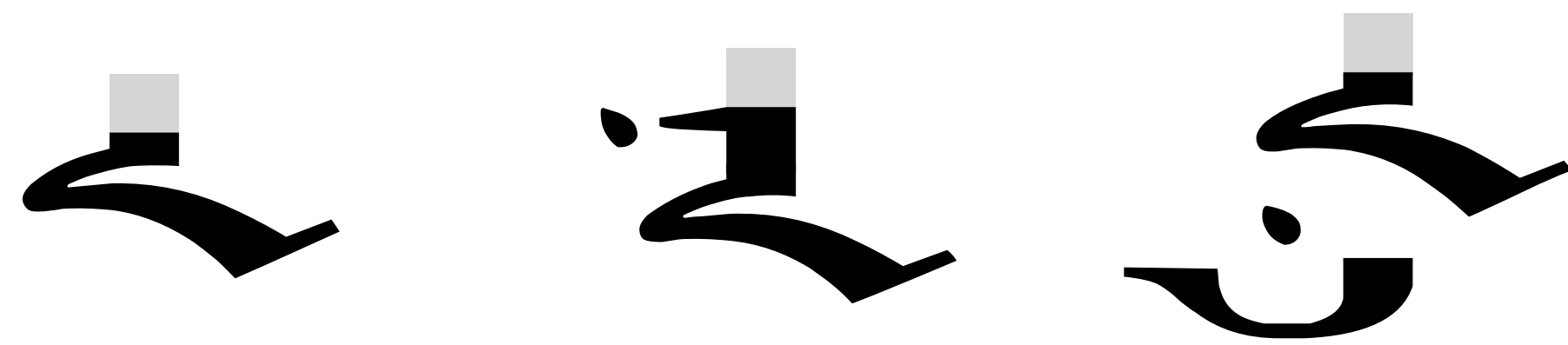
or**do** ur**tu** ur**du** | at**a** ad**a** en**d**e

I. Crash course: *Hudum-specific features*

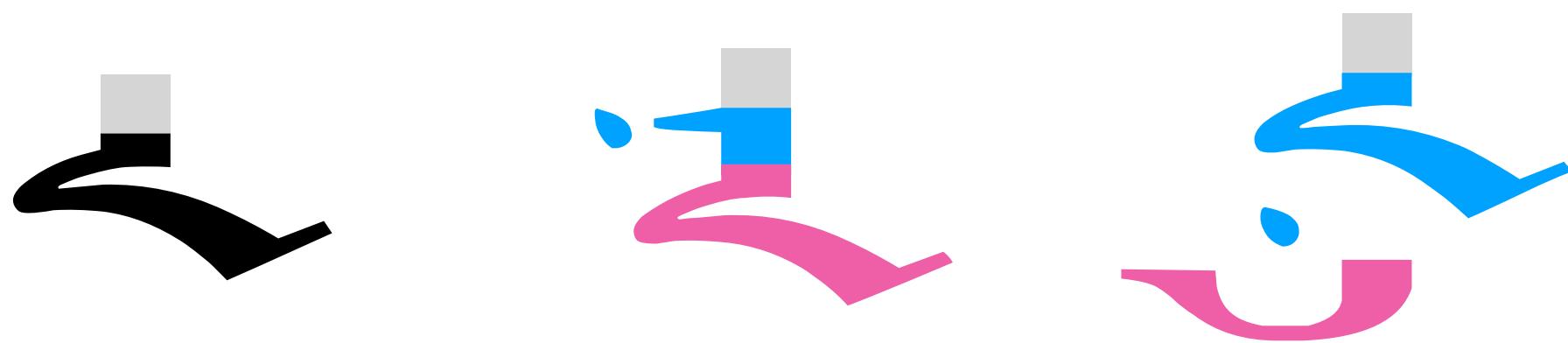
- **Disjointed tail** (*ćaculga*, detached *a/e*)
- **First-vowel forms** (*o, u, ö, ü*)
- **Controversial diphthongs**
- **Complex scopes**
 - One scope per word-stem (note compound words)
 - Word-stem boundaries affect syllable boundaries
 - Suffixes (including enclitics, which is disconnected) extend scopes
- **Purely lexical variants**

I. Crash course: *Hudum-specific features*

- **Disjointed tail** (*ćaculga*, detached *a/e*)
- **First-vowel forms** (*o, u, ö, ü*)
- **Controversial diphthongs**
- **Complex scopes**
 - One scope per word-stem (note compound words)
 - Word-stem boundaries affect syllable boundaries
 - Suffixes (including enclitics, which is disconnected) extend scopes
- **Purely lexical variants**



_n *_na* *_n[a]*



_n *_na* *_n[a]*

I. Crash course: *Hudum-specific features* [cont.]

- Disjointed tail (*ćaculga*, detached *a/e*)
- **First-vowel forms** (*o, u, ö, ü*)
- **Controversial diphthongs**
- **Complex scopes**
 - One scope per word-stem (note compound words)
 - Word-stem boundaries affect syllable boundaries
 - Suffixes (including enclitics, which is disconnected) extend scopes
- **Purely lexical variants**

