MWG/3-N7R2
L2/19-129

The refined phonetic model

The most convenient solution for recovering

the current Mongolian encoding

Bolorsoft team

2019, Ulaanbaatar

MWG3 meeting

Agenda

I. About Us (brief)
Who are we? Why are we getting involved?

II. Problems
Why is the Unicode Mongolian encoding broken?

III. Research & analysis
Should Mongolian script be encoded semantically or non-semantically?

IV. Solution & Procedure
What is the most convenient way to recover the current model?

V. Specification
We never had sufficient specification for Mongolian script encoding!

VI. Summary
What are the practical results?

VII. Future Work

- Part I -

Why are we getting involved

Who are we? A team for Mongolian script at Bolorsoft.

I. About Us: **Who are we?**

Bolorsoft LLC

Founded in 2008

Focus: NLP, AI

I. About Us: **Why are getting involved?**

Too many incomplete projects

2004: OTF enhancement of MongolianScript font

2008: Vertical direction support in OpenOffice

2012: Converter between Mongolian script and Cyrillic

2013: Spellchecker for Mongolian script

Reason: We have recognized that the Mongolian encoding is broken.

I. About Us: **Established a team at Bolorsoft**

After MWG#1 meeting in Huhhot 2017.09.24

We informed the true situation of Mongolian script to the president of Mongolia

Result: A working group at CITA, sub committee at MASM

2018.01.19: We have abandoned CITA WG.

I. About Us: The team at Bolorsoft

Ch. Munkhnaran (Coordinator)
T. Jamyansuren (Typography master)
S. Badral (Sofware architect)
T. Serchmaa (Software engineer)
E. Munkh-Uchral (Linguist)
D. Lhagvasuren (Tester)
D. Badarch (Adviser)
M. Erdenechimeg (Adviser)
Yu. Namstrai (Adviser)

- Part II -

Why is the Unicode Mongolian encoding broken?

Is the current model completely broken? What exactly are not working?

II. Problems

Is the current model really broken?

- The coarse-grained problems? [non-technical]
- The fine-grained problems? [technical]

II. Problems: **Is the model really broken?**

Unfortunately, YES.

We and all Unicode experts have acknowledged it.

However, it does not mean the current model is bad.

On the contrary, we confirm that the current Unicode model is most adequate for Mongolian script.

II. Problems: **Is the current model completely broken?**

Fortunately, NO!

We can recover it without any major changes except introducing two new letters and reducing control format characters.

Of course some reorganization of variants are necessary.

II. Problems: **The coarse-grained problems**

What problems the broken encoding cause?

- Decreased the use of Mongolian script
- Disabled IT development
- Lack of interoperability
- Insufficient documentation for implementers
- Increased corrupted data
- Increased social discontent, chaos

II. Problems: **The coarse-grained problems [cont.]**

Decreased the use of Mongolian script

Inner Mongolian scholars acknowledged that the usage of the Mongolian script dramatically decreased.

Reason:

The new generation couldn’t use Mongolian script flawlessly on their devices; they do/can not write by hand.

II. Problems: **The coarse-grained problems [cont.]**

Decreased the use of Mongolian script

In Mongolia, the official script is Cyrillic. However, Mongolian script is taught at every state schools in 6-12 classes.

There are many people, who want to use this script but due to the difficulties in computer environment they use Cyrillic.

II. Problems: **The coarse-grained problems [cont.]**

Disabled IT development

A lot of efforts, products, projects, fonts and web sites are discontinued or cancelled due to Mongolian script encoding problems.

II. Problems: **The coarse-grained problems [cont.]**

Interoperability

- All existing fonts are incompatible each other.
- Non-of them are faultless.
- Non-of them are stable between versions.
- All implementations are incompatible.

II. Problems: **The coarse-grained problems [cont.]**

Insufficient documentation for implementers

- Both national standards are different.
- Both standards are not synchronized with Unicode.
- No single documentation for font developers, however large number of complex OT rules are required.

(There was only one (insufficient) documentation from Microsoft but removed. The address was https://www.microsoft.com/typography/otfntdev/mongolot/)

- No documentation about directions.
  (glyphs-> texts-> frames)

II. Problems: **The coarse-grained problems [cont.]**

Increased corrupted data

- There exist several solutions, which has its own group of users.
- Not only legacy solutions.
- Some are based on Unicode .
- Some are based on PUA.

II. Problems: **The coarse-grained problems [cont.]**

Increased social discontent, chaos

Everybody says Unicode but nobody really understands what he/she says.

[Fact]: After our release, we have received several request to cooperate from some experts who worked on the Unicode encoding. ;-)

II. Problems: **The fine-grained problems**

What exactly is not working?

1. Architectural mistakes
2. Wrong direction of encoding
3. Complex design and poor specification
4. Limitations in the usability
5. Visual ambiguity

II. Problems: **The fine-grained problems [cont.]**

Architectural mistakes

- KE and GE characters are respectively unified to QA and GA.
- There are positional mismatches.
- Too many FVSs are encoded.
- There are inconsistent use of FVSs.

II. Problems: **The fine-grained problems [cont.]**

Architectural mistakes

KE and GE characters are respectively unified to QA and GA.

Don’t underestimate. It is a crucial point of the failure.

II. Problems: **The fine-grained problems [cont.]**

Architectural mistakes

Positional mismatches

Contradictory to the general cursive joining rules.

All digits, punctuations, MVS and NNBSP are non-joining characters.

ZWJ and NIRUGU are join causing characters.

II. Problems: **The fine-grained problems [cont.]**

Architectural mistakes

Too many FVSs

- Using more than one FVS is already known as exhaustive for end users.
- Opened the way to arbitrarily encode unnecessary variants.
- Increase ambiguities.
- Criticized from the beginning: http://www.unicode.org/L2/L1997/97028-n1497-mongolian.pdf

II. Problems: **The fine-grained problems [cont.]**

Architectural mistakes

Inconsistent use of FVSs

No clear rules, where to use FVSs.
Neither intuitive nor rational.

