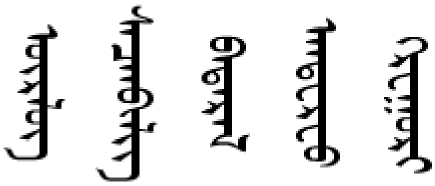
L2/17-036

PROPOSAL
Encode Mongolian Suffix Connector (U+180F) To Replace Narrow Non-Breaking Space (U+202F)

Greg Eck – greyson@postone.net
Andrew West - andrewcwest@gmail.com
Badral Sanlig - badral@bolorsoft.com
Siqinbilige - siqin@almas.co.jp
Ou Rileke - foximoyi@icloud.com



Relevant documents: L2/16-297 N4753 NNBSP Deficiency

This proposal requests the encoding of U+180F to be the new Mongolian Suffix Connector as a replacement for the currently-used NNBSP U+202F. The definition of this new character can follow that of the U+180E Mongolian Vowel Separator. All font developers are in agreement that the encoding of a replacement Mongolian suffix connector would be a good thing.

Background

There have been two problems that have plagued the drive to write in vertical Mongolian script since the beginning of the effort in the 1990s. The first has been the lack of clear specification and agreement on the encoding of positional and variant glyph forms. Also, the availability of such a specification has been an issue. Given that such a specification actually was found, it was many times under-specified. Further, given the incredible complexity of this 5-script encoding containing more than 1500 ligatures, errors easily crept in.

There have been several events that have arrested and are rectifying this situation however …

- Professor Quejingzhabu of Hohot, Inner Mongolia has generously and untiringly offered his professional time and counsel to many in the encoding process. He has offered his transform rules to font developers. His university department has held numerous conferences that have benefited companies, organizations and individual font developers alike.
- Font developers are talking together to share ideas, struggles, questions, and solutions. Richard Ishida's font comparator site (https://r12a.github.io/mongolian-variants/) has been a great help in comparing major vendor font positional glyphs, variant glyphs and other display choices. The online forum that Richard maintains helps all involved to stay in touch with the issues, provide discussion and find solutions to various development problems.
- Unicode members have from inception in the early 1990s until now presented a vision for a solid encoding greatly benefiting the Mongolian font user community. Major mile-stones include …
  - Unicode v2.0: Stable Mongolian code block from U+1800 – U+18A9
  - Unicode v6.2: Stable definition of the MVS format character allowing for a proper display of the A/E tsatslag final variants
  - Unicode v7.0: Variant forms and FVS sequences were added to code-charts
  - Unicode v9.0: Positional forms were added to the code charts
  - Unicode v10.0: Hopefully a joint effort by US/China/Mongolia will be submitted to include all variant and positional forms currently agreed upon

- Unicode v11.0: …

All of these events are leading to an ever-increasingly stable Mongolian encoding.

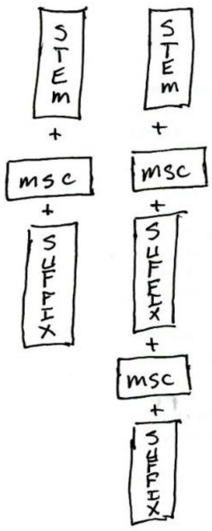
But there is still one problem area affecting up to 20% of Mongolian text that needs to be resolved. That is the issue of a stable suffix display across platforms and across the entire Mongolian font repertoire. There will be new code-points proposed and accepted, implemented. There will be new variant forms and associated variation selector sequences brought in. But there will be no single change that will stabilize the Mongolian encoding more than a solid solution to this problem. The Mongolian font user community need a robust connector character which will join suffixes to stems, suffixes to suffixes in a sure and consistent fashion across all platforms. If we can find a solution to this one issue, it will go a long way towards stabilizing the entire Mongolian script usage effort.

Problems

The current problems found with the use of the NNBSP (U+202F) as a suffix connector are at least two-fold. The most troubling is that many times the shaping of the suffix following the NNBSP is incorrect. The second is that the NNBSP-connected word often “breaks”. Both of these issues are discussed below in further detail.

1. Incorrect Shaping

In order to understand the issues, let’s first look at how a word is composed in vertical Mongolian. As an agglutinative language, every word that can be inflected starts with a stem. There are no prefixes, other than one very colloquial word-play structure. There are no infixes. There are only suffixes used to modify a stem. And a stem can take multiple suffixes of the form …



The unique aspect of connecting a suffix is that some suffixes require a bit of space in the connection. A special connector character is needed for these suffixes requiring space. Those suffixes connecting without space connect directly with no display problems.