يٰۤاَيُّهَا الَّذِيْنَ اٰمَنُوْا

يٰۤاَيُّهَا الَّذِيْنَ اٰمَنُوْا

يٰۤاَيُّهَا الَّذِيْنَ اٰمَنُوْا

يٰۤاَيُّهَا الَّذِيْنَ اٰمَنُوْا

يٰۤاَيُّهَا الَّذِيْنَ اٰمَنُوْا

يٰۤاَيُّهَا الَّذِيْنَ اٰمَنُوْا

يٰۤاَيُّهَا الَّذِيْنَ اٰمَنُوْا

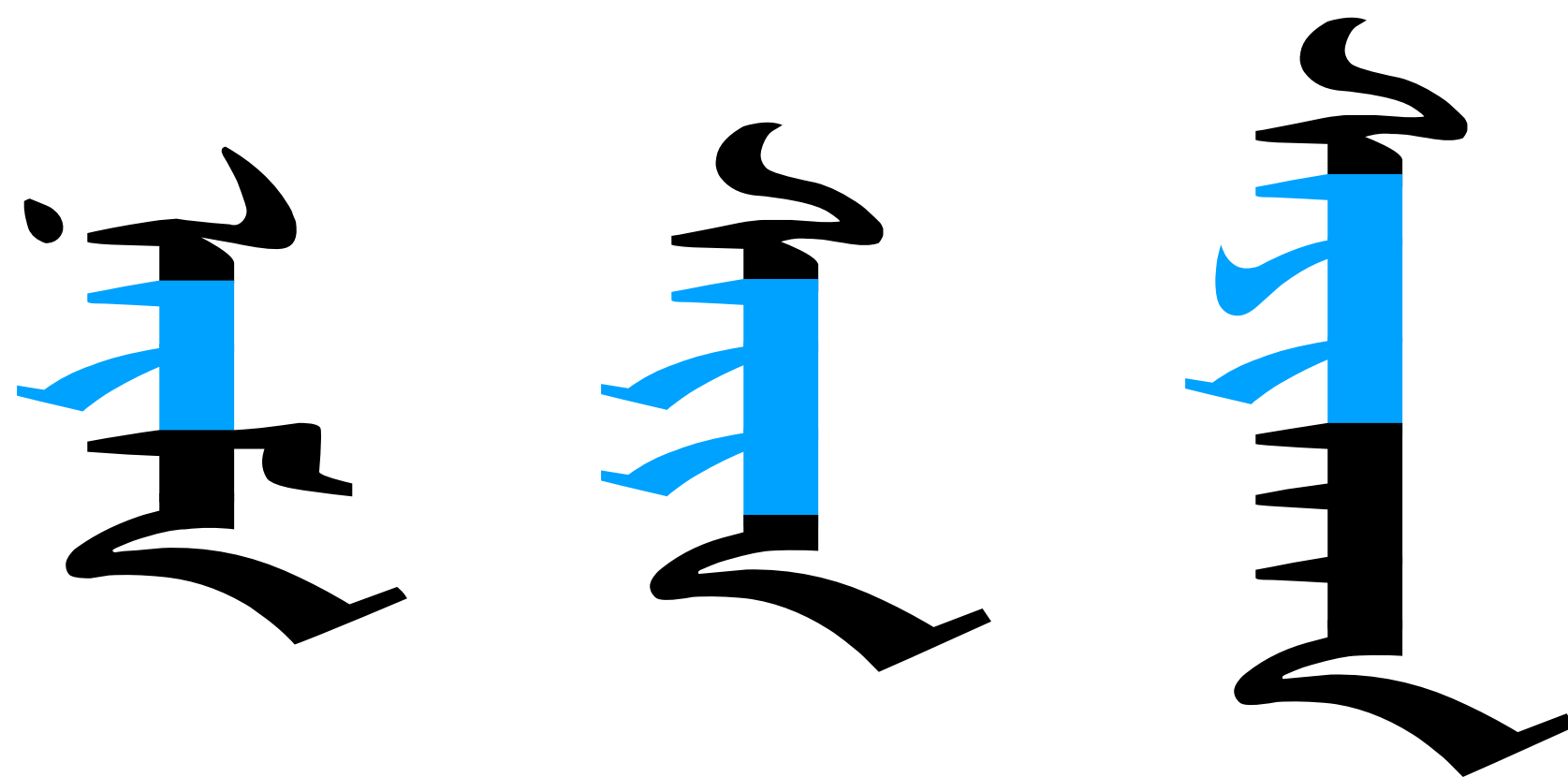
يٰۤاَيُّهَا الَّذِيْنَ اٰمَنُوْا

يٰۤاَيُّهَا الَّذِيْنَ اٰمَنُوْا

i u ii

I. Crash course: *Hudum-specific features* [cont.]

- Disjointed tail (*ćaculga*, detached *a/e*)
- First-vowel forms (*o, u, ö, ü*)
- **Controversial diphthongs**
- **Complex scopes**
 - One scope per word-stem (note compound words)
 - Word-stem boundaries affect syllable boundaries
 - Suffixes (including enclitics, which is disconnected) extend scopes
- **Purely lexical variants**



naima *sain/sayin/sayn* *sayiban*

I. Crash course: *Hudum-specific features* [cont.]

- **Disjointed tail** (*ćaculga*, detached *a/e*)
- **First-vowel forms** (*o, u, ö, ü*)
- **Controversial diphthongs**
- **Complex scopes**
 - One scope per word-stem (note compound words)
 - Word-stem boundaries affect syllable boundaries
 - Suffixes (including enclitics, which is disconnected) extend scopes
- **Purely lexical variants**

I. Crash course: *Todo-specific features*

- **Long vowels, diphthongs, and consecutive vowels**
 - Long-vowel sign
 - Diphthongs written with, for non-final VY, a linking glide (γ) or aleph; for VW, an offglide-specific form of -W; or as-is.
- **A single enclitic (*ni*)**

I. Crash course: *Manchu–Sibe–specific features*

- **Diphthongs and consecutive vowels**
 - For VY, with offglide-specific forms of -Y; for AW, -W (like o).
 - Or written with a linking aleph
- **A single enclitic (*i*)**
- **Irregular syllables**
- **Complex behavior of circle–dot modifiers**

• Part II •

The Unicode Mongolian, an encoding model from hell

Origin and encoding principles.

II. Model: *Origin*

Handwriting

Woodblock

Movable type

Various legacy encodings

The Unicode encoding for Mongolian

..... 1999 [Unicode 3.0; ISO/IEC 10646-1: 1993 / Amd. 29: 1999 (E)]

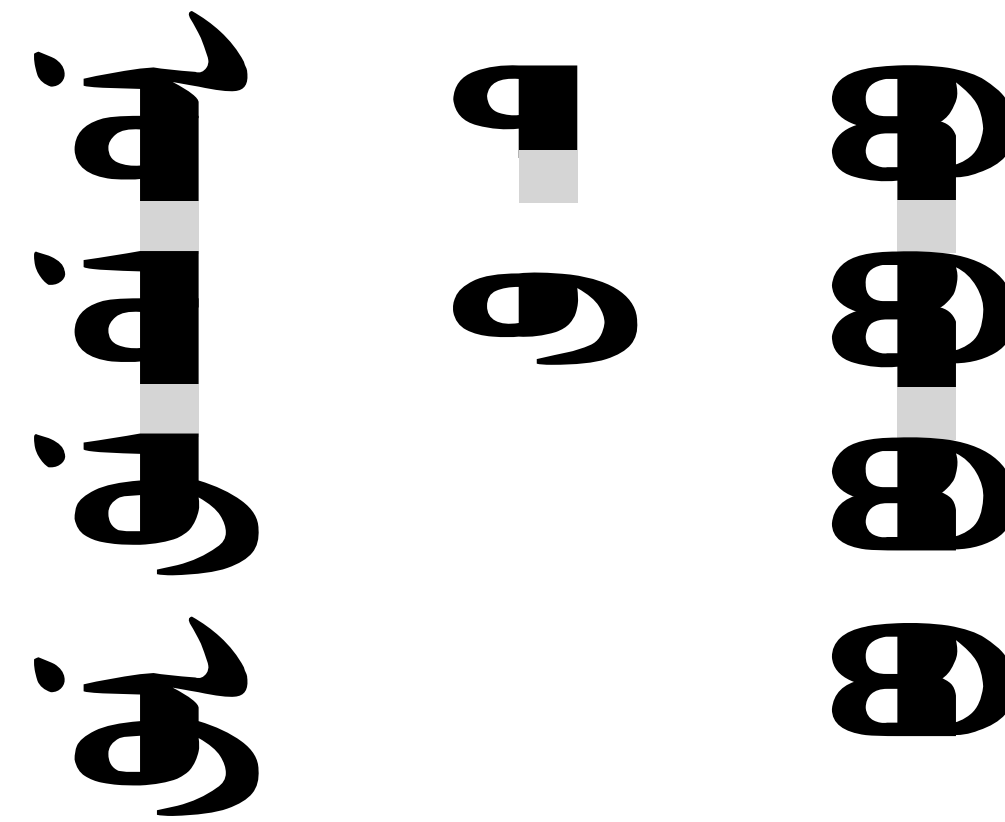
II. Model: *Origin*

Handwriting

Woodblock

Movable type

Various legacy encodings



The Unicode encoding for Mongolian

..... 1999 [Unicode 3.0; ISO/IEC 10646-1: 1993 / Amd. 29: 1999 (E)]