II. Problems: **The fine-grained problems [cont.]**

Architectural mistakes

NNBSP issues

The NNBSP is broken in Unicode standard. It never works flawlessly.

This character is defined as a space character and has additional trio functions.

It is used to display a narrow space.

It is involved in Mongolian shaping.

No reaction for our proposal L2/18-293.

https://www.unicode.org/L2/L2018/18293-nnbsp-solution.pdf

II. Problems: **The fine-grained problems [cont.]**

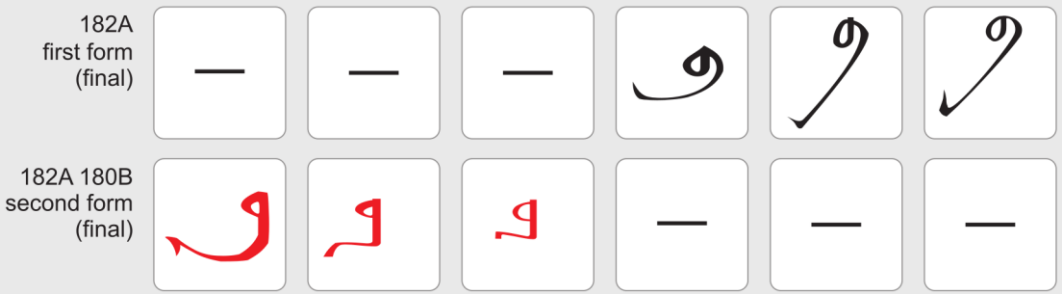
Wrong direction of encoding

- Historical variants and styles are encoded as variants.
- Variant encoding should not be underestimated.
- Increased number of FVSs.
- Mixed aspects of different abstraction levels.

II. Problems: **The fine-grained problems [cont.]**

Encoding of historical and stylistic variants

- It blocks historical and stylistic font development.



(for details: see our “Mongolian aesthetic 800 years”)

- It is one of the main reason to increase the number of FVSs.

