To: **Unicode Technical Committee**
From: **Ken Whistler, Peter Constable, Roozbeh Pournader, Liang Hai, Shen Yilei, Zheng Weizhe, Richard Ishida, and Debbie Anderson**
Date: **21 July 2017**
Subject: Script Ad Hoc report and recommendations on Mongolian

The script ad hoc held meetings on Mongolian on 2 June[1] and 14 July 2017. The following summarizes the outcome of the discussion and our recommendations to the UTC, based on the 14 July meeting, with additional input provided at the script ad hoc meeting on 21 July.

The high-level goal is to make Mongolian text and documents interchangeable, the script implementable, and platforms interoperable.

Liang Hai and colleagues presented their analyses of different alternatives to fix Hudum. The term "Hudum" explicitly refers to the modern writing system of the modern Mongolian language using the Mongolian script. The Hudum writing system is particularly complex and ambiguous, and most of the lingering problems for the encoding of the Mongolian script in the Unicode Standard, and for the OpenType-based implementations of Mongolian stem from this complexity. Other writing systems (Todo, Manchu, Sibe, etc) using the Mongolian script generally have evolved to use much more unambiguous orthographies (either alphabetically or syllabically unambiguous), and therefore might need different treatment. Analyses and discussions in this report focus on the Hudum writing system, and are incomplete in terms of side effects on other writing systems.

The attached appendix (pp. 7-8) is based on their work, and shows the changes required for one approach (Graphetic). The chart by Liang Hai on page 9 illustrates a different model (Ideal Phonetic).

The following summarizes some of the key features of these approaches:

1. Graphetic approach

This new approach identifies the letters of Mongolian simply based on their shapes and shaping behavior, rather than based on their readings and phonetic values in the modern Mongolian language (or other languages). This is the typical encoding approach used in the Unicode Standard for all cursive joining scripts descended from Aramaic, including the Syriac, Sogdian and Old Uyghur historical precursors to Mongolian. With this approach, a simple adaptation of the Arabic cursive-joining model can unambiguously handle all of the required shaping behavior, with a relatively small set of contextual rules.

An analysis of this approach turns up two strategies to take regarding any additional encoding requirements:

> a. Minimal new characters: This strategy minimizes the number of new Mongolian characters to be encoded: only 2 additional Mongolian characters would be required for Hudum Mongolian. It would not require use of variation sequences for modern Hudum text. This strategy will re-use

[1] Participating in the 2 June meeting were: Deborah Anderson, Liang Hai, Ken Whistler, Roozbeh Pournader, Peter Constable, Greg Eck, Marek Jeziorek, Lisa Moore, Orlog, Jiromtu, Kat Momoi, Ma Xudong, Shen Yilei, Wang Yihua, and Ben Yang.

already-encoded characters but will define a new set of shaping rules for those characters. The new rules would be much simpler than rules currently being used by implementers, but would also produce significantly different display results in many cases (as shown in examples on page 10). A consequence is that this would be a breaking change for existing documents in that character sequences would have different presentation under the new rules. A corollary is that existing documents could only be displayed as intended when using specific, existing fonts. (This level of non-interoperability is, however, already a problem with the existing encoding model for Mongolian.) It results in alternate spellings of Hudum text using a minimally-different character repertoire. The reuse of most characters provides fewer cues when examining plain text as to which encoding model – the old one or the new one – is in use; this could create confusion for implementations in interpreting text.

b. Maximal new characters: This strategy requires more character additions for encoded representation of Hudum Mongolian: 15 new characters. It would not change any rendering rules or variation sequences for existing characters, so in that sense is not a breaking change for existing data and fonts. It would not use those characters that do not, in their current definition, align with a graphetic model; new recommendations for representation of Hudum Mongolian would use those existing characters that are consistent with the graphetic approach, together with the 15 new characters. New or updated implementations could potentially support existing data and fonts (*modulo* the limits in interoperability of the current encoding model) as well as new data and fonts. It results in alternate representation of Hudum text: using the old model, with inherent limits in interoperability, and using the new, recommended model, presumably with eventual, widespread interoperability. Because many common-use characters are encoded differently using this strategy, more cues are available in plain text to easily distinguish old text from new text.

Either the minimal or maximal strategy would reuse a significant number of the existing consonant letters which currently do not cause problems for the model. (See pages 7-8 for a chart showing changes to Mongolian for both the Minimal and Maximal Graphetic approaches.)