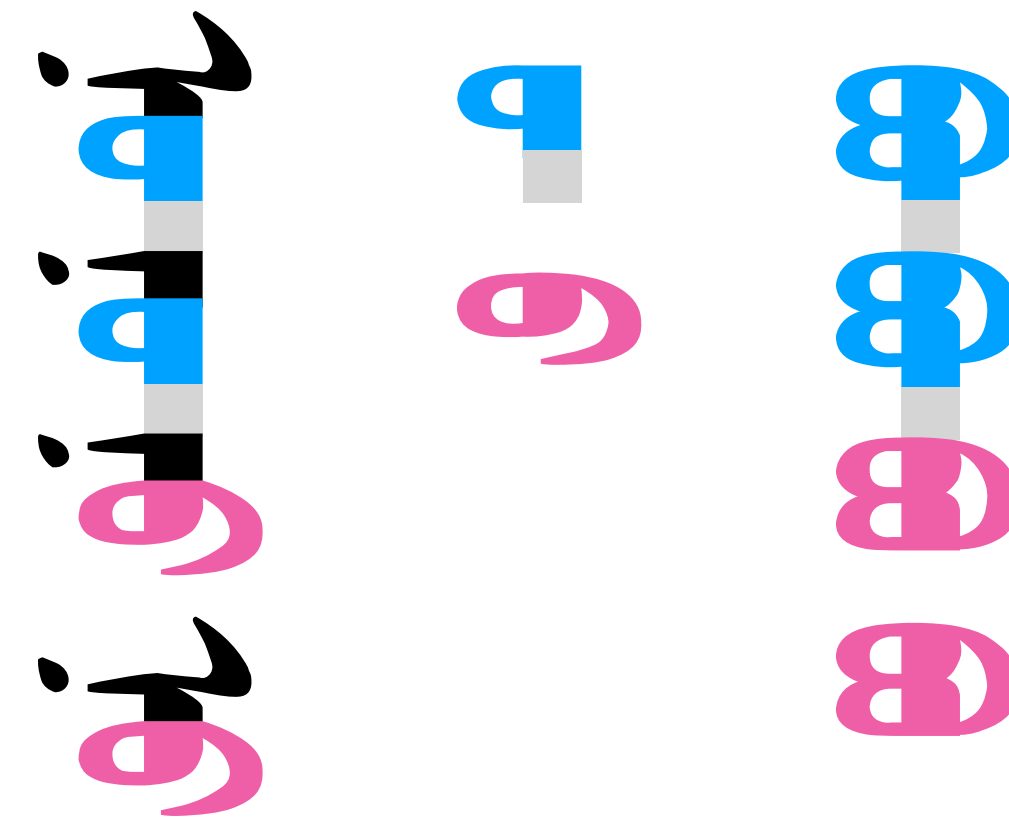
II. Model: *Origin*

Handwriting

Woodblock

Movable type

Various legacy encodings



The Unicode encoding for Mongolian

..... 1999 [Unicode 3.0; ISO/IEC 10646-1: 1993 / Amd. 29: 1999 (E)]

II. Model: *Early proposals*

Graphically duplicated *phonetic glyphs* encoded as characters, including *fragments of bowed-consonant ligatures*:

- **GH/90** *Code system for the Mongolian script* ↗Mongolia, 1993
 - Hudum-only
- **WG2 N1011** *A proposal about installing the Mongolian, Todo, Xibe (Manchu included) scripts into ISO/IEC 10646 BMP* ↗China, 1994
 - *Unification across writing systems*: corresponding context-specific glyphs of related phonetic letters

II. Model: *Early proposals* [cont.]

Graphically duplicated *phonetic glyphs* encoded as characters:

- **WG2 N1368** *Joint proposal draft on encoding Mongolian character set* ↗
.....China and Mongolia, the first joint proposal, 1996
 - *Unification across writing systems*: identical glyphs
 - *Bowed-consonant ligatures*: dynamically formed from characters

II. Model: *Early proposals* [cont.]

Graphically confusable *phonetic letters* encoded as cursive characters:

- **WG2 N1711** *The working meeting on Mongolian encoding attended by representatives of China and Mongolia* ↗China and Mongolia, joint proposal, 1998
 - *Unification across writing systems*: phonetic letters that appear identical in any contexts
 - *Bowed-consonant ligatures*: dynamically formed from characters
 - **Prototype of the Unicode encoding model**

II. Model: *Encoding principles*

- P1 Underlying **phonetic letters** are encoded as characters.
- P2 Characters are **cursive** with **word-wise** positional forms.
- P3 **Bowed consonants** are ligated to the immediately following vowels.
- P4 When **multiple forms** are possible on a position, additional mechanisms apply.
 - P4a **Contextual rules** select generally expected forms.
 - P4b **MVS** triggers special spellings for the lexical feature of detached *a/e*.
 - P4c **NNBSP** triggers special spellings for the grammatical feature of enclitics.
 - P4d **FVSes** request forms that are not selected by the mechanisms above.
- P5 [*de facto*] **In-isolation** and **in-word** forms are decided with different processes.

II. Model: *Encoding principles*

- P1 Underlying **phonetic letters** are encoded as characters.
- P2 Characters are **cursive** with **word-wise** positional forms.
- P3 **Bowed consonants** are ligated to the immediately following vowels.
- P4 When **multiple forms** are possible on a position, additional mechanisms apply.
 - P4a **Contextual rules** select generally expected forms.
 - P4b **MVS** triggers special spellings for the lexical feature of detached *a/e*.
 - P4c **NNBSP** triggers special spellings for the grammatical feature of enclitics.
 - P4d **FVSes** request forms that are not selected by the mechanisms above.
- P5 [*de facto*] **In-isolation** and **in-word** forms are decided with different processes.

رَسْمِيٌّ

رَسْمِيَّةٌ

رَسْمِيَّةَانِ

رَسْمِيَّةَانِ

رَسْمِيَّةَانِ

رَسْمِيَّةَانِ

رَسْمِيَّةَانِ

↓

V

V_

nV

nV_nV_nV

→

a

e

i

o

u

ö

ü

II. Model: *Encoding principles* [cont.]

- P1 Underlying **phonetic letters** are encoded as characters.
- P2 Characters are **cursive** with **word-wise** positional forms.
- P3 **Bowed consonants** are ligated to the immediately following vowels.
- P4 When **multiple forms** are possible on a position, additional mechanisms apply.
 - P4a **Contextual rules** select generally expected forms.
 - P4b **MVS** triggers special spellings for the lexical feature of detached *a/e*.
 - P4c **NNBSP** triggers special spellings for the grammatical feature of enclitics.
 - P4d **FVSes** request forms that are not selected by the mechanisms above.
- P5 [*de facto*] **In-isolation** and **in-word** forms are decided with different processes.