Given the case where space in required in the suffix connection and where the connector character is used, the definition of a word is very important. The space is actually a part of the word. Therefore, a word with one case suffix and one plural suffix, for example, will have 2 bits of space in the actual word. The space is not a full-width space, but instead about 1/3 of a regular U+0020 space. It cannot be a U+0020 space however, because it is actually a character just as much as the other letters. It must be a character which has space and which joins the preceding letter with the following letter. A string of letters with such a connector character internal to the string is defined as a word. Though this example looks like 3 words because of the spaces, it is actually defined to be one word in Mongolian albeit composed of one stem and two suffixes.

A simple example used in our testing is NOM+Suffix_Connector+UN. This is represented by the sequence <U+1828><U+1823><U+182E><U+202F><U+1824><U+1828> ᠨᠣᠮ ᠤᠨ. The meaning of NOM is “book”. UN is the genitive case suffix. Looking at *APPENDIX IV*, we can see that this common noun NOM can take all of the case suffixes listed. In some situations, two case suffixes can be attached back to back.

One problem that may result when the suffix is added, is that the first letter of the suffix often changes form. The font’s Open Type rulings must change the form accordingly. Though difficult in the early years of Mongolian font development, most of these issues have already been resolved.

Proper Display ᠨᠣᠮ ᠤᠨ (Mongolian Baiti – Shipped version)

Improper Display ᠨᠣᠮ ᠤᠨ (NNBSP is drawn using Times New Roman)

2. Incorrect breaking behavior

A second problem arises in that an NNBSP-connected word tends to "break" at this connector character juncture. It is not proper for the word to separate over line breaks, during searches and sorts, word jump, word count, etc. The connector character must never, never allow the fore and aft parts of the word to separate or be considered separate fore from the aft. The scope of this problem includes text processing applications, browsers, dialogue boxes (including the URL box), and utilities such as search and sort. *During the period of Unicode v9.0 development, the UTC considered changing the General Category of the U+202F from Zs to Cf.* The purpose of this change would have been to better hold the word together in the case described above. We see now that even if the change had gone through, the third issue as described below would still not be resolved. *(see APPENDIX III)*

3. Font Fall-Back

A third major problem with the NNBSP acting as the suffix connector deals with the font fallback system used by operating systems and various applications. Where a utility processing Mongolian text should be designed to specify the desired Mongolian font, a general purpose application such as a browser will process all sort of scripts and languages. If a given web page does not specify a specific Mongolian font to display the Mongolian text, then the system must choose a font to use for each Mongolian character. Such a fallback choice is commonly fine for non-suffixed Mongolian words. The problem comes when the system attempts to display a suffixed word where the suffix connector is the NNBSP.

Moving back to the example string given earlier NOM+NNBSP+UN – the first character U+1828(N) is paired most likely with the Mongolian Baiti font and will be drawn without a problem. The second and third letters, in the same way have no problem being displayed. However, when the NNBSP is processed, it will commonly be paired with a font higher up in the fallback list such as Arial or Times New Roman. Both of these fonts register and process the U+202F NNBSP. But we want the NNBSP to be paired with a Mongolian font – for example Mongolian Baiti. This situation gives the problems as described below. The space inherent to the NNBSP is drawn up fine. It is at the processing of the 1824 where we encounter a problem. The 1824 should change shape from the default to a variant at this point. The context as registered in the Open Type substitution rule will match on a <U+202F><U+1824> sequence. As the NNBSP has been processed through another font such as Arial or Times New Roman, the possibility of a match on the context of <U+202F> <U+1824> is apparently lost. The U+1824 as paired with a Mongolian font such as Mongolian Baiti, finds no contextual match with the OT rule and so draws up the U+1824 in its default form. The display of the word is spoiled at this point no matter what happens afterwards. This third situation is very disturbing as it covers many different areas. To solve these issues at the font fallback level seems untenable due to time and man-power limitations. *(see APPENDIX I/II)*

Proposed Solution

Up until this summer, I personally had been hopeful that the performance of the NNBSP as a Mongolian Suffix Connector could still be resolved. With this third major issue arising, I think we cannot risk any more time in trying to resolve the Mongolian Suffix Connector by continuing to work with the NNBSP. There are still more unknowns down the road as Mongolian utilities develop and operating systems change and applications evolve. We know that there is a firm solution in the encoding of a single-use character that has just one solitary function as a suffix connector. We have seen the U+180E MVS evolve through development issues to the point where it is now completely stable. The MVS is a format character with the General Category defined to be Cf and carries only one function – to act as a connecting format letter between the early part of the word and the separate but connected A/E (the Mongolian name for this form is "tsatslag"). We have built prototype fonts using the MVS as the suffix connector and the resultant suffix display was fine. The requirements for the tsatslag connector and the requirements of the suffix connector are identical. A sure solution for the Mongolian suffix is a format character that performs only one function for one language just as the MVS performs one function and performs it well.