II. Problems: **The fine-grained problems [cont.]**

Mixed aspects in different abstraction level

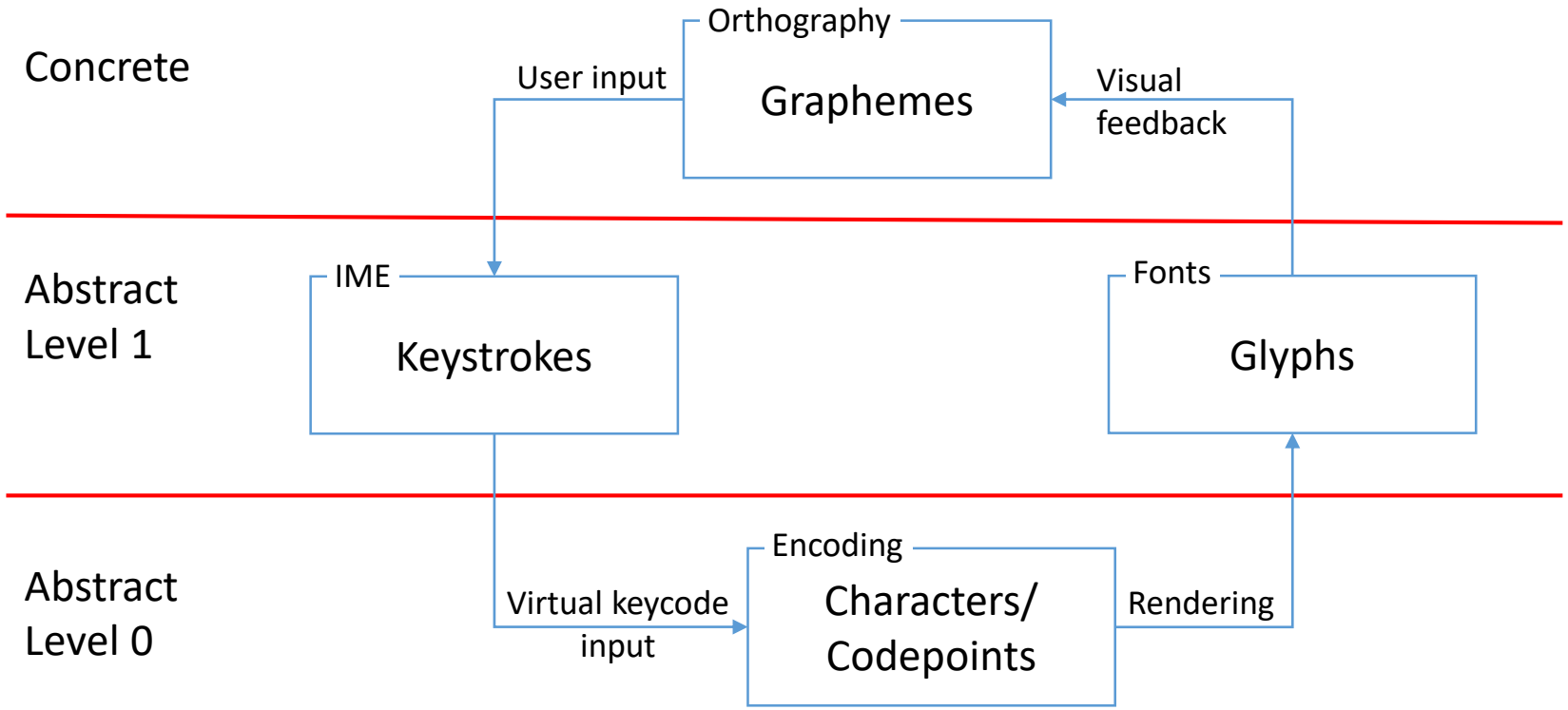


Figure adapted from: https://scripts.sil.org/IWS-Chapter02

II. Problems: **The fine-grained problems [cont.]**

Mixed aspects in different abstraction level

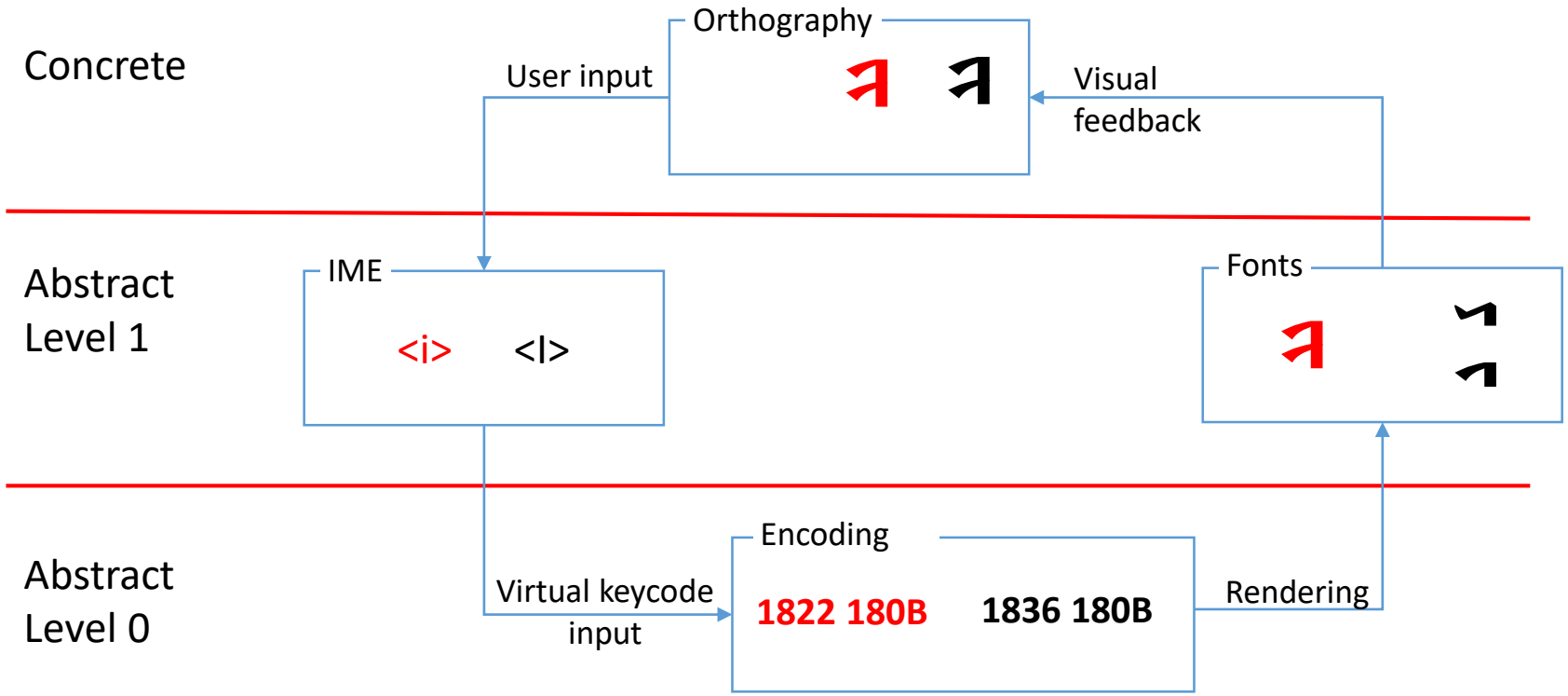


Figure adapted from: https://scripts.sil.org/IWS-Chapter02

II. Problems: **The fine-grained problems [cont.]**

Mixed aspects in different abstraction level

- Phonetically, that doesn’t mean to be encoded the orthographical rules. (diphthongs)
- Phonetically, that doesn’t mean the text renderer checks morphological rules. (gender determination)

II. Problems: **The fine-grained problems [cont.]**

Complex design and poor specification

- Semantic encoding
- Cursive joining
- Too many special characters such as NNBSP, MVS, FVS1, FVS2, FVS3, ZWJ, ZWNJ
- Large number of complex open type rules
- The only vertical LTR orientation
- Texts are horizontally rendered at first then rotated

II. Problems: **The fine-grained problems [cont.]**

Limitations in the usability

- Too many FVSs.
- The usage of FVSs are different in all font.
- Too many unnecessary characters.
- Important variants are well hidden.
- Usage of ZWJ, ZWNJ in Mongolian are not clear.
- Fonts are non-interoperable.

II. Problems: **The fine-grained problems [cont.]**

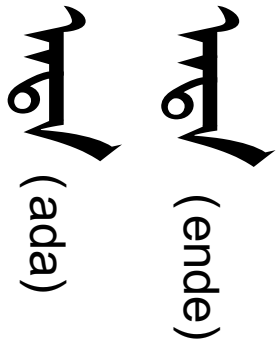
Visual ambiguity

Ambiguity free representation is almost not possible in Mongolian script even in the use of the graphetic model.

However, continuously attempted to eliminate of letter ambiguity.

"Todo" was an essential attempt. Experienced is:

The significant changes to glyphs probably not resistant but distinguishing without a destruction of aesthetics of graphemes is not harmful.



An example

- Part III -

Research and analysis of the Mongolian script

Should Mongolian script be encoded semantically or non-semantically?

III. Research: **Phonetic vs. Graphetic?**

As we know: The Mongolian script is phonetical.
However, we need to do more research to ensure it.

- Are Mongolian characters really phonetical?
- If yes, why current encoding doesn’t work well?
- If no, is the graphetic model a suitable solution?
- Existing encoding solutions for the Mongolian script

III. Research: **Are Mongolian characters really phonetical?**

It is recognized that Mongolian writing system is based on phonetic letters.

