WG2 #65 Mongolian Discussion Points
San Jose, CA
Sept. 26-30 2016
Greyson Translation Services
Greg Eck (by telephone) / Orlog Ou Rileke (in person)
greyson@postone.net

1.) UTC Request: NNBSP Deficiency – Proposal for new mechanism **[1 new code-point – U+180F]**
2.) UTC Request: U+1887 Deficiency – Change of Variation Selector Assignment **[U+FE00]**
3.) UTC Request: Unicode CDR 9.0 Corrections
4.) FYI: Nomenclature – Proposal for a Standard Canonical Nomenclature
5.) FYI: MVS+A/E+Suffix – Problem & Solution
6.) FYI: Difficulty in Displaying of Mongolian Web Pages – Bookmarklet Work-Around
7.) FYI: Microsoft Word 2016 – Word-Count & Word-Jump
8.) FYI: DS01 Objectives
9.) FYI: Baiti Font enhancements – to be shown in person by Mr. Ou
   - MVS+A/E+Suffix Solution – Proof of Concept
   - Baluda / Third Baluda – Implementation in Baiti (still working on this)
10.) Discussion: DS01 Violet Points as time allows
11.) Discussion: FVS Mismatch if time allows

Call for New Format Character to Replace NNBSP Usage in Mongolian Block

Even with the changes that were effected in the NNBSP re-specification (Unicode version 9.0), there are still deficiencies in proper functionality. These deficiencies include the following:

- Display problem if Fallback Font includes its own version of the NNBSP (U+202F). This problem occurs at the Application level.
- Certain applications such as PDF viewers will drop the NNBSP upon cut & paste as well as other operations.

As much as we fix the NNBSP in Mongolian to display correctly in different situations, we continue to find more areas where it will fail. The most recent finding over the summer was that given an application that is displaying Mongolian text AND given a font selection that is not a proper vertical Mongolian font AND given NNBSP+Suffix forms in the text AND given a fall-back font that uses its own NNBSP, the display of the suffixal forms will fail. The reason is that the fall-back font definition of the NNBSP does not include the OT rulings to do proper and necessary substitutions based on the NNBSP context.

There is considerable frustration in the Mongolian user community with the ongoing problems we are having with this one control character. At the same time, we see much momentum in the Mongolian community to adopt the Unicode encoding if it will work correctly. This is significant as there are huge swaths of Mongolian text using ASCII legacy era vertical Mongolian fonts. We cannot miss this moment of opportunity to provide a simple solution to this one biggest trouble of our Mongolian encoding – the NNBSP. It is felt that once an NNBSP-replacement is encoded within the Mongolian block, both rendering engines as well as fonts will be able to exercise adequate control as to display suffixal forms correctly.

I am calling for a new control character "Mongolian Suffix Connector" placed at U+180F to replace the current functionality of the NNBSP (U+202F) in Mongolian contexts. Transitional fonts, of course, would have to implement both the NNBSP as well as the replacement control character for backwards compatibility. Keyboards could be changed to use the new control character instead of the old NNBSP mechanism. Even if it took a considerable period of time to cast the encoding, I think all parties dealing with Mongolian vertical fonts will embrace this move.

The character should probably be a format character of General Category Cf rather than a space character. It should not decompose as the NNBSP does. It should not have any breaking features within the word or within a sentence or within a line of text. It should however exhibit 1/3 of a regular U+0020 space from the preceding character to the following character. In most contexts, the preceding character and the following character will change their shape.