Therefore, I am proposing the encoding of U+180F to be the new Mongolian Suffix Connector as a replacement for the currently-used NNBSP U+202F. The definition of this new character can follow that of the U+180E Mongolian Vowel Separator, with properties as below. I have spoken with all of our font developers on the matter. All are in agreement that the encoding of a replacement Mongolian suffix connector would be a good thing.

Properties: `180F;MONGOLIAN SUFFIX CONNECTOR;Cf;0;BN;;;;;N;;;;;`

APPENDIX I
U+202F RELEVANT FACTS

- The NNBSP is registered in all Mongolian fonts
- The NNBSP is registered in some non-Mongolian fonts (as seen listed below)
- Fallback font chain would be expected to place non-English fonts such as Mongolian lower in the list
- Fallback font chain places some fonts which register the NNBSP higher in the list than Mongolian fonts; examples as shown below ***(in yellow highlighting)*** include Arial, most SIL fonts, Courier New, Calibri
- Only a Mongolian font is going to carry the Open Type rules to handle Mongolian suffix substitutions
- NNBSP related suffix OTL substitutions are essential to the proper display of Mongolian running text

Therefore Mongolian text which includes NNBSP-sequenced suffixes will never display correctly given a non-Mongolian fallback font is selected which registers and processes the NNBSP.

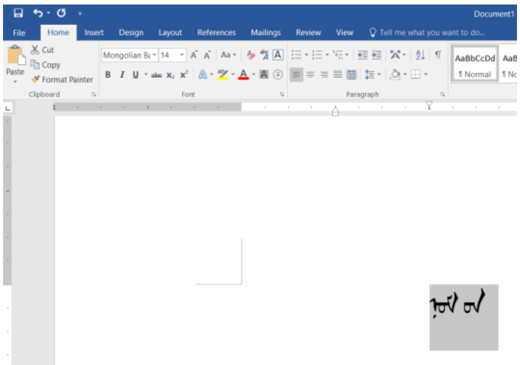
Legend
Yellow – carries the NNBSP; Mongolian suffix display is incorrect
Blue – carries the NNBSP, but the NNBSP seems to be non-addressable; Mongolian suffix display is fine
No_Color – does not carry the NNBSP; Mongolian suffix display is fine

This is a listing of the first ~50 fonts in the Windows 10 fonts directory:

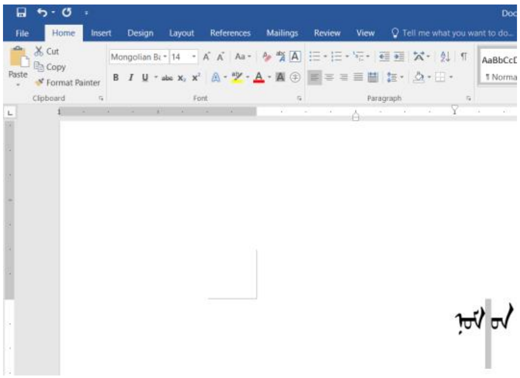
Agency FB (no NNBSP)
Algerian (no NNBSP)
Arial (has NNBSP)
Arial Black (no NNBSP)
Arial Narrow (no NNBSP)
Arial Rounded MT Bold (no NNBSP)
Arimo (has NNBSP – but not useable)
Baskerville Old Face (no NNBSP)
Bauhaus 93 (no NNBSP)
Bell MT (no NNBSP)
Berlin Sans FB (no NNBSP)
Berlin Sans FB Demi (no NNBSP)
Bernard MT Condensed (no NNBSP)
Blackadder ITC (no NNBSP)
Bodoni MT (no NNBSP)
Bodoni MT Condensed (no NNBSP)
Bodoni MT Poster Compressed (no NNBSP)
Book Antiqua (no NNBSP)
Bookman Old Style (no NNBSP)
Bradley Hand ITC (no NNBSP)
Brittanica Bold (no NNBSP)
Broadway (no NNBSP)
Brush Script MT (no NNBSP)
Calibri (has NNBSP)
Calibri Light (has NNBSP)
Californian FB (no NNBSP)
Calisto MT (no NNBSP)
Cambria (no NNBSP)
Cambria Math (no NNBSP)
Candara (no NNBSP)
Castellar (no NNBSP)
Centaur (no NNBSP)
Century (no NNBSP)
Century Gothic (no NNBSP)
Century Schoolbook (no NNBSP)
Charis SIL (has NNBSP)
Chiller (no NNBSP)
Collona MT (no NNBSP)
Comic Sans MS (no NNBSP)
Consolas (no NNBSP)
Constantia (no NNBSP)
Cooper Black (no NNBSP)
Copperplate Gothic Bold (no NNBSP)
Copperplate Gothic Light (no NNBSP)
Corbel (no NNBSP)
Courier New (has NNBSP)
Curlz MT (no NNBSP)