- Pluses
    - Cleaner, unambiguous representation of text
    - Vastly simpler font implementation, with only local contextual rules
    - No variation sequences required for modern Mongolian (except to support old model for backwards compatibility)
    - More straightforward user experience; type what you see

- Minuses:
    - Radical reorganization of the existing Mongolian model
    - Departure from Chinese standard for Mongolian
    - Long-term migration issue for data and documents
    - Uncertain reception from users and implementers in China familiar with the existing model
    - Would require encoding new Mongolian letters (2 for the minimal strategy, up to 15 for the maximal)

2. Phonetic approach

The current model of the Hudum writing system for Mongolian distinguishes letters based on their readings (current or historic), rather than their shapes per se. In some cases, as for vowels, and notoriously complex pairs, such as Q/G and T/D, this results in very complex shaping rules that sometimes depend on morphological analysis, on long-distance contextual rules (to account for Mongolian vowel harmony), and which also require the definition of numerous variation sequences for cases where the contextual rules fail to produce the expected form, or to deal with historical variant forms.

The current model for the Hudum writing system in Unicode employs the phonetic approach. This approach is inherently complex, ambiguous, and difficult to implement successfully. However, there are a couple of strategies that could be taken to improve it incrementally, without changing the overall model.

a. The first of these is the Improved Phonetic strategy, which adds no characters, but makes the minimal set of changes and additions to the existing set of variation sequences to make them more consistent

b. The second of these is the Ideal Phonetic strategy, which also requires encoding no new characters, but which would effectively deprecate all existing Mongolian variation sequences in favor of a completely re-rationalized set of variation sequences (using FE00..FE0F VSs) to make the required processing and shaping more comprehensible and consistent than the current state of affairs.

The chart on page 9 drafted by Liang Hai shows a rationalization of all the glyph forms needed to represent the characters for the Hudum writing system, independent of the contextual rules. In the Ideal Phonetic model, these glyph forms would be associated with a new set of variation sequences, using the VSes from the FE00..FE0F block, rather than the Mongolian Free Variation Selectors (FVS) from the Mongolian block. With a complete and consistent set of variation sequences available, the contextual rules required for the current phonetic model could be rationalized.

3. The pros and cons for the three phonetic strategies are listed below:

3.1 Status Quo

- Pluses
  - No change to the existing model
  - Basically consistent with the current Chinese standard for Mongolian

- Minuses
  - Model is very complex, is underspecified, and has resulted in ambiguous representation of text
  - Assumes very complex contextual rules that have to be implemented in fonts and rendering engines, but without clear specification of those rules
  - Requires large set of variation sequences to force particular outcomes for gaps in the rules

- Very difficult for implementers and font providers
- Results in de facto non-interoperability of different implementations and fonts
- Is subject to attempts to "fix" interoperability by an unstable stream of patches with unstable rationales

3.2 Improved Phonetic

- Pluses
  - No radical change to the existing model
  - Could be consistent with an updated Chinese standard for Mongolian
  - Requires no new characters and only relatively small list of variation sequence additions
  - Fixes for variation sequences lead to a chance for marginally better interoperability
  - Has a chance to be seen as the "consensus" fix, clamping down on the dribbling of patches

- Minuses
  - Basically retains the existing model, with most of its problems
  - To be useful would still require extended documentation of the implied contextual rules
  - Has the chance of just becoming another "flavor" of Mongolian, incompatible with the other incompatible implementations, without true convergence
  - Would still change rendering of existing text because of differing interpretation of the use of VSes

3.3 Ideal Phonetic

- Pluses
  - Introduction of a rationalized set of variation sequences could simplify implementation somewhat
  - Marginally easier to explain
  - Marginally better user experience
  - Would allow compatibility implementations for old behavior without disturbing existing variation sequences

- Minuses
  - Basically retains the existing model, with many of its problems
  - To be useful would still require extended documentation of the implied contextual rules
  - Much less likely than the Improved Phonetic strategy to gain support in the Chinese standard for Mongolian
  - Would still change rendering of existing text because of differing interpretation of the use of VSes

The assessments of the Script Ad Hoc resulting from this review of Mongolian models are:

1. We have unanimous agreement that the status quo is unsustainable. Even with no change to the model, existing text is not interoperable. Furthermore, the requirement for patches to contexual

rules, fonts, and sequences, will make text unstable for the foreseeable future.

2. The Graphetic approach is the preferable model; it is far simpler, is internally coherent, and allows unambiguous representation of Mongolian for the Hudum writing system without use of variation sequences. However, it may be difficult to get support for it from experts and companies in China. And there are migration problems for it. And there are significant migration problems for the graphetic approach, which differ greatly in detail, depending on whether the minimal new character strategy or the maximal new character strategy is taken.