- Danjindagba (fl. 1723-1736) - The Space mantra for eliminating of letter ambiguity: “Commentary to Auricle of the heart“
  Transliteration: Jirüken-ü tolta-yin tayilburi üsüg-ün endegürel-i arilɣaɣci Oɣtaryui-yin mani
- The Mongolian grammar: “Jirüken-ü tolta” Choiji-Odser (1307-1321)
- Sakya Pandita Kunga Gyaltsan (1182-1251)

III. Research: **Are Mongolian characters really phonetical?**

What is a phonetic letter in Mongolian script?

- If a character is a minimal unit of text that has semantic value. On that note, an orthographical character is a letter in Mongolian script.
- A Mongolian character is well constructed by some elements/pictograms that don’t have semantic value.
- Any analysis based on these elements leads to non-semantical encoding.

III. Research: Are Mongolian characters really phonetical? [cont.]

Conclusion

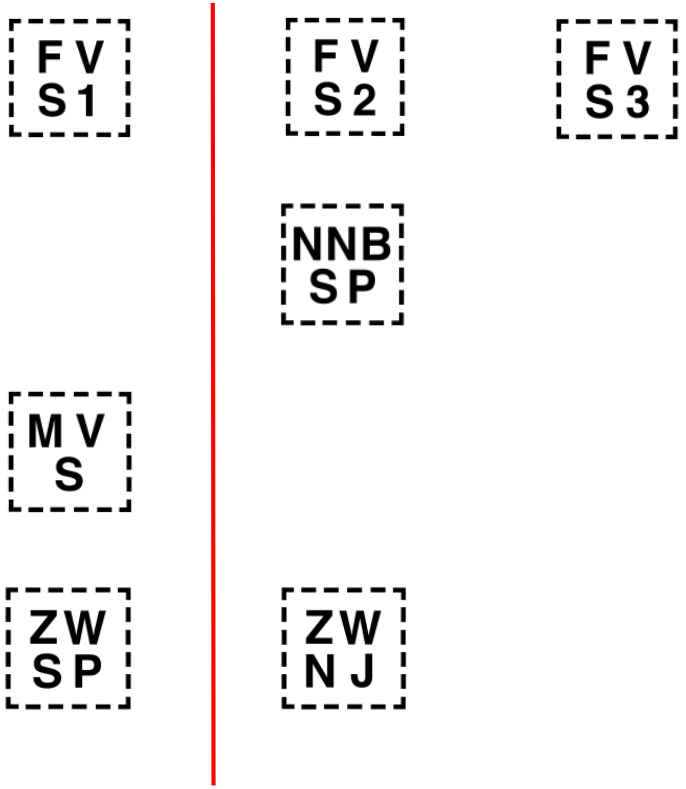
The Mongolian script is really phonetical.
It is not artificially formulated.

Mongolian script should be encoded semantically same as now is!

III. Research: **Then, why current encoding doesn't work well? [cont.]**

Are all format control characters really required?

- FVS1, FVS2, FVS3
- NNBSP
- MVS
- ZWJ, ZWNJ



III. Research: **Then, why current encoding doesn't work well? [cont.]**

Are KE and GE characters really necessary?

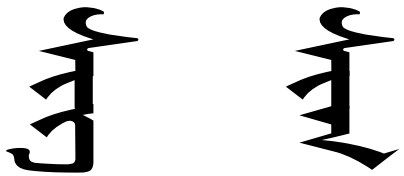
[Analysis] QA and GA are the most frequently used characters

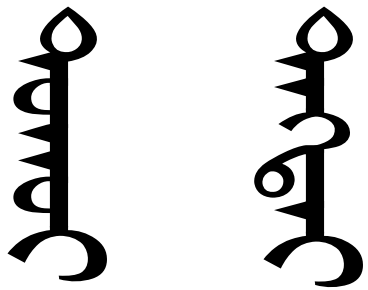
No.	Letter	Name	Code point	Occurrence
1	ᠠ	MONGOLIAN LETTER A	U 1820	44245
2	ᠢ	MONGOLIAN LETTER I	U 1822	29090
3	ᠭ + ᠎	MONGOLIAN LETTER GA	U 182D	11105+16507
4	ᠡ	MONGOLIAN LETTER E	U 1821	23269
5	ᠤ	MONGOLIAN LETTER U	U 1824	19942
6	ᠯ	MONGOLIAN LETTER LA	U 182F	19865
7	ᠷ	MONGOLIAN LETTER RA	U 1837	18383
8	ᠺ + ᠎	MONGOLIAN LETTER QA	U 182C	7069+6243
9	ᠳ	MONGOLIAN LETTER DA	U 1833	11596
10	ᠦ	MONGOLIAN LETTER UE	U 1826	10988

III. Research: **Then, why current encoding doesn’t work well? [cont.]**

Are KE and GE characters really necessary?

[Analysis] KE and GE are also used to identify the gender of a word more reliable and faster than vowels.





III. Research: **Then, why the current encoding doesn't work well? [cont.]**

Are KE and GE characters really necessary?

Nobody wants to type FVS for QA and GA due to its frequency.

It brings to font developers a challenge of handling these characters without FVSs.

However, it is impossible with limited number of OT rules.

III. Research: **Then, why the current encoding doesn't work well? [cont.]**

Are KE and GE characters really necessary?

Unification of KE and QA, GE and GA had treated <u>unbelievably bad</u>

182C MONGOLIAN LETTER QA
→ 0445 х cyrillic small letter ha
~ 182C first form (isolate)
~ 182C first form (initial)
~ 182C first form (medial)
~ 182C first form (final)
~ 182C 180B second form (isolate)
~ 182C 180B second form (initial)
~ 182C 180B second form (medial)
~ 182C 180C third form (medial)
~ 182C 180D fourth form (medial)

182D MONGOLIAN LETTER GA
→ 0433 г cyrillic small letter ghe
~ 182D first form (isolate)
~ 182D first form (initial)
~ 182D first form (medial)
~ 182D first form (final)
~ 182D 180B second form (initial)
~ 182D 180B second form (medial)
~ 182D 180B second form (final)
~ 182D 180C third form (medial)
~ 182D 180D fourth form (medial)

III. Research: **Then, why the current encoding doesn't work well? [cont.]**

Are KE and GE characters really necessary?