APPENDIX II
Display of NNBSP as selected by various fonts

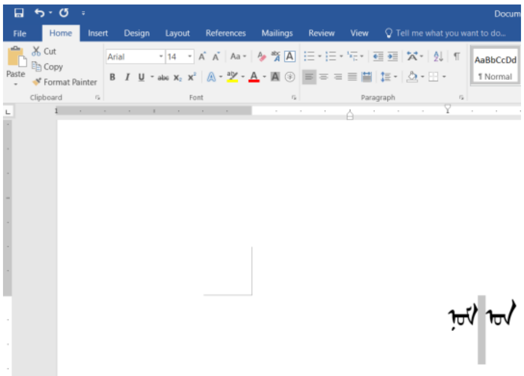
Proper Display by Shipped version of Mongolian Baiti



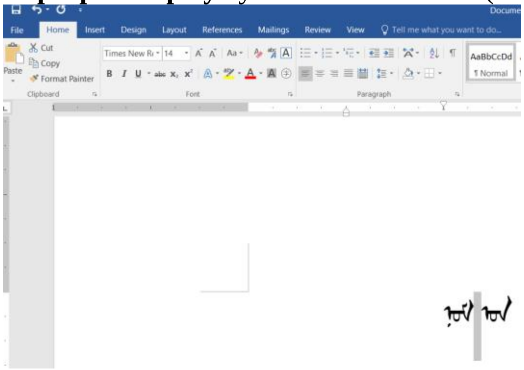
Proper Display by pre-Shipped version of Mongolian Baiti



Improper Display by Arial (Arial carries its own version of the NNBSP)

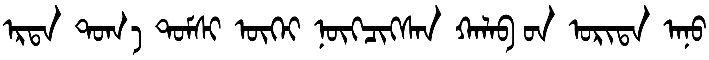


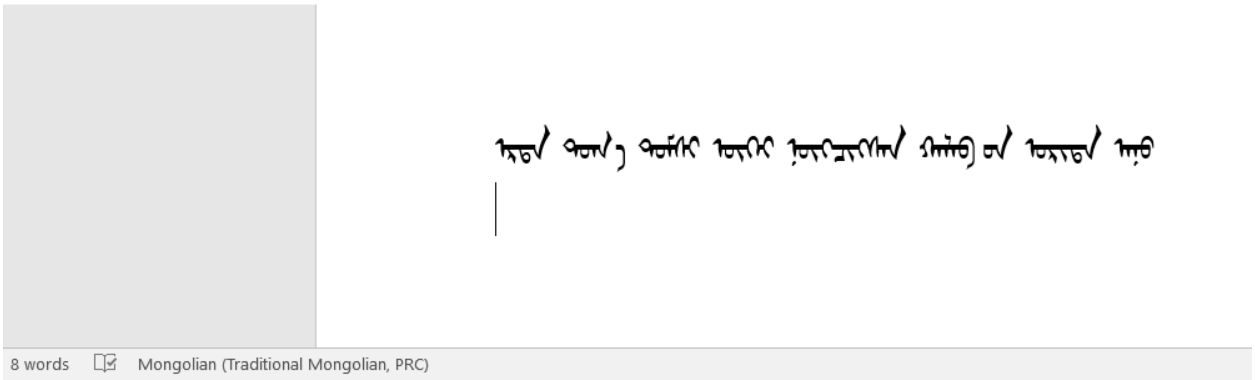
Improper Display by Times New Roman (Time New Roman carries its own version of the NNBSP)