3. The Improved Phonetic strategy retains most of the complexity of the existing Mongolian model, but would improve its interoperability if widely adopted. For successful interchange and convergence of implementations, it would require clearly specified contextual rules for fonts and rendering engines to be defined and enforced. This may be the easiest alternative to gain support for in China.

4. The Ideal Phonetic strategy would simplify parts of the existing Mongolian model and make implementations marginally easier. It would, however, require abandoning the current hodge-podge of variation sequences for a new, more coherent set. It would also require clearly specified contextual rules for fonts and rendering engines to be defined and enforced. Because of the radical departure from use of currently defined variation sequences, it likely would meet more pushback from Chinese stakeholders. The main function of the Ideal Phonetic strategy might be to simply serve as a reference tool to constrain what kinds of additions and redefinitions for variation sequences could actually be considered true "improvements" for the Improved Phonetic strategy.

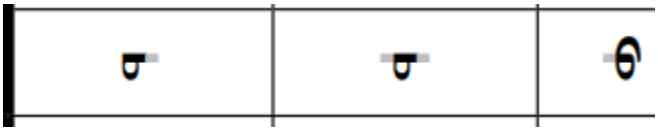
5. Regardless of the strategy taken for improving the representation of modern Mongolian for the Hudum writing system, variation sequences may still be needed to represent historic variant glyphs in historic materials. These variation sequences, however, would represent appropriate use of the mechanism to pick out specific historic variants, rather than complicated set of contextually driven glyph choices for modern text forced by the Phonetic approach of the current Mongolian model.

6. In any case, a review of the impact of any model changes should be undertaken to verify their impact on the writing of other languages using the Mongolian script (Todo, Sibe, Manchu). Our initial assessment, however, is that the use of the script for those languages is far simpler than for the Hudum writing system, and it is unlikely that model changes for Hudum will impact them very much.

Recommendations: We recommend that the UTC discuss the implications of the various strategies for changes in the Mongolian text model for the Hudum writing system.

We also recommend, if possible, an improvement for the glyph display in the Mongolian code charts, to change the cursive connecting portions of glyphs to be shown. This could be in black or another shade, as illustrated below, or another font. (See Liang Hai's chart of glyph forms on p. 9 for a complete

example.) Doing so would help everybody to better understand the relationship between the basic letter shapes and their contextual shaping in cursive contexts.



Changes to Current Mongolian Block
for Minimal and Maximal Graphetic Approaches

init	medi	fina	isol	Graphetic characters	**Minimal** 2 new chars	**Maximal** 15 new chars
O group:						
ᠥ	ᠥ	ᠥ [ᠣ]	-	O-U-Ö-Ü tailless	U+1824	**new**
-	-	ᢀ	ᢀ	O-U-Ö-Ü tailed	U+1825	**new**
-	ᢁ	ᢁ [ᢀ]	-	Ö-Ü	U+1826	**new**
T group:						
ᡩ	ᡩ	ᡩ	-	T-D taw	U+1832	**new**
-	ᡮ	ᡮ	-	D taw	**new**	
ᡱ	ᡱ	ᡱ	-	T-D lamed	U+1833	**new**
E group (the Aleph part is separated from vowel letters):						
ᡷ	-	-	-	E long stem	U+185D	**new**
ᠠ	ᠠ	ᠠ [ᡂ]	ᡂ	A-E, Aleph, N dotless, Cap of H	U+1821	**new**
ᡃ	ᡃ	ᠠ.	-	N dotted	U+1828	**new**
×	×	×	ᡄ	A-E disjointed tail	**new**	
I group:						
ᠢ	ᠢ	ᠢ [ᡃ]	ᠢ	I-Y-J	U+1822	**new**
ᡯ	ᡯ	-	-	Y	U+1836	**new**
-	ᡭ	ᡭ	-	J	U+1854	
W group:						
ᡅ	ᡅ	ᡅ	-	W, Ë	U+1838	**new**
X group:						
ᠺ	ᠺ	ᡀ	-	X-G	U+1889	
ᡰ	ᡰ	ᡰ	-	X-G tense	U+182C	**new**
ᡰ᠈	ᡰ᠈	ᡰ.	-	G tense	U+182D	**new**

H group:

ᡂ	ᡂ	ᡂ	-	H capless, ZH	U+1841

NG group (decomposed):

-	ᡅ	ᡅ	-	<N dotless, X-G>	<1821, 1889>	<**new**, 1889>
ᡇᡂ	ᡇᡂ	-	-	<L, H capless>	<182F, 1841>	

Other letters (no changes needed):