زفلفق زق

زفلفق

زف زق

زف* زق*

زف زق

و و*

و و*

ଆ ର ଓ

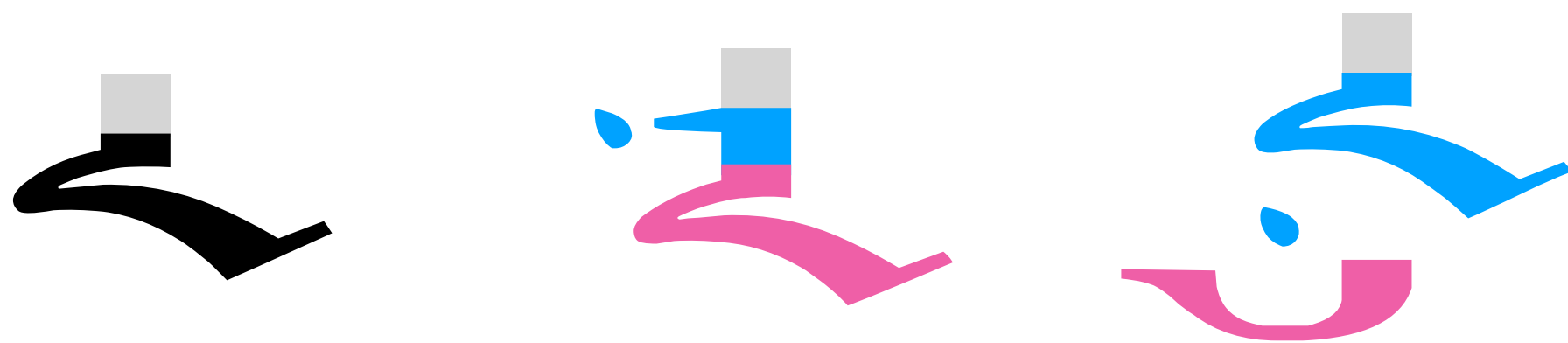


... ଆ ର ଓ

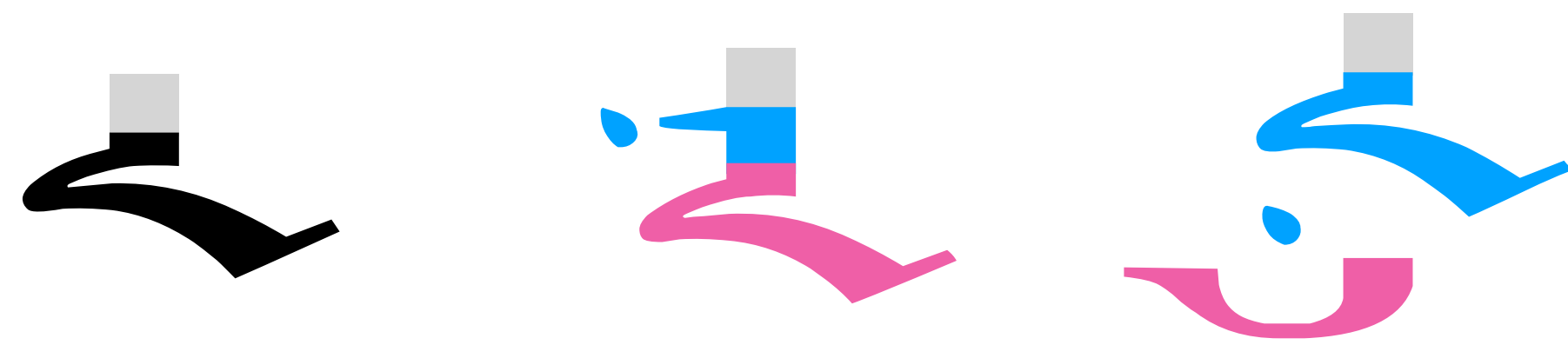


II. Model: *Encoding principles* [cont.]

- P1 Underlying **phonetic letters** are encoded as characters.
- P2 Characters are **cursive** with **word-wise** positional forms.
- P3 **Bowed consonants** are ligated to the immediately following vowels.
- P4 When **multiple forms** are possible on a position, additional mechanisms apply.
 - P4a **Contextual rules** select generally expected forms.
 - P4b **MVS** triggers special spellings for the lexical feature of detached *a/e*.
 - P4c **NNBSP** triggers special spellings for the grammatical feature of enclitics.
 - P4d **FVSes** request forms that are not selected by the mechanisms above.
- P5 [*de facto*] **In-isolation** and **in-word** forms are decided with different processes.



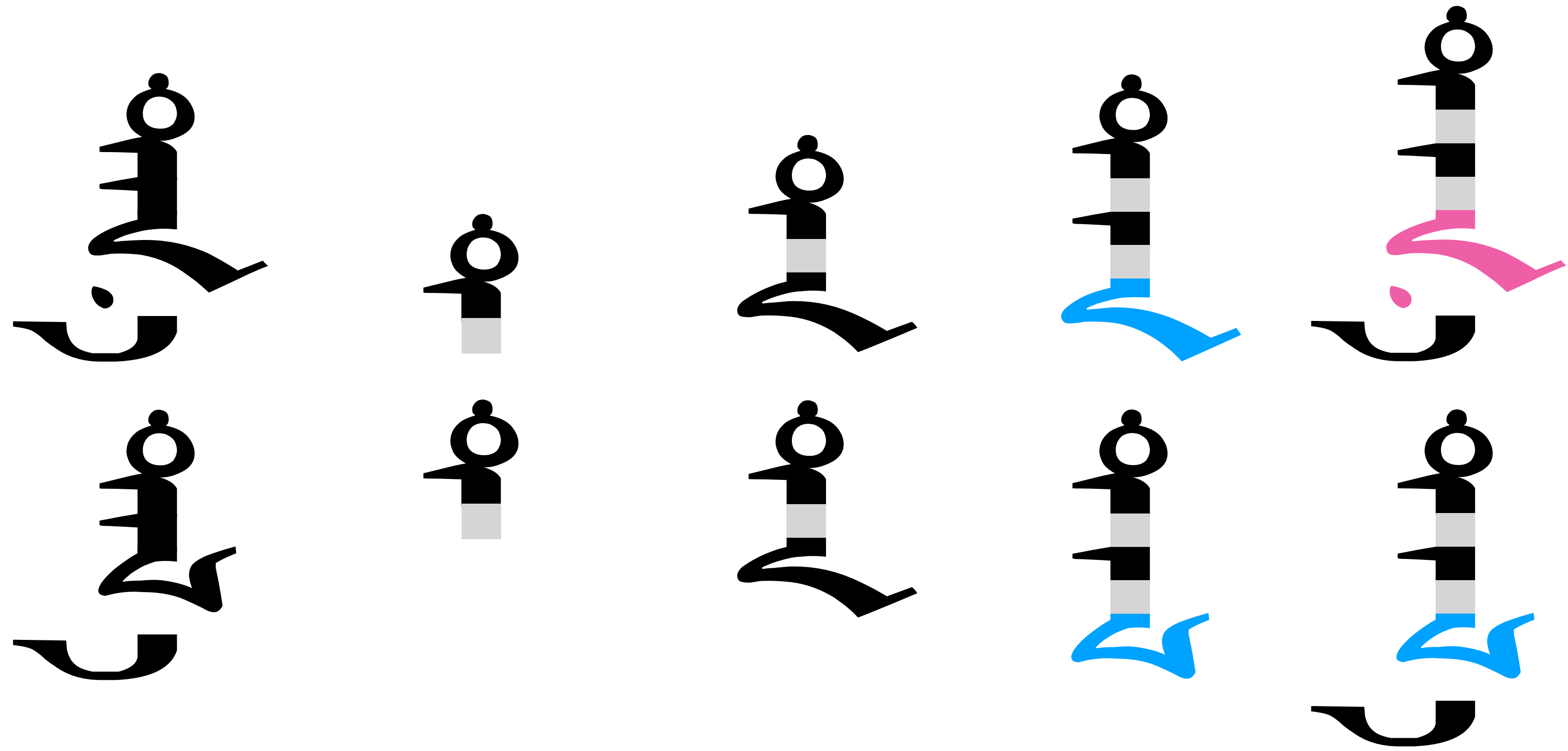
_n *_na* *_n[a]*



_n *_na* *_n*<MVS>*a*

II. Model: *Encoding principles* [cont.]

- P1 Underlying **phonetic letters** are encoded as characters.
- P2 Characters are **cursive with word-wise positional forms**.
- P3 **Bowed consonants** are ligated to the immediately following vowels.
- P4 When **multiple forms** are possible on a position, additional mechanisms apply.
 - P4a **Contextual rules** select generally expected forms.
 - P4b **MVS** triggers special spellings for the lexical feature of detached *a/e*.
 - P4c **NNBSP** triggers special spellings for the grammatical feature of enclitics.
 - P4d **FVSes** request forms that are not selected by the mechanisms above.
- P5 [*de facto*] **In-isolation** and **in-word** forms are decided with different processes.