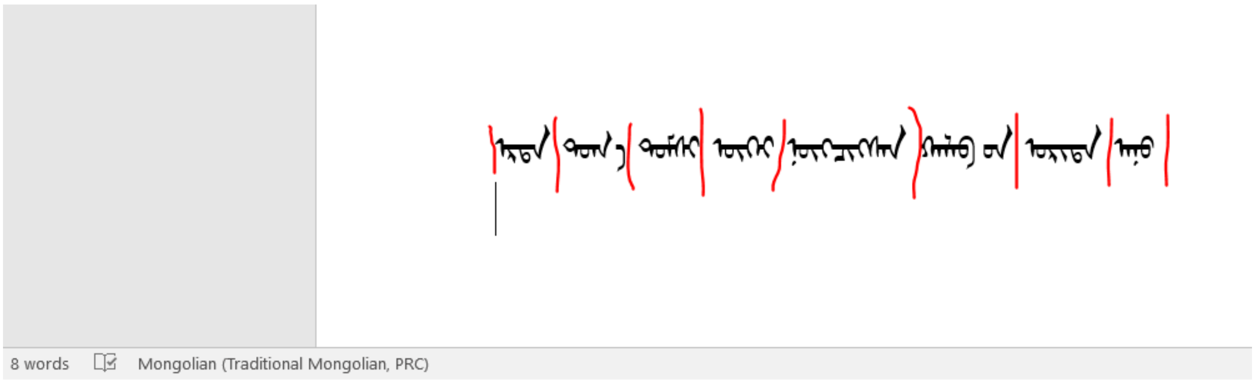
APPENDIX III
Microsoft Word 2016 – Word Count Feature Fixed

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 



8 words Mongolian (Traditional Mongolian, PRC)

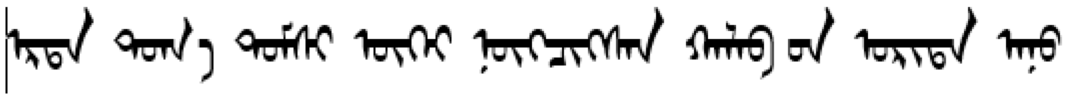


8 words Mongolian (Traditional Mongolian, PRC)

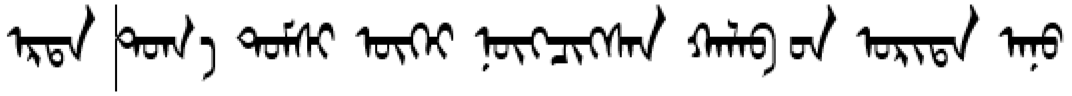
The word count is correct at 8 words. The second word includes an MVS before the tsatslag_A The seventh word includes an NNBSP before the genitive suffix UN. This is correct. Kudos to Zoey Fan and her Microsoft Word team on this fix.

Microsoft Word 2016 – Word Jump Fix in Process

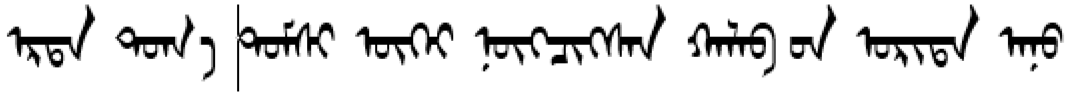
The word-jump feature using <CTL>+<RIGHT/LEFT CLICK> is still in process as seen below ..



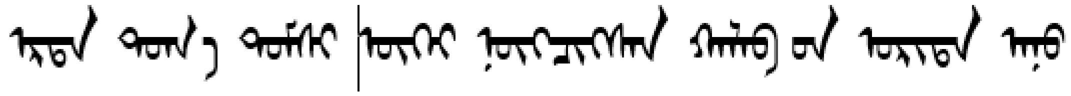
Cursor starts at beginning of the sentence



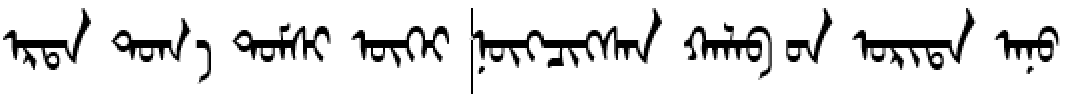
First <CTL><RIGHT_CLICK> - CORRECT – one word jump



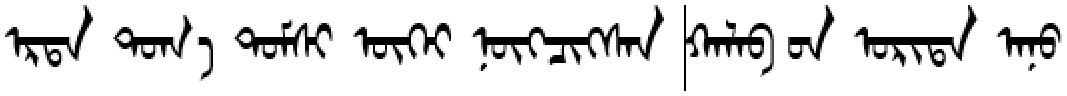
Second <CTL><RIGHT_CLICK> - CORRECT – two word jumps