QA and GA are the most frequent and most important characters.
However, these characters contain all possible mismatches!

- Isolated forms are defined as feminine but masculine are actually used.
- Historical/periodic forms are encoded at FVS1, FVS2 slots.
- No final form is defined for QA.
- Positional mismatches for both letters.
- Totally different graphemes are incorporated.

III. Research: **Then, why the current encoding doesn't work well? [cont.]**

Are KE and GE characters really necessary?

All these problems lead to:

- Adding unnecessary FVSs to Mongolian block

  The main fault of usage complexity
- Unlimited number of OT rules

  The main fault of non-interoperability of fonts and implementations

III. Research: **Then, why current encoding doesn't work well?**

Unravelling of a mystery to visual ambiguity

- In brain the letters O, U, OE, UE are separately encoded.
- On paper O and U, OE and UE are identically encoded.
- The reason is to be liberally accepted by all parties, which are unified by the Mongol Empire.
- The trick worked well until computerization era and then it is transferred to a bug.

III. Research: Then, why current encoding doesn't work well? [cont.]

Conclusion

The reason is not because of a bad model
but because of some small bugs with big impacts.

Generally, a small bug at encoding level can generate anyways huge impacts.

The crux of the matter is:

- The (badly treated) unification of KE and QA, GE and GA
- Encoding of historical characters
- Positional mismatches

III. Research: **Is the graphetic model a suitable solution?**

There are notable good points

- Excellent analysis of graphemes
- Correcting the positional mismatches is required for any attempt at improving the current encoding whatever the model is.
- The intermediate table for variant set is close to the refined model table.
- Great decompositions

III. Research: **Is the graphetic model a suitable solution? [cont.]**

Why is this model unacceptable?

- Performed excessive minimization to remove visual ambiguities, however could not be solved completely due to the Mongolian script nature.
- Non-semantical encoding but a cursive joining model
- Probably a non-cursive joining model more suitable for non-semantical encoding
- Leads to disadvantages of non-semantic (inaccuracy, looseness, approximations for further processing)

III. Research: **Is the graphetic model a suitable solution? [cont.]**

There are some similarities to the refined model

Refined model	**Graphetic model**
Clean up historical and stylistic variants to reduce FVSs.	Fix positional mismatches
Fix positional mismatches	Eliminate visual ambiguities.
Disunify KE and GE from QA and GA and reintroduce KE and GE	Deep minimizations by decomposing and merging graphemes
Result: Inflict minimal changes to the current encoding	Result: introduce a non-semantical new encoding

III. Research: Is the graphetic model a suitable solution? [cont.]

Conclusion

The graphetic model is not a suitable solution for Mongolian script due to unusual behavior of cursive joining effect.

III. Research: About the existing encoding solutions for Mongolian script [cont.]

Case study

- Myanmar disaster
- Arabic encoding

III. Research: About the existing encoding solutions for Mongolian script [cont.]

A brief tour of previous attempts

- ASCII/ANSI based old solution (Peter Chang)
- PUA based legacy solution (Menkhgal)
- Unicode based solution (Microsoft, Google, Almas, Menkhgal, Bolorsoft, …)
- Modified Unicode solution (Others in inner Mongolia)

- Part IV -

Solution & Procedure

What is the most convenient way to recover the current model?

Which procedures are required to achieve our goal?

IV. Solution

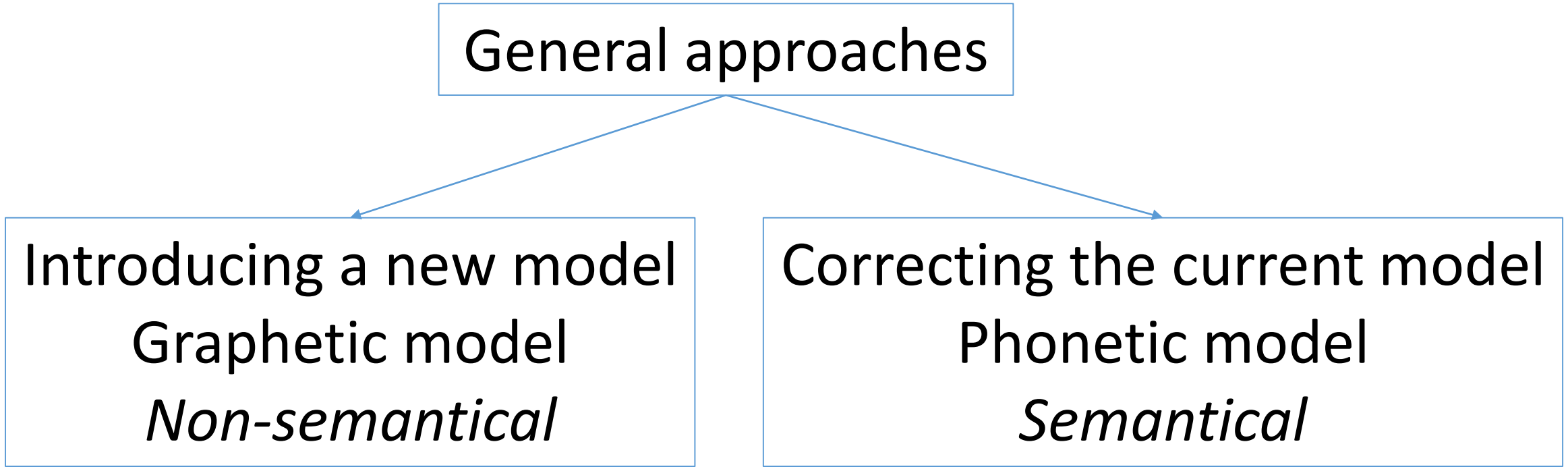
- Vision
- Possible approaches
- Comparisons of solutions
- Procedures