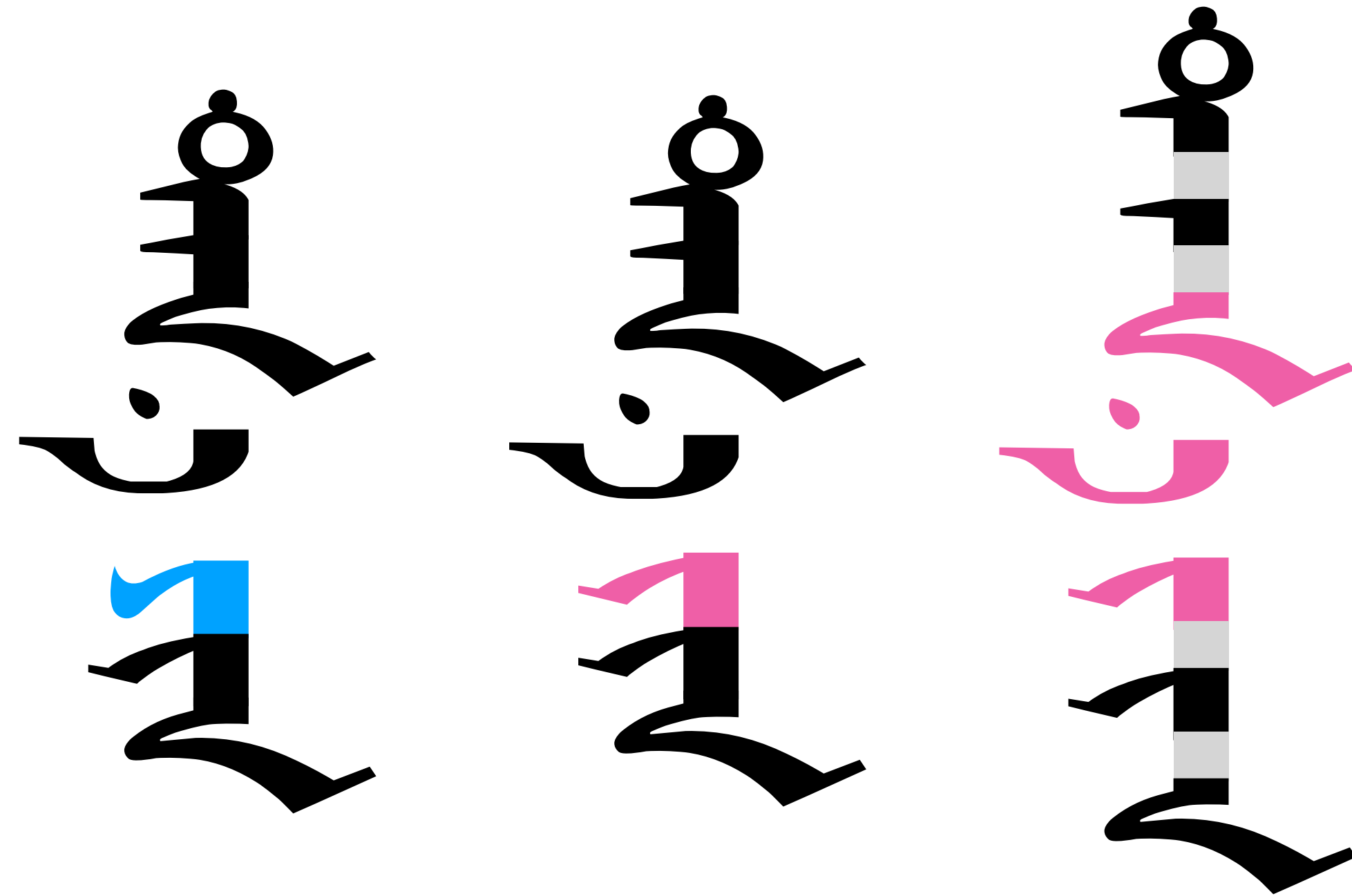


tan<MVS>*a* *t_* *t_a* *t_a_n* *t_a_n*<MVS>*a*

tal<MVS>*a* *t_* *t_a* *t_a_l* *t_a_l*<MVS>*a*

II. Model: *Encoding principles* [cont.]

- P1 Underlying **phonetic letters** are encoded as characters.
- P2 Characters are **cursive** with **word-wise** positional forms.
- P3 **Bowed consonants** are ligated to the immediately following vowels.
- P4 When **multiple forms** are possible on a position, additional mechanisms apply.
 - P4a **Contextual rules** select generally expected forms.
 - P4b **MVS** triggers special spellings for the lexical feature of detached *a/e*.
 - P4c **NNBSP** triggers special spellings for the grammatical feature of enclitics.
 - P4d **FVSes** request forms that are not selected by the mechanisms above.
- P5 [*de facto*] **In-isolation** and **in-word** forms are decided with different processes.



tan<MVS>*a*<SP>*yin*

tan<MVS>*a*<NNBSP>*yin*

t_a_n<MVS>*a*<NNBSP>*y_i_n*

ᠮᠣᠩᠭᠣᠯᠤᠯᠤᠰᠤᠨᠠᠨ

ᠮᠣᠩᠭᠣᠯᠤᠯᠤᠰᠤᠨᠠᠨ

ᠮᠣᠩᠭᠣᠯᠤᠯᠤᠰᠤᠨᠠᠨ

ᠮᠣᠩᠭᠣᠯᠤᠯᠤᠰᠤᠨᠠᠨ

mongol<SP>*un*

...<NNBSP>*un*

...<SP>*tai*

...<NNBSP>*tai*

II. Model: *Encoding principles* [cont.]

- P1 Underlying **phonetic letters** are encoded as characters.
- P2 Characters are **cursive** with **word-wise** positional forms.
- P3 **Bowed consonants** are ligated to the immediately following vowels.
- P4 **When multiple forms are possible on a position, additional mechanisms apply.**
 - P4a **Contextual rules** select generally expected forms.
 - P4b **MVS** triggers special spellings for the lexical feature of detached *a/e*.
 - P4c **NNBSP** triggers special spellings for the grammatical feature of enclitics.
 - P4d **FVSes** request forms that are not selected by the mechanisms above.
- P5 [*de facto*] **In-isolation** and **in-word** forms are decided with different processes.