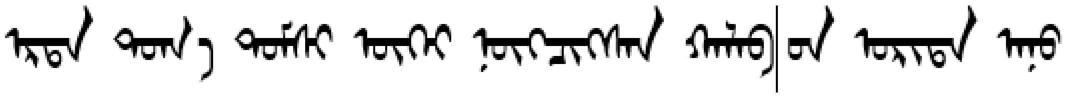
Third <CTL><RIGHT_CLICK> - CORRECT – three word jumps



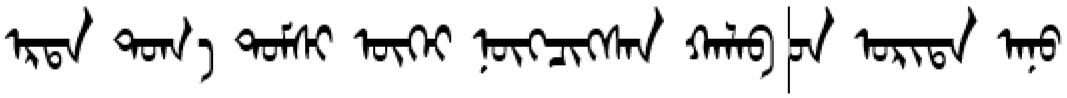
Fourth <CTL><RIGHT_CLICK> - CORRECT – four word jumps



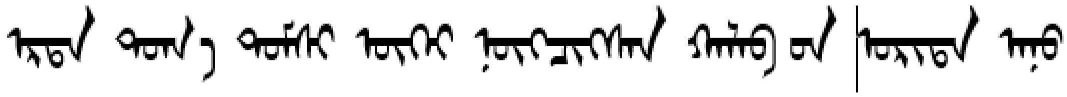
Fifth <CTL><RIGHT_CLICK> - CORRECT – five word jumps



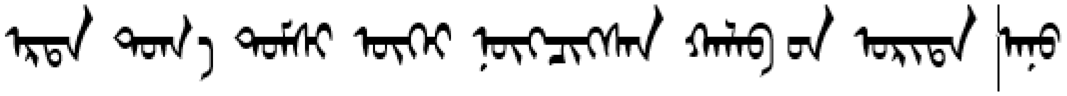
Sixth <CTL><RIGHT_CLICK> - WRONG – 5 ½ word jumps



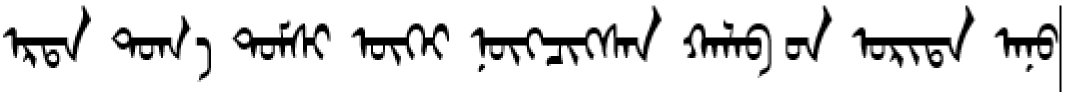
Seventh <CTL><RIGHT_CLICK> - WRONG – 5 ¾ word jumps



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



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



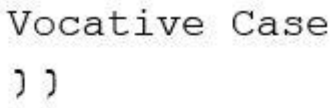
Tenth <CTL><RIGHT_CLICK> - CORRECT– eight word jumps

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.

APPENDIX IV
MONGOLIAN SUFFIXES
AS CONNECTED BY NNBSP

VOCATIVE CASE

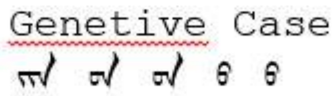
NNBSP+1820	ᠠ
NNBSP+1821	ᠡ


Vocative Case

ᠠ ᠡ

GENITIVE CASE

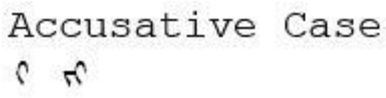
NNBSP+1836+1822+1828	ᠶᠢᠨ
NNBSP+1824+1828	ᠤᠨ
NNBSP+1826+1828	ᠦᠨ
NNBSP+1824	ᠤ
NNBSP+1826	ᠦ


Genetive Case

ᠶᠢᠨ ᠤᠨ ᠦᠨ ᠤ ᠦ

ACCUSATIVE CASE

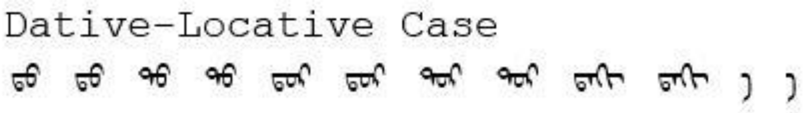
NNBSP+1822	ᠢ
NNBSP+1836+1822	ᠶᠢ


Accusative Case

ᠢ ᠶᠢ

DATIVE-LOCATIVE CASE

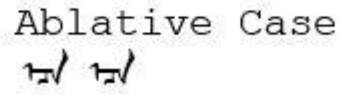
NNBSP+1833+1824	ᠳᠤ
NNBSP+1833+1826	ᠳᠦ
NNBSP+1832+1824	ᠲᠤ
NNBSP+1832+1826	ᠲᠦ
NNBSP+1833+1824+1837	ᠳᠤᠷ
NNBSP+1833+1826+1837	ᠳᠦᠷ