᠊	᠊	᠊	᠊	Nirugu	U+180A
ᡆ	ᡆ	ᡆ	-	B	U+182A
ᡉ	ᡉ	ᡉ	-	P	U+182B
ᡊ	ᡊ	ᡊ	-	F	U+1839
ᡉ (ᡍ)	ᡉ (ᡍ)	ᡉ (ᡍ)	-	K	U+183A
ᡋ	ᡋ	ᡋ	-	M	U+182E
ᡌ	ᡌ	ᡌ	-	L	U+182F
ᡎ	ᡎ	ᡎ	-	S	U+1830
ᡏ	ᡏ	ᡏ	-	Š	U+1831
ᡐ	ᡐ	ᡐ	-	Č	U+1834
ᡑ	ᡑ	ᡑ	-	TS	U+183C
ᡒ	ᡒ	ᡒ	-	DZ	U+183D
ᡔ	ᡔ	ᡔ	-	R	U+1837
ᡕᡕ	-	-	-	CH	U+1842
ᡖ	-	-	-	RH	U+183F

Ideal Phonetic Model

Liang Hai; Draft Part 1: Required Variants.

U+1820 A

U+1821 E

U+1822 I

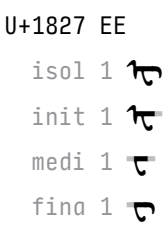
U+1823 O

U+1824 U

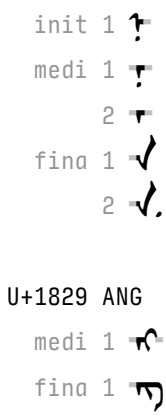
U+1825 OE

U+1826 UE

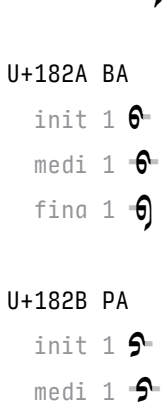
U+1827 EE



U+1828 NA

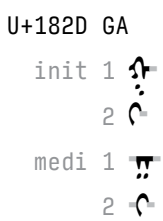


U+1829 ANG



U+182A BA

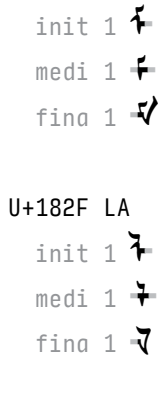
U+182B PA



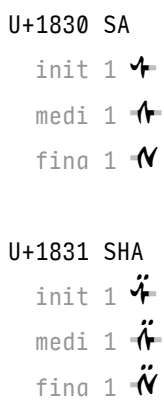
U+182C QA

U+182D GA

U+182E MA

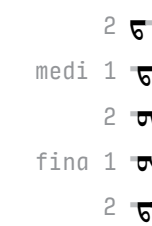


U+182F LA



U+1830 SA

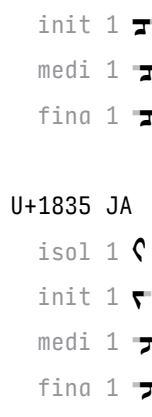
U+1831 SHA



U+1832 TA

U+1833 DA

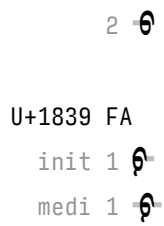
U+1834 CHA



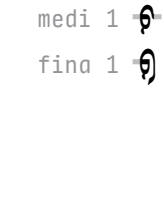
U+1835 JA

U+1836 YA

U+1837 RA



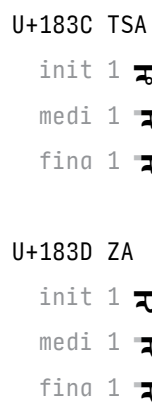
U+1838 WA



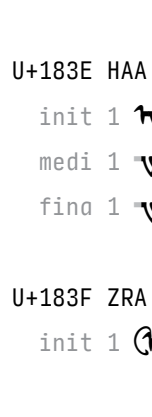
U+1839 FA

U+183A KA

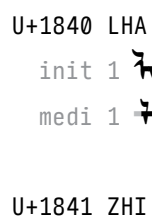
U+183B KHA



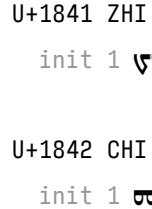
U+183C TSA



U+183D ZA



U+183E HAA



U+183F ZRA

U+1840 LHA

U+1841 ZHI

U+1842 CHI

Examples of Different Displays between Legacy Encoding vs. Minimal Graphetic Model

		Legacy Encoding	**Minimal Graphetic**
(Inner)	öbür		
(Mongolia)	monggol		
	-un		
(Autonomous)	öbertegen		
	jasaxu		
(Region)	orun		
(Hohhot)	xöxexota