IV. Solution: **Vision**

To begin with the end in mind

- Time to market (without delay after implementation)
- Reliable representation (cursive but no shaping problems)
- Robust encoding (not switchable by regional use or by other reason)
- Reliable operating (no approximation for collation, sorting, …)
- Easy for the end users (maximal one FVS, workaround for invisible characters)
- Easy for the development (using standard string functions)
- Rapid migration (a tool, which converts from other existing encoding to our encoding)
- Interoperable implementations (reducing font rules significantly up to 25 lookups,)

IV. Solution: **Possible approaches**



IV. Solution: **Available solutions**

- The graphetic model (Liang Hai)
- The improved phonetic model (Prof. Que)
- The refined phonetic model (Bolorsoft)

IV. Solution: **Comparison**

Criteria [fn]	**Improved model**	**Graphetic model**	**Refined model**
Reliable representation	★☆☆☆☆	★★★★☆	★★★★☆
Robust encoding	★★☆☆☆	★★★★☆	★★★★☆
Reliable operating	★★★☆☆	☆☆☆☆☆	★★★★☆
Interoperable	★☆☆☆☆	★★★★☆	★★★★☆

IV. Solution: **Comparison [cont.]**

Criteria [non-fn]	**Improved model**	**Graphetic model**	**Refined model**
Time to market	★★★★★	☆☆☆☆☆	★★★★★
Easy for the end users	★☆☆☆☆	★★★☆☆	★★★★☆
Easy for the further development	★★★☆☆	★★☆☆☆	★★★★☆
Rapid migration	★★★☆☆	☆☆☆☆☆	★★★★☆

IV. Solution: **Procedures**

1. Cleaning up stylistic and historic variants
2. Fixing positional mismatches is mandatory regardless of which solution will be applied
3. Rentroducing two letters
4. Replacing NNBSP by MVS
5. Specification and documentation
6. Prototyping
7. Preparation of migration
8. Launch / Updating Unicode standard

IV. Solution: **Cleaning up stylistic and historic variants**

Code point	Char	Variant	Reasons
1820 180B		Second isolate form of A	It is Mongolian Ali Gali form.
1821 180B		Second initial form of E	It is a historical char. (pre-classical)
1826 180B		Second isolate form of UE	Used for Chinese Wu syllable. Should be defined at “W” if required.
1828 180B		Second initial form of NA	It is a historical char. (pre-classical)
1828 180D		Fourth medial form of NA	Todo separated suffix.
182A 180B		Second final form of BA	It is a historical char. (pre-classical)

IV. Solution: **Cleaning up stylistic and historic variants [cont.]**

Code point	Char	Variant	Reasons
182C 180B		Second isolate form of QA	It is a historical char. (pre-classical)
182D 180B		Second initial form of QA	It is a historical char. (pre-classical)
182C 180B		Second initial form of QA	It is a historical char. (pre-classical)
182C 180B		Third medial form of QA	It is a historical char. (pre-classical)
182D 180B		Second initial form of GA	It is a historical char. (pre-classical)
182D 180B		Second final form of SA	It is a historical char. (pre-classical)

IV. Solution: **Fixing positional mismatches**

Code	Char	Name	Currently	To be
1820	ᠠ	Mongolian A	Medial (third form)	Initial (second)
1820	ᡀ	Mongolian A	Final (second form)	Isolate (second)
1821	ᡁ	Mongolian E	Final (second form)	Isolate (second)
1822	ᠡ	Mongolian I	Missing (used medial)	Initial (second)
1822	ᡂ	Mongolian I	Missing (used final)	Isolate (second)
1824	ᡅ	Mongolian U	Missing (used final)	Isolate (second)
1824	ᡄ	Mongolian U	Missing (used medial)	Initial (second)
1826	ᡆ	Mongolian UE	Missing (used final)	Isolate (second)

IV. Solution: **Fixing positional mismatches [cont.]**

Code	Char	Name	Currently	To be
1826	ᠦ	Mongolian	Missing (used medial)	Initial (second)
1828	ᠨᠠ	Mongolian NA	Medial (third form)	Final (second)
182C	ᠬᠠ	Mongolian QA	Medial (fourth form)	Final (second)
182D	ᠭᠠ	Mongolian GA	Medial (third form)	Final (second)
1835	ᠵᠠ	Mongolian JA	Medial (second form)	Isolate (second)
1836	ᠶᠠ	Mongolian YA	Medial (third form)	Final (first)

IV. Solution: **Reintroducing two letters**

Letter KE and GE

- In 2018, we have already proposed to disunify QA and GA and reintroduce KE and GE.

  L2/18-294: https://www.unicode.org/L2/L2018/18294-two-mongolian-ltrs.pdf

- In 2019, we have discussed about code points and permitted to use code points from Mongolian Ali Gali block.

  L2/19-058: https://www.unicode.org/L2/L2019/19058-mongolian-ad-hoc-rept.pdf

IV. Solution: **Reintroducing two letters [cont.]**

Code points for letter KE and GE

- We have carefully examined Mongolian Ali Gali section and determined that the code points 1887-188A non-essential.
- The 1887, 1888 and 188A are styles and **1889** MONGOLIAN LETTER ALI GALI KA is actually Mongolian **GE** letter. The Form is also identical. Thus, we put MONGOLIAN GE to its own place and just selected the previous neighbor **1888** as MONGOLIAN LETTER **KE**.

IV. Solution: **Replacing NNBSP by MVS**

About proposal

- In 2018, we have already proposed to solve NNBSP issues introducing the MSC - Mongolian Suffix Connector.

  L2/18-293: https://www.unicode.org/L2/L2018/18293-nnbsp-solution.pdf

- In “Proposed solution” on page 14 we have already mentioned that in meantime the MVS could be used as NNBSP.
- However, we didn’t receive any approval or constructive feedbacks to make decision.