Dative-Locative Case

ᠳᠤ ᠳᠦ ᠲᠤ ᠲᠦ ᠳᠤᠷ ᠳᠦᠷ ᠲᠤᠷ ᠲᠦᠷ ᠳᠠᠬᠢ ᠳᠡᠬᠢ ᠠ ᠡ

NNBSP+1832+1824+1837	ᠲᠤᠷ
NNBSP+1832+1826+1837	ᠲᠦᠷ
NNBSP+1833+1820+182C+1822	ᠳᠠᠬᠢ
NNBSP+1833+1821+182C+1822	ᠳᠡᠬᠢ
NNBSP+1820	ᠠ
NNBSP+1821	ᠡ

ABLATIVE CASE

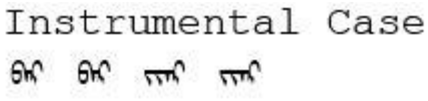


Ablative Case

ᠠᠴᠠ ᠡᠴᠡ

NNBSP+1820+1834+1820	ᠠᠴᠠ
NNBSP+1821+1834+1821	ᠡᠴᠡ

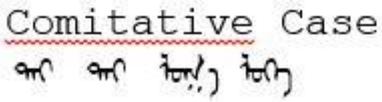
INSTRUMENTAL CASE



Instrumental Case

ᠪᠠᠷ ᠪᠡᠷ ᠢᠶᠠᠷ ᠢᠶᠡᠷ

NNBSP+182A+1820+1837	ᠪᠠᠷ
NNBSP+182A+1821+1837	ᠪᠡᠷ
NNBSP+1822+1836+1820+1837	ᠢᠶᠠᠷ
NNBSP+1822+1836+1821+1837	ᠢᠶᠡᠷ

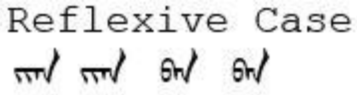


Comitative Case

ᠲᠠᠢ ᠲᠡᠢ ᠮᠤᠭ᠎ᠠ ᠮᠦᠭᠡ

COMITATIVE CASE

NNBSP+1832+1820+1822	ᠲᠠᠢ
NNBSP+1832+1821+1822	ᠲᠡᠢ
NNBSP+182F+1824+182D+180E+1820	ᠮᠤᠭ᠎ᠠ
NNBSP+182F+1826+182D+1821	ᠮᠦᠭᠡ



Reflexive Case

ᠢᠶᠠᠨ ᠢᠶᠡᠨ ᠪᠠᠨ ᠪᠡᠨ

REFLEXIVE CASE

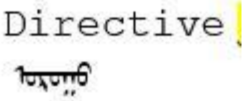
NNBSP+1822+1836+1820+1828	ᠢᠶᠠᠨ

NNBSP+1822+1836+1821+1828 ᠢᠶᠡᠨ

NNBSP+182A+1820+1828 ᠪᠠᠨ

NNBSP+182A+1821+1828 ᠪᠡᠨ

DIRECTIVE CASE (may or may not use NNBSP)