ᠭᠠᠳᠠ

ᠠᠭᠠ

ᠠᠭᠳᠠ

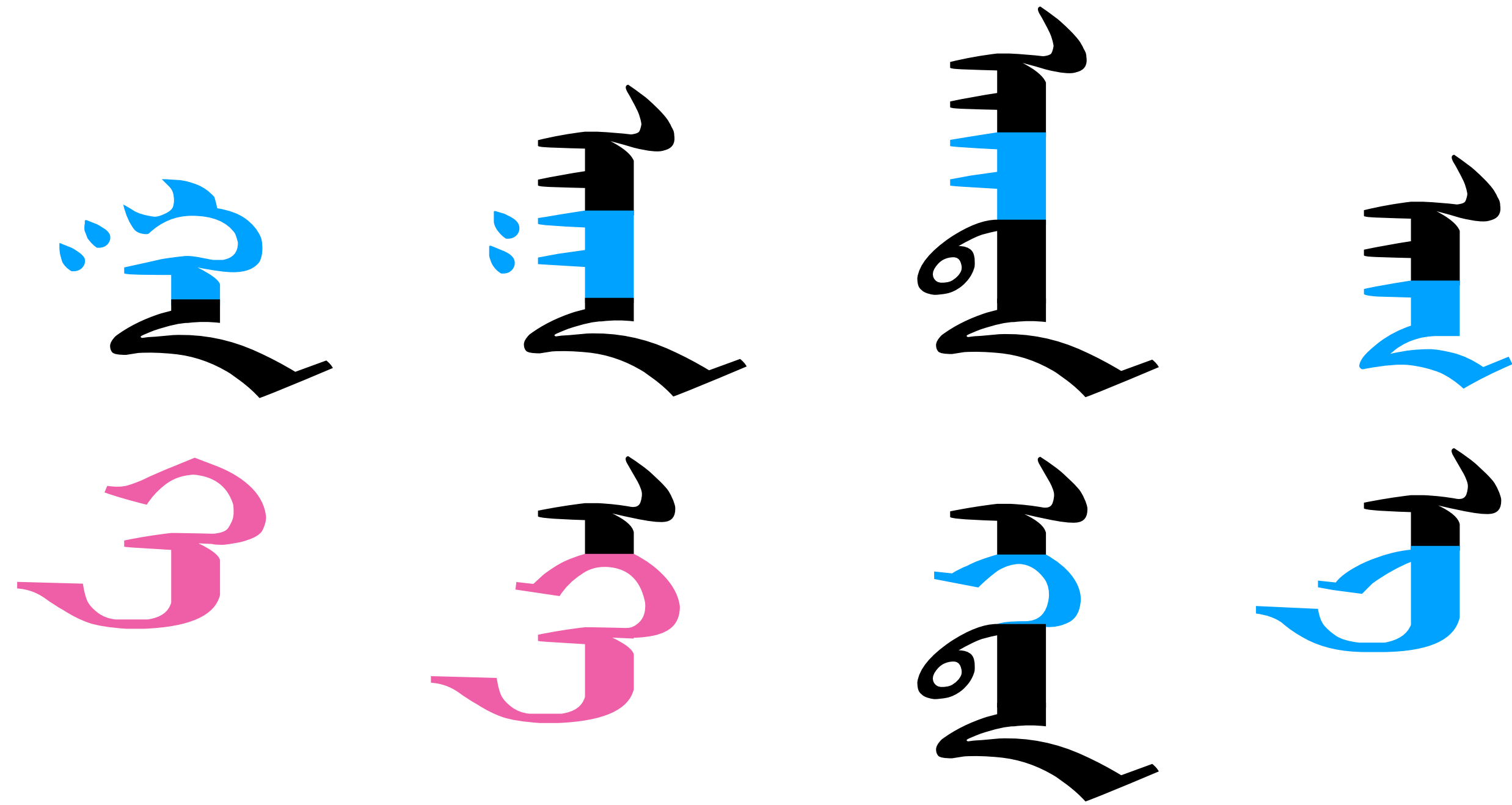
ᠠᠭᠳᠠᠳᠠ

ᠠᠭᠳᠠᠳᠠᠳᠠ

ᠠᠭ

ᠠᠭ

ga a.ga ag.da ag
ge e.ge eg.de eg



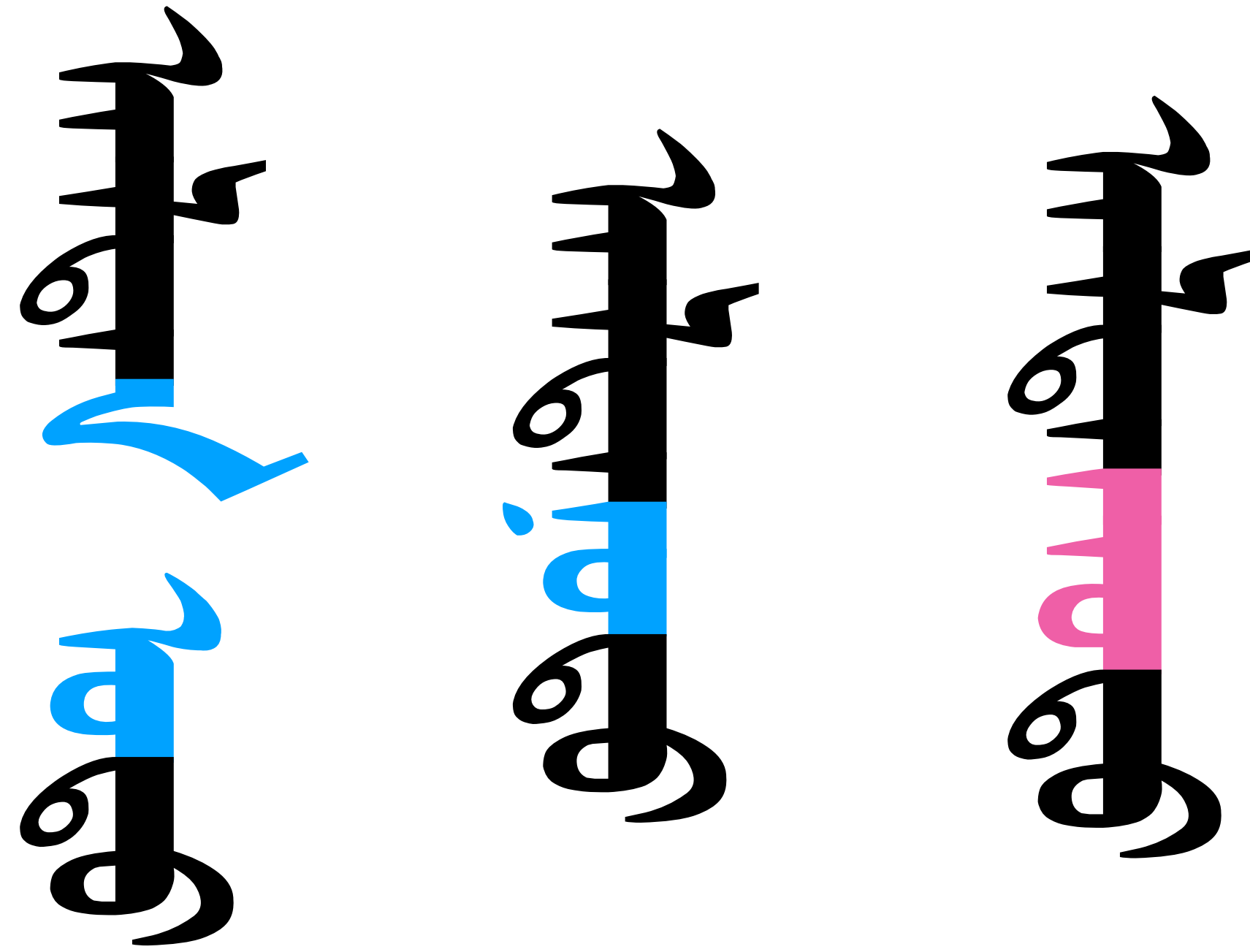
ga *a.ga* *ag.da* *ag*
ge *e.ge* *eg.de* *eg*

ed

ed

ed

ed<FVS>



al.tan<SP>*o.do*

al.ta.no.do

al.tan<FVS>.*o*<FVS>.*do*

ايتو

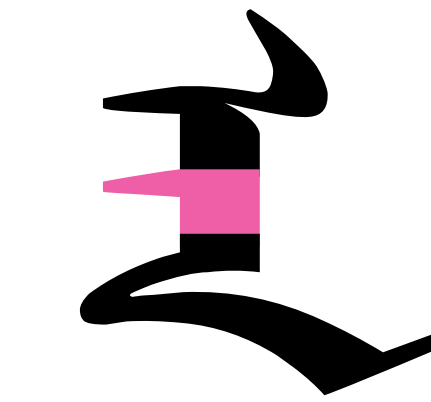
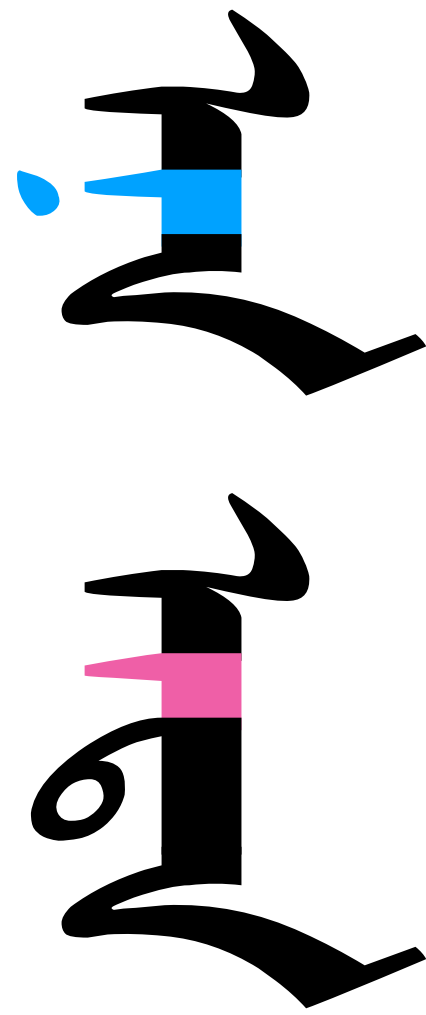
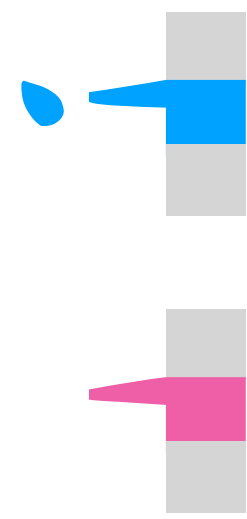
ايتو

aïto

aï<FVS>*t*<FVS>*o*<FVS>

II. Model: *Encoding principles* [cont.]

- P1 Underlying **phonetic letters** are encoded as characters.
- P2 Characters are **cursive** with **word-wise** positional forms.
- P3 **Bowed consonants** are ligated to the immediately following vowels.
- P4 When **multiple forms** are possible on a position, additional mechanisms apply.
 - P4a **Contextual rules** select generally expected forms.
 - P4b **MVS** triggers special spellings for the lexical feature of detached *a/e*.
 - P4c **NNBSP** triggers special spellings for the grammatical feature of enclitics.
 - P4d **FVSes** request forms that are not selected by the mechanisms above.
- P5 [*de facto*] **In-isolation** and **in-word** forms are decided with different processes.



n<FVS1>_

e.ne

en<FVS1>.*de*

n

en.de

e.n<FVS1>*e*

II. Model: *Various (de facto) standards*

The Users' Convention:

- **"TR #170"** ↗ (table A ↗ · table B ↗)Myatav Erdenechimeg, et al.

UNU/IIST (The United Nations University / International Institute for Software Technology) Report No. 170: *Traditional Mongolian Script in the ISO/IEC 10646 and Unicode Standards*. August 1999.