IV. Solution: **Replacing NNBSP by MVS [cont.]**

Explanatory statement

Similarity:

- The functionality of MVS and NNBSP is very similar.
- MVS joins disjoint A and E.
- NNBSP joins the suffixes to its stem word or preceding suffix.

Differences:

MVS is a format control character, NNBSP is a space separator.

IV. Solution: **Replacing NNBSP by MVS [cont.]**

Why is it not included in pre-release

- Pre-release was only for windows and mac OS.
- It was an chance to investigate.
- Respecting existing users.
- No response from UTC regarding our proposal.
- We have not yet informed to UTC.
- We wanted to discuss about it at this meeting.
- Anyways, short time update is not harmful.

IV. Solution: **Replacing NNBSP by MVS [cont.]**

Study results

- The majority of users was new.
- Almost every new user asked about how to write suffixes.
- Almost 70% of users didn’t distinguish between MVS and NNBSP. NNBSP is simply used instead of MVS.
- Few users confused between KE and GE.
- The existing users are rather critical. They are still testing our solution.
- The majority of users were happy for the simple writing.
- NNBSP is still not supported well by major vendors like Webkit for MacOS and iOS. (Older versions were flawless. It means NNBSP implementation instable.)

IV. Solution: **Replacing NNBSP by MVS [cont.]**

Can MVS fulfill all requirements of MSC?

As mentioned in our proposal:

- In encoding level just one marker is enough for suffix joining.
- MVS solves all word boundary problems and word selection problems.
- For the shaping, MVS is far reliable than NNBSP.
- All other functionality solved by fonts and space is manipulated by fonts.
- The only deficit is line breaking. Anyways, NNBSP has also line breaking issue.

IV. Solution: **Replacing NNBSP by MVS [cont.]**

Decision and Release plan

- From our study, we have decided to replace NNBSP by MVS.
- We have planned to release the Linux version in next week.
- Mobile versions are in demand. Release comes in four weeks.

IV. Solution: **Specification & documentation**

- Cleaning up unnecessary variants.
- KE and GE letters are disunified.
- Reducing format control characters: FVS1, MVS, ZWJ, NNBSP, ZWNJ, FVS3, FVS2
- Basic rules for writing

- Documentation: https://wiki.mngl.net

IV. Solution: **Prototyping**

Keyboard (for windows, MAC OS)

IV. Solution: **Prototyping**

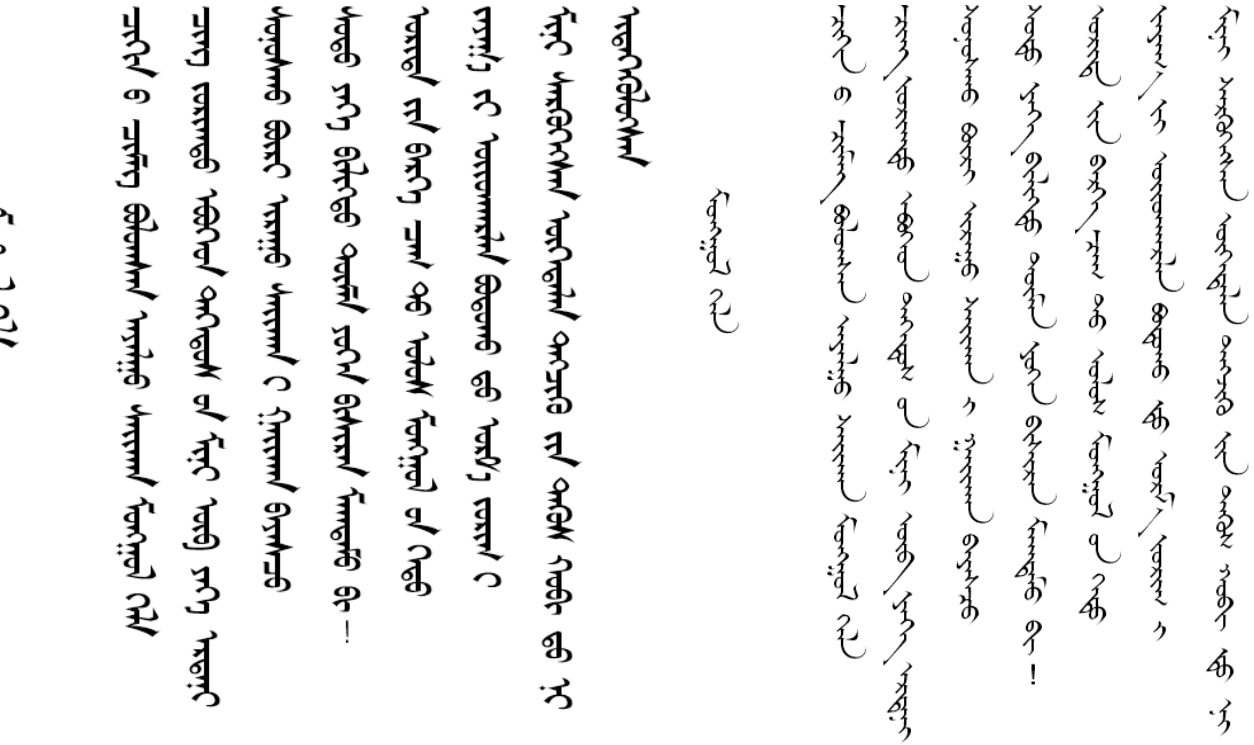
Keyboard (for mobile platforms)

- Prepared simple mobile keyboards (waiting for apple and google approvals).
- Planning to release smart input methods.
- Planning to release a spellchecker with keyboards.