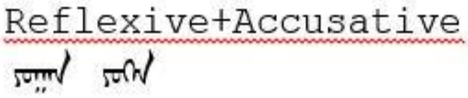


Directive

ᠤᠷᠤᠭᠤ

NNBSP+1824+1837+1824+182D+1824 ᠤᠷᠤᠭᠤ

REFLEXIVE + ACCUSATIVE CASE



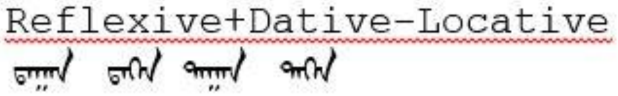
Reflexive+Accusative

ᠶᠤᠭᠠᠨ ᠶᠦᠭᠡᠨ

NNBSP+1836+1824+182D+1820+1828 ᠶᠤᠭᠠᠨ

NNBSP+1836+1826+182D+1821+1828 ᠶᠦᠭᠡᠨ

REFLEXIVE+DATIVE-LOCATIVE CASE



Reflexive+Dative-Locative

ᠳᠠᠭᠠᠨ ᠳᠡᠭᠡᠨ ᠲᠠᠭᠠᠨ ᠲᠡᠭᠡᠨ

NNBSP+1833+1820+182D+1820+1828

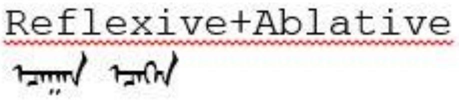
ᠳᠠᠭᠠᠨ

NNBSP+1833+1821+182D+1821+1828 ᠳᠡᠭᠡᠨ

NNBSP+1832+1820+182D+1820+1828 ᠲᠠᠭᠠᠨ

NNBSP+1832+1821+182D+1821+1828 ᠲᠡᠭᠡᠨ

REFLEXIVE+ABLATIVE CASE



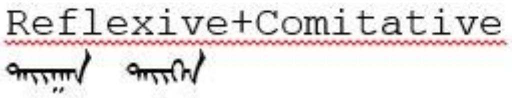
Reflexive+Ablative

ᠠᠴᠠᠭᠠᠨ ᠡᠴᠡᠭᠡᠨ

NNBSP+1820+1834+1820+182D+1820+1828 ᠠᠴᠠᠭᠠᠨ

NNBSP+1821+1834+1821+182D+1821+1828 ᠡᠴᠡᠭᠡᠨ

REFLEXIVE+COMITATIVE CASE



Reflexive+Comitative

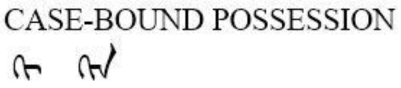
ᠲᠠᠶᠢᠭᠠᠨ ᠲᠡᠶᠢᠭᠡᠨ

NNBSP+1832+1820+1836+1822+182D+1820+1828 ᠲᠠᠶᠢᠭᠠᠨ

NNBSP+1832+1821+1836+1822+182D+1821+1828 ᠲᠡᠶᠢᠭᠡᠨ

CASE-BOUND POSSESSION CASE

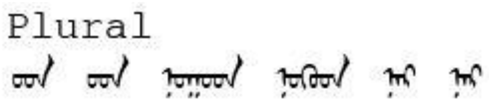
NNBSP+182C+1822	ᠬᠢ
NNBSP+182C+1822+1828	ᠬᠢᠨ


CASE-BOUND POSSESSION

ᠬᠢ ᠬᠢᠨ

PLURAL (first form may connect directly also)

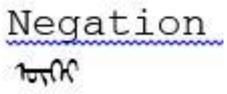
NNBSP+1824+1833	ᠤᠳ
NNBSP+1826+1833	ᠦᠳ
NNBSP+1828+1824+182D+1824+1833	ᠨᠤᠭᠤᠳ
NNBSP+1828+1826+182D+1826+1833	ᠨᠦᠭᠦᠳ
NNBSP+1828+1820+1837	ᠨᠠᠷ
NNBSP+1828+1821+1837	ᠨᠡᠷ


Plural

ᠤᠳ ᠦᠳ ᠨᠤᠭᠤᠳ ᠨᠦᠭᠦᠳ ᠨᠠᠷ ᠨᠡᠷ

NEGATION (may or may not use NNBSP)

NNBSP+1826+182D+1821+1822	ᠦᠭᠡᠢ

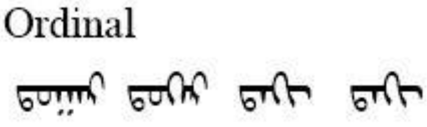

Negation

ᠦᠭᠡᠢ

ORDINAL

NNBSP+1833+1824+182D+1820+1837	ᠳᠤᠭᠠᠷ
NNBSP+1833+1826+182D+1821+1837	ᠳᠦᠭᠡᠷ
NNBSP+1833+1820+182C+1822	ᠳᠠᠬᠢ
NNBSP+1833+1821+182C+1822	ᠳᠡᠬᠢ

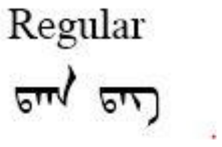
Note: may follow Mongolian or Latin digit as well as the spelled-out digit


Ordinal

ᠳᠤᠭᠠᠷ ᠳᠦᠭᠡᠷ ᠳᠠᠬᠢ ᠳᠡᠬᠢ

REGULAR ACTION (may or may not use NNBSP)

NNBSP+1833+1820+182D	ᠳᠠᠭ
NNBSP+1833+1821+182D	ᠳᠡᠭ


Regular

ᠳᠠᠭ ᠳᠡᠭ

PARTICLES NOT USING THE NNBSP

1824+1824	ᠤᠤ
1826+1826	ᠦᠦ