Solution for NNBSP Issues

Badral Sanlig
badral@bolorsoft.com
Bolorsoft LLC, Mongolia
Munkh-Uchral Enkhtur
munkhuchral@bolorsoft.com
Bolorsoft LLC, Mongolia

Abstract: The Mongol script[1] shows all of the essential characteristics of Mongolian language which is often classified as an agglutinative language. Taking some of the essential characteristics of Mongolian language that are embodied in the Mongol script, the character NNBSP (Narrow No-Break Space) was introduced in 1999 to join suffixes to basic stem words and also to join more suffixes to words ending with other suffixes. However, the NNBSP has never been problem-free until today. Therefore, our team aims to make an examination on the Unicode specification of NNBSP, and to analyze problems of this character and the existing solutions attempted to solve those problems. Also our team aims to demonstrate some features of Mongolian suffixes and propose an ideal solution for the NNBSP.

Keywords: Narrow No-Break Space, NNBSP, Mongolian space, Mongolian Suffix connector

1 Introduction

The Mongol script uses a gap to join some suffixes with stem words, or stem words end with other suffixes, which is a unique feature compared to many other writing systems. For this reason, the NNBSP was introduced in the Version 3.0 of the Unicode Standard in 1999. The WG2 group selected NNBSP as Mongolian suffix connector in the standard when Professor Quejingzhabu had proposed a new code for Mongolian suffix connector [MWG/2-N1]. The adapted functionality of this character is defined as follows in the TR170 [TR170, P. 10-11].

“The Mongolian space is not coded explicitly in the standards, but its functionality is provided by character 202F, NARROW NO-BREAK SPACE. The Mongolian space occurs frequently in Mongolian language: many words are formed by an addition of one or more suffixes (which indicate for example, different case endings of nouns and pronouns, ownership, and negation) to a basic stem word, and each individual suffix is separated from the stem or from the preceding suffix by the Mongolian space. Visually, this appears as a small white space, though it also affects the forms of the letters preceding and following it, the preceding character adopting its final form. However, it does not mark a

[1] We use the term ’Mongol script’ to indicate the traditional Mongolian script that is used in Mongolia and outside the country among other Mongolians. While the term ’Mongolian script’ implies to scripts that is used only in the country Mongolia.

break between words, the stem word together with all its suffixes is considered to form a single word.”

Several works published lists of NNBSP practices ([Que 2001], [TR170, P. 10-11], [GB/T 26226-2010] and [MWG/2-N1]). We didn't find any other specifications of NNBSP, even in "The Mongolian Encoding" book of Professor Quejingzhabu [Que 2001].

In the above specification, there is a lack of information about positional forms of first letters of suffixes that follow the NNBSP. In other words, it is not clear which positional form should be used as first letter of suffixes that follow the NNBSP. It is impractical and illogical to apply sometimes initial form of a letter, and sometimes medial form of a letter to start suffixes. We need to define one positional form, either initial or medial for the suffix beginning.

The NNBSP is defined as follows in current version of Unicode standard.

```
202F;NARROW NO-BREAK SPACE;Zs;0;CS;<noBreak> 0020;;;;N;;;;;
```

Property	Value	Description
Code point:	202F	
Name:	NARROW NO BREAK SPACE	English name
Version:	3.0 (August 1999)[2]	
Script:	Common[3]	
Category:	Zs[4]	Separator, Space
Combining:	0[3]	Not combined
BIDI:	CS (initially, it was WS)	Common Number Separator (COLON, COMMA, FULL STOP, NO-BREAK SPACE, ...)
Decomposition:	<noBreak> 0020	
Mirror:	N	N means No
Block:	General Punctuation	

Table 1: The definition of character 202F.

Use of NNBSP makes it possible to recognize suffixes from word structure, to select word, to determine word boundary or to parse. For example, «ᠠᠵᠢᠯ ᠤᠨ ᠬᠢᠨ ᠲᠠᠢ ᠪᠠᠨ» ajil-un-kin-tai-ban («ажлынхантайгаа» in Cyrillic, means “with colleagues” in English) is written in five separate parts, but it is still considered as one word. Unfortunately, the NNBSP does not completely work on the basis of the above mentioned functionality and specification. Thus, further investigation must be done.

[2] http://www.unicode.org/Public/UCD/latest/ucd/DerivedAge.txt

[3] UNIDATA, ftp://ftp.unicode.org/Public/UNIDATA/Scripts.txt

[4] UNIDATA, ftp://ftp.unicode.org/Public/UNIDATA/UnicodeData.txt

2 Problems and current solutions

In this section, we re-examine currently known NNBSP problems and consider previous attempts to fix it. Until now, there has been several attempts to fix the problem related to the NNBSP. For example, a change in the NNBSP properties, a proposal for character replacement etc.. We have discussed these issues with representatives of active user groups in Mongolia and classified the problems in four levels, as it is not clear to end users why and where these problems occurred.

- **Level 1**: Encoding
- **Level 2**: Font, Open type rules
- **Level 3**: Keymap or rendering engines
- **Level 4**: Applications

Until now, there has been several attempts to fix the problem related to NNBSP. For example, a change in NNBSP properties, a proposal for character replacement etc..