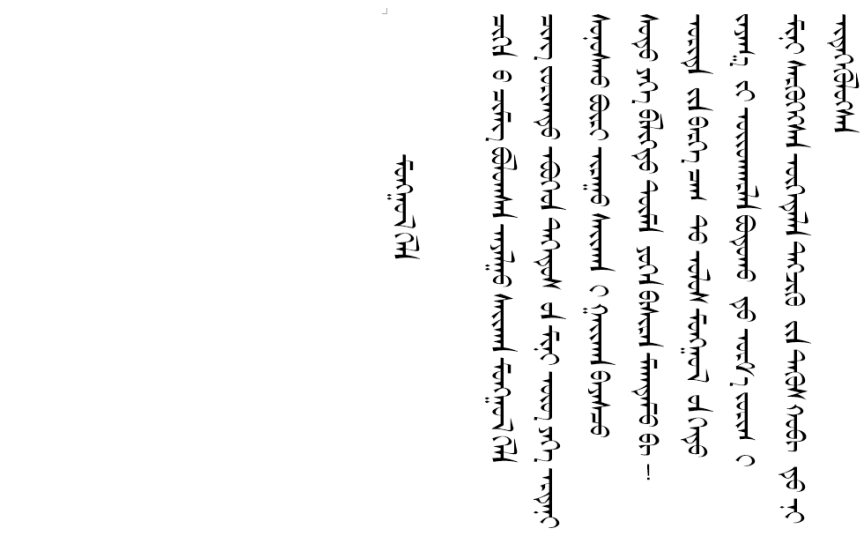
IV. Solution: **Prototyping**

Fonts

Two fonts are currently released.



More than 10 fonts will be released in the next 2 months.



IV. Solution: **Preparation of migration**

- Text encoding conversion
- Documentation
- Including in school programs

IV. Solution: **Preparation of migration**

Text encoding conversion

We are developing free text conversion utilities for all platforms.
https://tools.mngl.net

Progress:

The existing Unicode encoding (Mongolian Baiti, Noto Sans Mongolian, older MongolianScript) → The refined Unicode encoding (completed)

PUA (Menkhsoft and similar solutions) → The refined Unicode encoding

ASCII (Ulaanbaatar, Urguu, etc.) → The refined Unicode encoding

IV. Solution: **Preparation of migration**

Documentation

Documentations and manuals will be released under:
https://wiki.mnlg.net

The project is already started.

IV. Solution: **Preparation of migration**

Including in school programs

Cooperation of educational institutes

Bolorsoft has already signed Memorandum of Agreement with the **Mongolian Institute for Education** of Ministry of Education Culture and Science of Mongolia.

- Part V -

The specification of the Refined Phonetic Model

Core specification, …

V. The refined model

- Terminology and definitions
- Encoding principles
- The character set
- Specific characteristics
- Directionality
- Basic rules

V. The refined model: **Terminology and definitions**

Some clarifications

- Mongolian Writing Systems – Hudum
- Separated suffixes, Not enclitics!
- Cursive joining

V. The refined model: **The Encoding principles**

The principles

- Semantical encoding
- Separation of concerns

V. The refined model: **The Hudum character set**

The character set

- Format controls
- Punctuation
- Digits
- Basic letters

V. The refined model: **Specific characteristics**

Specific characteristics

- MVS
- FVS
- ZWJ
- ~~NNBS~~

V. The refined model: **Directionality**

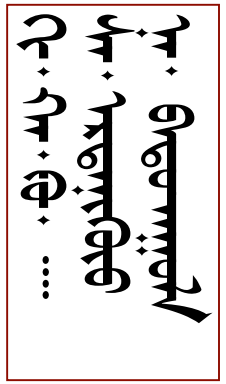
Developer guidelines are necessary

- Font glyphs are rotated 90° counterclockwise.
- Text frames rotate the text line-wise to express vertical orientation

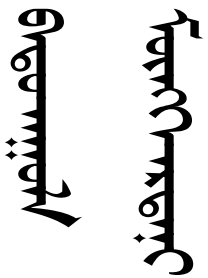
V. The refined model: **Basic rules**

Some basic rules

Abbreviations



Compound words



ML_BA ML_A ML_TA
ML_U ZWJ ML_A ML_GA
ML_U ML_LA ML_A.
The ZWJ is substituted by toothed forms by open type rules.

- Part VI -

Summary

Implementations, proof of concept, tests,

live action, products, …

VI. Summary

- *If the cursive joining model is applied for Mongolian script, then it has to be encoded semantically.*
- *Current model is broken, not because of bad model but because of some bugs. In contrary, it is most adequate model for Mongolian.*
- *The refined model is just the bug fixes of current encoding model.*
- *In any case, the cleaning up of stylistic or historical variants is necessary to recover current model.*
- *The visual ambiguities could be significantly reduced by font manipulation.*

VI. Summary [cont.]

- *Current specification covers only Hudum block.*
- *We have fully implemented our solution.*
- *We have tested our solution in-house.*
- *We have released our solution as “proof of concept” (only for Windows and Mac OS) to collect end user feedbacks (for MVS vs NNBSP, FVS reduction, KE, GE).*
- *The refined model has accelerated our product development significantly.*

- Part VII -

Future Work

What is next?

VI. Future Work

- *Enhancing the analysis for other script blocks such as “Todo”, “Sibe”, “Manchu” and “Ali Gali”*
- *Documentation and wiki updates*
- *User handbook, font developer guidelines*
- *Updating Unicode standard, if our solution succeed*
- *Verifying phags-pa, soyombo and vagindra script*
- *Developing fonts and keyboards for historical scripts*
- *We will provide and support all the products, which will use the refined phonetic model.*

References

- https://scripts.sil.org/IWS-Chapter02
- All documents under: *https://unicode.org/L2/topical/mongolian/*
- https://www.unicode.org/Public/12.0.0/ucd/ArabicShaping.txt

Acknowledgements

- *Special thank to UTC stuff for the topical document registry. https://unicode.org/L2/topical/mongolian/*
- *Myatav Erdenechimeg*
- *Lisa Moore*
- *Debroh Anderson*
- *Liang Hai*