- **MNS 4932: 2000** ↗Mongolia

Монголжин бичгийн кодыг хэрэглэх дүрэм / Use of Mongolian character encoding.
2000.

II. Model: *Various (de facto) standards [cont.]*

The Users' Convention, altered:

- **"MGWBM"** ↗Quejingzhabu*

(*  *choijongjab / coyijongjab / coyjongjab*; 确精扎布 *què jīng zhā bù*; Choijinzhab)

“蒙古文编码” (*měng gǔ wén biān mǎ*), literally “Mongolian script encoding”. August 2012.

- **GB/T 26226–2010**China

“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 controlling characters*”. 10 January 2011.

Subsets: **GB/T 25914–2010** (Hudum) · **GB/T 36331–2018** (Uyghur Mongolian)

II. Model: *Various (de facto) standards [cont.]*

Specifications with contextual rules:

- **"Specification 9"** ↗Quejingzhabu

“传统蒙古文名义字符到变形显现字符的转换规则（供微软用）”，literally “traditional Mongolian script nominal-characters-to-presentation-characters conversion rules (for Microsoft’s use)”.

The ninth draft. 25 December 2012.

- **"EAC Project Standard"** ↗EAC of Inner Mongolia,* China

(* *ᠠᠨᠢᠮᠤᠩᠭᠣᠯᠠᠯᠤᠯᠤᠰ ᠤᠨ ᠨᠠᠭᠤᠨ ᠨᠠᠭᠤᠨ ᠨᠠᠭᠤᠨ ᠨᠠᠭᠤᠨ ᠨᠠᠭᠤᠨ ᠨᠠᠭᠤᠨ ᠨᠠᠭᠤᠨ ᠨᠠᠭᠤᠨ ᠨᠠᠭᠤᠨ ᠨᠠᠭᠤᠨ* / 内蒙古自治区民族事务委员会 / Ethnic Affairs Committee of the Inner Mongolia Autonomous Region)

“信息技术 传统蒙古文名义字符到变形显现字符转换规则 / *Information technology — The Transferring Rules of Traditional Mongolian Nominal Form to Variant Form*”.

Version 1.0.2. Dated 17 June 2018, published on 8 August 2018.

• Part III •

What exactly are not working?

" ~ ! @ # \$ % ^ & * () _ + ? ! "

III. Issues: *Problematic principles*

Refresh:

- P1 **Phonetic letters...**
- P2 **Word-wise cursive...**
- P3 **Bowed consonants...**
- P4 **Multiple forms...**
 - P4a **Contextual rules...**
 - P4b **MVS...**
 - P4c **NNBSP...**
 - P4d **FVSes...**
- P5 **[*de facto*] In-isolation vs in-word...**

III. Issues: *Problematic principles* [cont.]

General methodology:

- **Poor separation of concerns**
- **No coherent abstraction layers**
- **Non-sequential execution of rules**
 - Enumerated rules that only cover common cases. Trying to directly transform characters (“nominal characters”) into final glyphs (“presentation characters”).

III. Issues: *P1. Phonetic letters*

How to **segment** written forms and **identify** underlying phonetic letters is highly controversial.

- Phonetic information is *theoretically good to have*, but the problem of input errors was underestimated.
 - Scholars designed with idealism. Users suffer from reality.
- As the *text representation* principle, P1 is heavily coupled to and affected by *rendering* principles P2–P4.

III. Issues: *P1. Phonetic letters [cont.]*

Usability:

- Users **can't consistently identify** underlying phonetic letters.
- Users **don't care** about orthodoxly correct phonetic letters.
- Users **can't trust** text for reliable phonetic information.
 - Phonetic normalization is practically required for any processes that involve phonetic information.
- Users suffer from **visual confusability and text spoofing**.

III. Issues: *P2. Word-wise cursive*

The word-wise model **conflicts** with the standard cursive joining Model.

- Vendors are driven to *patch* implementations with self-invented rules.
 - Inconsistent implementations
- [*myth*] MVS and NNBSF need dictionary-based complicated effects on cursive positions? Well, it's largely a result of analyzing with word-wise positional forms.

III. Issues: *P3. Bowed consonants*

Hard-coded character interaction *parallel* to all other contextual processes.

- Neither ligature segments or their underlying allographs are identified as variants.
- Causing contextual rules to be unnecessarily complicated and incoherent.
- [*note*] Ligation is just a special case of contextual variation.

III. Issues: *P4a. Contextual rules*

No agreement on a stable set.

- Not built systematically from the ground up with well agreed-on principles (eg, *the twelve syllabaries*).
 - *Arbitrarily* cover common cases, leaving marginal cases undefined.
 - Involve dictionary-based and phonological rules.
- *Syllabification* is crucial for defining the rules and is helpful for other text processes, but is not clearly defined.

III. Issues: *P4b. MVS*

As syntactic sugar, its behavior is undefined when used in unintended environments, eg, when typing.

- <..., C, **FVS**, **ZWNJ**, A/E, **FVS**>
- <..., C, **MVS**, A/E>

III. Issues: P4c. NNBS

Another *syntactic sugar*, relying on a predefined *dictionary* which in turn is result of *controversial grammar theories*.

- Width and line-breaking behavior are defined to suit a certain grammatical understanding's preference, instead of meeting the general public's need.

Usability:

- Fails in script run segmentation and font fallback.

III. Issues: P4d. FVSeS

No agreement on FVS assignment for in-word shaping.

- A certain *Mongolian variation sequence's* positional forms are irrelevant to each other.
 - When typing, a user needs to *predict* an FVS's effect if the base character is not on the desired cursive position yet.
- The *de facto* behavior in many implementations is *context-dependent*, allowing users to mostly stick to FVS1 when requesting an alternative form. However this logic is not coherent when it comes to marginal cases.

Usability:

- Users have difficulty with manual keyboards and largely rely on smart input methods.

III. Issues: *P5*. [*de facto*] *In-isolation vs in-word*

Different sets of contextual rules and FVS assignment apply to the two processes, *in-isolation* and *in-word*.

- The departed *in-word* rules tend to be exploited by specification authors and developers to include incoherent rules, allowing fewer FVSes to be used in common words.

III. Issues: *Poor specification*

The originally planned *Users' Convention*, which was meant to be the shaping specification, was not internationally reviewed and was not freely published.

- The *Users' Convention* doesn't include the crucial contextual rules.
- Experts and vendors are forced to develop private specifications.

Poor coordination between national bodies.

- The standards are unstable and not synchronized.
- Authors change content without consensus from the community.

III. Issues: *Fragile implementations*

Developers *don't have access* to a proper specification.

- Forced to interpret with private, inconsistent understandings.
- Implementations are not interoperable.

Users see *inconsistent, unreliable* rendering between fonts and shaping engines, and don't get support from major OSes and applications.

- Restricted to vendor-specific, non-interoperable ecosystems.

• Part IV •

Tough lessons learned

Very educational. :)

IV. Lessons: *The concept of letters*

The concept of letters can be very misleading.

- Mongolian “letters” shouldn’t be compared to English letters, since they don’t directly correspond to graphemes.

IV. Lessons: *Unicode architecture*

We need to be accurately and repeatedly explain and discuss the Unicode architecture to native experts.

- Misunderstood “*presentation forms shall not be encoded*”. (cf. **WG2 N1368**)
- The relationship between characters and glyphs is widely misunderstood, while itself also evolves.
- The separation of encoding, input, and display layers.

IV. Lessons: *Cursive joining*

The cursive joining model is often misunderstood.

- Experts tend to confuse *word-wise* positions with the plain *cursive* positions.
- Mongolian experts didn't understand that *in-isolation* forms are not special in the cursive joining model.
- [*lost in translation*] The *word-wise* positional forms are added to the standard and named like normal *cursive* ones.
- [*myth*] The Mongolian variants (positional forms of both atomic characters and standardized variation sequences) in the names list are practically only relevant to *in-isolation* forms. Limited value, and misleading.

IV. Lessons: *Prototyping*

Designing a new encoding model without prototyping is be dangerous.