Usage of a New Variation Selector VS01 (U+FE00)

Discussions with Professor Quejingzhabu have shown that the specification of the U+1887 Initial first variant is incorrect. The glyph is actually a final – the fifth of 5 final forms. Early in the specification period, a decision was made to incorrectly specify it as an initial in a work-around to include the form since there was no fourth variant selector. This should be corrected. The easiest solution is to use a generic Variation Selector. We are suggesting the use of VS01 – U+FE00. As this is actually a variant assignment, we are looking for consensus at the WG2 in the assignment of this variation selector to U+1887 Fifth Final Form. The specification would then look like the following:

ZWJ+1887	⊓ (x—x)	*First Final Form*
ZWJ+1887+FVS1	′ (x__)	*Second Final Form*
ZWJ+1887+FVS2	] (xx__)	*Third Final Form*
ZWJ+1887+FVS3	⊓ (xx_x)	*Fourth Final Form*
ZWJ+1887+VS01	⊃ (x__)	*Fifth Final Form*

STATUS: Proposal #2

Unicode CDR 9.0

There is one area that is considered to be incorrect in the Unicode charts for the Mongolian block. We can argue from the Chinese Standard as the basis for the incorrectness.

U+182D – The Unicode charts from the beginning have placed the final feminine form at 180B (FVS1). The MGWBM, the TR170, the DS01, the Chinese Standard all place the final feminine form at 180C as below.

ZWJ+182D	ᠭ (x__) *First Final Form* (NOTE: error in Chinese Standard - missing)
ZWJ+182D+FVS1	ᠭ (x__) *Second Final Form* (needed to over-ride default context) (SIG+FVS1)
ZWJ+182D+FVS2	ᠭ (x__) *Third Final Form* (context-driven) (ABSIG+FVS2)

SIG (which is by nature feminine) needs to be both masculine and feminine therefore we need to be able to over-ride the OT contextual substitutions. In the same way, ABSIG (which is by nature masculine) also needs to be able to over-ride the OT contextual substitutions.

<h1 style="text-align:center"><u>Nomenclature Standardization</u></h1>

There is a great need to standardize our notation into some canonical form in discussing the Mongolian encoding. I suggest the following standards using the following order:

Specification of a code-point form, positional form or variant form

Index 1 – Unicode code-point

Index 2 – Position

- Isolate
- Initial
- Medial
- Final

Index 3 – Variant

- Default
- First variant (applies FVS1)
- Second variant (applies FVS2)
- Third variant (applies FVS3)
- Fourth variant (applies FVS4)

For example, if we are talking about the U+1821 with all of its positional forms and variant forms it would look like this …

1821	ꛫ	UNICODE MONGOLIAN LETTER E
1821	ꛫ	*Isolate Form*
1821+ZWJ	ꜜ	*First Initial Form*
1821+FVS1+ZWJ	ꜞ	*Second Initial Form*
ZWJ+1821+ZWJ	ꜝ	*Medial Form*
ZWJ+1821	ꜟ	*First Final Form*
ZWJ+1821+FVS1	꜠	*Second Final Form*

There is no variant at the initial, so there is no need to specify first or second. There is a variant at the initial position, so the first and second are called out in the specification. Again, there is no variant at the medial position, so the medial position is called simply the "medial position". Etc., etc.

Gender should not be specified at this level as it will only clutter the definition. Unambiguous and discrete definition can be made using just the three indices given. Further detail including gender, the particular script being applied to, etc. can be made in the footnotes.

Specification of a ligature can proceed in the following order:

- Initial + Medial
- Initial + Final

- Medial + Medial
- Medial + Final

One could argue for Initial + Medial followed by Medial + Medial, etc. It is not so important which order is adopted. We should all just try to adopt one order for ease of comparison and usage.

Standardization leading towards a canonical nomenclature will help us understand each other's work quicker as well as encourage us to use each other's documentation and diagnostics. This in turn will reduce duplication of effort and move the Mongolian Font Development forward more quickly in turn.

<h1 style="text-align:center">Stem + MVS + Suffix – Problem & Solution</h1>

There is an advanced shaping case in regard to the MVS (U+180E) which is not being handled correctly by any of the existent Mongolian vertical fonts yet. This is the case where you have a STEM + MVS + A/E + SUFFIX. Proper shaping behavior in this case is like this … Given the STEM + MVS + A/E string, there is no problem. The word will form in most fonts correctly showing the left-ward sweeping tsatslag. By definition of the Mongolian script, no letter attaches to the tsatslag in final position. If the tsatslag is followed directly by an "attaching suffix", then the tsatslag A/E will transform back to its common A/E "shuud/single-tooth" medial form with no space in the shaping of the word. Final forms preceding the the tsatslag will revert to their default medial forms. If the suffix following the tsatslag A/E is removed, then the space preceding the tsatslag A/E will reappear as well as the tsatlag; final forms preceding the tsatslag will also reappear.
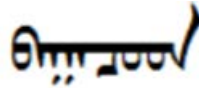
Given the following word and suffix:

BAG-MVS-A – noun meaning "small"   CHUD – plural suffix
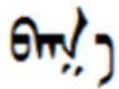
Proper shaping of BAG-MVS-A is

Proper shaping of BAG-MVS-A + CHUD(plural suffix) is
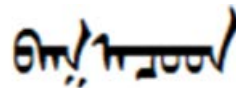
Current shaping of BAG-MVS-A + CHUD – plural suffix is   which is incorrect (current BAITI font rendering).

The solution that we have found for Baiti is like this …

Let the current OT rulings for STEM + MVS + A/E stand as they are. It is important to understand that <u>this will be considered the default shaping behavior.</u>

Modified behavior to handle a suffix after the Tsatslag_A/E can be effected by the following two rules …
1.) Change the Tsatslag A/E to the default medial A/E
- uni1820.init -> uni1820.medi in the context of MVS + uni1820.init + any_glyph_starting_a_suffix_which_can_be_appended_to_the_tsatslag_A
- uni1821.init -> uni1821.medi in the context of MVS + uni1821.init + any_glyph_starting_a_suffix_which_can_be_appended_to_the_tsatslag_E

NOTE: The lone tsatslag is tagged as an isolate by the rendering engines; however as soon as another letter is appended to the tsatslag, it is tagged as an initial and is no longer tsatslag in form but initial.

2.) Change the final forms of U+1828 / U+182E / U+182F / U+182C / U+182D / U+1835 / U+1836 / U+1837 / U+1838 preceding the MVS to the default medial forms

- uni18xx.fina -> uni18xx.medi in the context of   uni18xx.fina + MVS + uni1820.medi + any_glyph_starting_a_suffix_which_can_be_appended_to_the_tsatslag_A
- uni18xx.fina -> uni18xx.medi in the context of   uni18xx.fina + MVS + uni1821.medi + any_glyph_starting_a_suffix_which_can_be_appended_to_the_tsatslag_E

NOTE: The previously_isolate_tsatslag_then_turned_initial_A/E has by this time been changed to a medial by rule#1 above.

NOTE: The MVS will be happily and invisibly sitting in the middle of the string, ready to do its tsatslag-marking life job again if all suffixing following the A/E is removed.

If the post-MVS_A/E suffix string is removed, then the default shaping behavior mentioned above should return.

Questions that this has raised include:

- What do you say on the spelling of a dictionary entry like SARATAI? It clearly contains an MVS-A in the middle. Do we burden the user with the responsibility of thinking through each word and whether it has a tsatslag hidden in the stem?
- There will be other issues that this raises as well and should be discussed more in depth.

Work-Around to Anomalous Web-Site Display of Traditional Mongolian

There have been several in the Mongolian user community who have complained of not being able to view good Mongolian on their pages. After looking at the pages, we found that they were indeed written in Unicode-compliant code-point. The deficiency was found to be at the font as called out by the particular browser. Depending on whether the page includes CSS specifications that call for an actual Mongolian font (rather than just a fall-back font), the page will either display well or display improperly. One of the main problems found is in the display of the NNBSP and associated suffixes.

At a meeting on the Microsoft campus in Redmond this summer, the matter was discussed. A suggestion came up as to using a bookmarklet to control the actual font used to display the page. Microsoft employee, Aaron Bell, designed a bookmarklet, as attached that can be added to your favorites bar. Once executed, the display font is changed to Mongolian Baiti. Of course, the font could be Noto Sans Mongolian, MongolianWhite, or any other as desired. It is a work-around no doubt and still deficient at that. Only one font can be selected for a given page. But it is much better than having no control of the font on the given page. The bookmarklet can be found at the Greyson page (http://greyson.postone.net).

Microsoft Word 2016 – Word Count Feature Fixed

For some time, the word-count feature in Microsoft Word has been broken in the area of the NNBSP. Although the Unicode initiative to redefine some of the word-break features of the NNBSP (U+202F) will fix some of this functionality across all platforms/applications, there was a fix applied late summer 2016 which did fix the word-count feature under Microsoft Word 2016.

Given the string ᠬᠣᠶᠠᠷ ᠴᠠᠭ᠎ᠠ ᠴᠤᠭᠯᠠᠭᠠᠨ ᠲᠤᠬᠠᠢ ᠲᠤᠷᠠᠰᠬᠠᠯᠲᠤ ᠰᠠᠢᠬᠠᠨ ᠤ ᠲᠡᠭᠦᠰᠭᠡᠯ ᠲᠦᠷᠦ

The word count is correct at 8 words. The second word includes an MVS before the tsatslag_A The seventh word includes an NNBSP before the genitive suffix UN. This is correct. Kudos to Zoey Fan and her Microsoft Word team on this fix.

Microsoft Word 2016 – Word Jump Fix in Process

The word-jump feature using <CTL>+<RIGHT/LEFT CLICK> is still in process as seen below ..
_____

Cursor starts at beginning of the sentence

_____

First <CTL><RIGHT_CLICK> - CORRECT – one word jump

_____

Second <CTL><RIGHT_CLICK> - CORRECT – two word jumps

_____

Third <CTL><RIGHT_CLICK> - CORRECT – three word jumps

_____

Fourth <CTL><RIGHT_CLICK> - CORRECT – four word jumps

_____

Fifth <CTL><RIGHT_CLICK> - CORRECT – five word jumps

_____

Sixth <CTL><RIGHT_CLICK> - WRONG – 5 ½ word jumps

_____

Seventh <CTL><RIGHT_CLICK> - WRONG – 5 ¾ word jumps

_____

Eighth <CTL><RIGHT_CLICK> - CORRECT – six word jumps

_____

ꓕ𐌙ꓔᏤ ꝯᴑᴎᏤ) ꝯᴑᐊᏒ ꝩᴑᴦᏒ)ᴑᴦꓛᴧᴦᏤᏋᴧᏤ Ꮥᴔᴔꓭ) ᴑᏤ ꝩᴑᴧᴧᏤ)ᴔᏒꓭ

Ninth <CTL><RIGHT_CLICK> - CORRECT – seven word jumps

_____

ꓕ𐌙ꓔᏤ ꝯᴑᴎᏤ) ꝯᴑᐊᏒ ꝩᴑᴦᏒ)ᴑᴦꓛᴧᴦᏤᏋᴧᏤ Ꮥᴔᴔꓭ) ᴑᏤ ꝩᴑᴧᴧᏤ)ᴔᏒꓭ

Tenth <CTL><RIGHT_CLICK> - CORRECT– eight word jumps

_____

The problem is that on the 6th and the 7th right-click, the NNBSP is being picked up as it still has work-break characteristics tied to it. The NNBSP is being picked up as a word. Proper jump behavior would be for steps 6/7 above to be skipped.

DS01 – for information only

TARGET AUDIENCE: Font developers

OBJECTIVE: Full specification of Mongolian/Todo/Sibe/Manchu/AliGali script base forms, positional forms, and variant forms. It should be noted that full specification of implementation logic is beyond the scope of this paper. It should be noted that the emphasis of this paper is stand-alone text rather than running text. However, all running text forms, aside from ligatures should be found here. This paper is not meant to be a replacement for the Chinese Standard nor the Ulaanbaatar Mongolian Standard. It is derived from these two standards. Suggestions such as in the yellow/violet highlighting are derived from external discussions such as the Unicode Forum.

The DS01 is in complete sync with the Unicode 9.0 charts. The DS01 is in complete sync (other than yellow/violet sections) with the current Chinese Standard. Comments showing where the DS01 is not in sync with these two documents are appreciated and should be sent to greyson@postone.net . The current DS01 can be found at http://greyson.postone.net