- Complex new models need to have working prototypes from multiple parties for cross-checking encoding principles.
- Experts need to review encoded sample text. Text engine and font prototypes need to be tested.
- Input methods should be prototyped too. (cf. FVS usability during typing.)

IV. Lessons: *Specifications*

Unicode–OpenType experts need to own specifications of complex scripts.

- A well-reviewed and frozen specification at the time of accepting characters is crucial.
- Deferring the specification is harmful to interoperability.
- The specification authors need to provide reference implementations.
 - Mongolian experts are not familiar with Unicode–OpenType technologies and failed to properly implement long-distance effect in OpenType.

IV. Lessons: *Interoperability*

Interoperability is often overshadowed by *seemingly conformant* implementations.

- Implementors tend to settle for a *implementable* model and not realize underlying major issues.
 - “I have implemented Mongolian shaping. It was not very difficult at all.”
 - Implementing once with a certain understanding is different from implementing it multiple times consistently.
- **Don't hesitate to call out** when noticing an encoding model problematic (or even just feeling weird) during implementation.

IV. Lessons: *National bodies*

National bodies tend to only submit a single, final proposal (supposedly an internal agreement) for international discussion.

- Valuable internal opinions are left behind, and opportunities for correcting internal misunderstandings are missed.
- Need to encourage national bodies to seek early, informal feedback from expert groups like Script Ad Hoc.

National standards are often not properly synchronized to international standards despite appearing so, which is misleading and harmful.

IV. Lessons: *Contextual shaping*

Some thoughts about contextual shaping.

- The standard cursive joining model might not be a good option for all cursive scripts
 - It relies on reasonable fallback forms.
 - Mongolian tends to not have positional forms well-defined on all positions (especially lacking distinct isolate forms), despite being dual-joining.
 - For absence of natural fallback, explicit and artificial warnings should be considered (cf. arrows in Abkai fonts that indicate invalid positions).

IV. Lessons: *Contextual shaping* [cont.]

- Text encodings shouldn't enforce a certain school of grammar and orthography.
- Relying on common, misleading characters (eg, NNBS) for required shaping is dangerous.
- Designing format control mechanisms from a static view (when a whole word is present then modify) can lead to confusing user experience when typing.
- When an encoding model already has a logically complete process (eg, FVSes), introducing incomplete (although convenient) syntactic sugar is duplicative, and is a warning that the model might be problematic.
- For complex shaping logic, one-step and parallel contextual rules are hard to design properly and implement accurately.

• Part V •

Ongoing efforts, and how to participate

Discussions, resources, and some (limited) progress.

V. Efforts: *Expert groups and meetings*

Unicode Consortium:

- **Script Ad Hoc**, more or less monthly, with occasionally topical meetings
 - Mongolian ad hoc, WG2 #65September 2016, **L2/16-297**
 - *Recommendations on Mongolian text model*.....August 2017, **L2/17-328**
 - Mongolian ad hoc (redesignated as MWG1), WG2 #66.....September 2017, **L2/17-347**
- **Mongolian Working Group**
 - Mongolian Working Group Meeting #2 (MWG2).....April 2018, **L2/18-108**
 - Mongolian Working Group Meeting #3 (MWG3).....*tentatively April or May 2019, Ulaanbaatar*
- **Unicode Technical Committee**, quarterly
 - Mongolian ad hoc, UTC #156, *established the latest goals*.....July 2018, **L2/18-254**

V. Efforts: *Expert groups and meetings* [cont.]

Unicode Consortium liaison members and representatives:

- **MASM, Mongolia**.....**B. Undraa** ↔ Debbie Anderson

Стандарт, хэмжил зүйн газар

Mongolian Agency for Standard and Metrology

- **EAC of Inner Mongolia, China**.....**Liang Jinbao** ↔ Liang Hai

ᠨᠢᠮᠤᠩᠭᠣᠯᠢ ᠶ᠋ᠢᠨᠨᠠᠨᠠᠭᠤᠨ ᠨᠠᠨᠢᠯᠠᠭᠤᠨ ᠨᠠᠨᠢᠯᠠᠭᠤᠨ ᠨᠠᠨᠢᠯᠠᠭᠤᠨ ᠨᠠᠨᠢᠯᠠᠭᠤᠨ ᠨᠠᠨᠢᠯᠠᠭᠤᠨ ᠨᠠᠨᠢᠯᠠᠭᠤᠨ ᠨᠠᠨᠢᠯᠠᠭᠤᠨ ᠨᠠᠨᠢᠯᠠᠭᠤᠨ

内蒙古自治区民族事务委员会

Ethnic Affairs Committee of the Inner Mongolia Autonomous Region

V. Efforts: *Expert groups and meetings* [cont.]

Other groups to get in touch with:

- **"China Mongolian Working Group"**

蒙古文信息技术国家标准工作组, literally "Mongolian script information technology national standard working group"

- **W3C Internationalization Interest Group: Mongolian** ↗

Note the *encoding discussion document log* ↗

This mailing list also serves the *Mongolian Layout Task Force* ↗

Also, **Liang Hai and his friends** have *continuous informal discussions* that are more accessible to experts who prefer Chinese to English as working language.

V. Efforts: *Noteworthy standard updates*

- Added Mongolian variation sequences and their positional forms.....
.....**L2/02-012**; published in Unicode 3.2 ([StandardizedVariants-3.2.0.html ↗](#))
- Clarified relative order of FVSes and ZWJ**L2/03-065**
- Changed MVS from gc = Cf to Zs, then back to Cf**L2/13-004**
- Added glyphs of positional forms (incl. originally undefined ones) to names list
.....**L2/14-031**; published in Unicode 9.0
- Removed glyphs of originally *undefined* positional forms from names list.....
.....**L2/17-368**; published in Unicode 11.0

V. Efforts: *Next steps*

- Investigate **NNBSP** behavior and properties.....Unicode 12.0
- Restructure the *Core Specification* **chapter**Unicode 12.0–13.0
- Unicode Technical Note for **documentation**.....draft for MWG3
- **MWG3** (Mongolian Working Group Meeting #3)
.....*tentatively April or May 2019, Ulaanbaatar*

V. Efforts: *Next steps* [cont.]

Long-term investigations:

- Investigate existing attempts of specification as well as potential directions of **improving the encoding model**
 - A set of special character properties for describing the contextual rules
- Explore **alternative encoding models** and, in particular, see whether they are applicable to writing systems beyond the modern Hudum
- ... punctuation usage ... MVS and NNBSF usability ... additional FVSes ... unification issues and new characters ...

V. Efforts: *Additional resources*

- *UTC Document Registry: Topical Document List: Mongolian* ↗
 - *ScriptSource: Unicode Status (Mongolian)* ↗
- *Asmus Freytag, et al.: Mongolian Unicode Project* ↗
- *Richard Ishida: Script links: Central Asia: Mongolian* ↗
- *Andrew West: Mongolian Script* ↗
- *Liang Hai: A summary of national standards related to the Mongolian script* ↗

About me

梁海 *liáng hǎi* · Liang Hai · ल्यांग हाइ · ལྷོ རྩེ

- Freelancing *multilingual font technician*, based in Beijing.
- As a participant of *Script Ad Hoc*, *UTC*, and the *Unicode Editorial Committee* meetings, I help Unicode and OpenType understand complex scripts — especially *Indic (Brahmic)* ones and *Mongolian*.
- I go by my surname *Liang* /liɑŋ/ in English.
- **lianghai.github.io** ↗

Acknowledgments

Debbie Anderson · Greg Eck · Richard Ishida

梁金宝 *Liang Jinbao* · Lisa Moore · روزبه پورنادر *Roozbeh Pournader*

沈逸磊 *Shen Yilei* · Ken Whistler · 郑维喆 *Zheng Weizhe*

Fonts

FF Basic Gothic · Prenton RP · ATF Garamond · 29LT Zarid

***Bolorsoft:* MongolianScript · *MenkSoft:* Menk Vran Tig · Menk Qagan Tig**

***Hasutai:* Sungar Ginggulere hergen · Sunggar Wencin durun · *Mingzai:* Todo Sudur Mingzei**