No.	Date	Subject	UTC No.
1.	**08 Jul 2015**	MONGOLIAN NNBSP-CONNECTED SUFFIXES	-
2.	**29 Jul 2015**	U+202F NNBSP Impact on Mongolian Options	L2/15-212
3.	**29 Sep 2015**	Proposal for property change from Zs to Pc.	
4.	**29 Jul 2015**	N4752 Mongolian Base Forms, Positional Forms, Variant Forms	L2/16-258
5.	**24 Sep 2016**	N4763 Comments on Mongolian, Small Khitan, and other WG2 #65 document	L2/16-266
6.	**26-30 Sep 2016**	N4753 WG2 #65 Mongolian Discussion Points	L2/16-259
7.	**29 Sep 2016**	N4753 NNBSP Deficiency	L2/16-297
8.	**15 Jan 2017**	Proposal to Encode Mongolian Suffix Connector (U+180F) To Replace NNBSP (U+202F)	L2/17-036
9.	**25 Jan 2017**	Comments on L2/17-036 (MSC)	L2/17-052
10.	**16 Mar 2018**	The Proposal for deprecation of MSC/NNBSP Mongolian Suffix Form Controlling Behavior	MWG/2-N1

Table 2: Previous proposals and comments.

In the following subsections, we will discuss existing problems and attempts to solve those problems.

2.1 Word boundary problem

Statement of the problem

There are two specificities in the Mongol script word boundary, in comparison to other writing systems. The first one is the "tsatsalga/orkitsa", which occurs separate from the consonants that end words, to indicate final forms of vowels "A" and "E". The second one is about writing some suffixes, that are connected with a small white space to stem words and stem words end with other suffixes. (See [TR170, P. 10-11]) To implement these logics in the computer environment, the MVS and NNBSP were introduced to Unicode standard. With the introduction of the MVS, the first issue regarding the "tsatsalga/orkitsa" has been successfully solved. But second specificity has not been solved completely.

Regardless of the space length between suffixes and stem, or even broken in multiple lines, in Mongol script a word that is connected with suffixes using NNBSP is considered as one word.

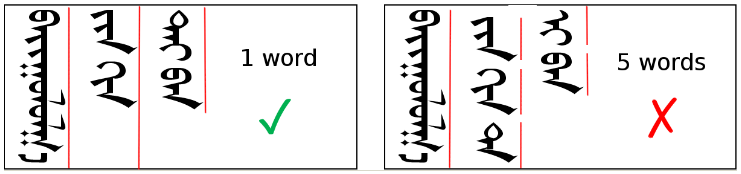


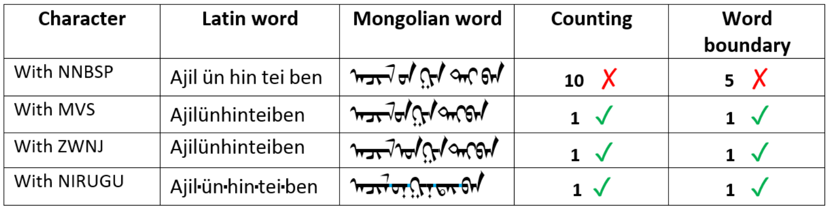
Figure 1: One long word in limited area

The aim, to introduce NNBSP to Unicode standard, was to recognize stem words followed with suffixes as one word. Unfortunately, it has never fully reached it's objective. It has certain characteristics for shaping the preceding and following characters, which mostly affects encoding experts. This side functionality will be discussed in section 2.2.

The word boundaries should be handled by higher level text processors, whereas the encoding should provide the basic feature to recognize a word. Thus, NNBSP has been introduced. It's visual representation is a small space and its length is defined as 1/3 em in [Que 2001]. This fixed length is always criticized by some experts as it is found frequently in practice with longer space. Also, it can be stretchable and breakable in limited print-area height due to agglutinative word structures and typography. (see figure 1).

To demonstrate word boundary problem of NNBSP, we will write the word ajil-un-kin-tai-ban intentionally wrong as ajil-ün-kin-tei-ben.

We use Microsoft Word as it is the best word processing software for Mongol script document creation. They update Mongol script features from version to version yet the current status of Version 2016 is unsatisfactory. The reason is, word counting functionality of MS-Word doesn't work correctly. However, it worked fine between version 2007 and 2013. The word boundary never worked correctly on all applications of every platform. (see figure 2) Every internal function to perform operations on individual words returns 5 separate words "ajil", "ün", "kin", "tei" and "ben". The "ajil" is a masculine word thus the last



Character	Latin word	Mongolian word	Counting	Word boundary
With NNBSP	Ajil ün hin tei ben	ᠠᠵᠢᠯ ᠦᠨ ᠬᠢᠨ ᠲᠡᠢ ᠪᠡᠨ	10 ✗	5 ✗
With MVS	Ajilünhinteiben	ᠠᠵᠢᠯ᠎ᠦᠨ᠎ᠬᠢᠨ᠎ᠲᠡᠢ᠎ᠪᠡᠨ	1 ✓	1 ✓
With ZWNJ	Ajilünhinteiben	ᠠᠵᠢᠯ‌ᠦᠨ‌ᠬᠢᠨ‌ᠲᠡᠢ‌ᠪᠡᠨ	1 ✓	1 ✓
With NIRUGU	Ajil·ün·hin·tei·ben	ᠠᠵᠢᠯ᠊ᠦᠨ᠊ᠬᠢᠨ᠊ᠲᠡᠢ᠊ᠪᠡᠨ	1 ✓	1 ✓

Figure 2: NNBSP behavior in MS Word 2016

suffix should be "ban" although "ban" and "ben" looks the same. How can we detect the difference between "ban" and "ben"? Also how can we check the correctness of the spelling, when someone mistypes the last suffix "ban" as "bal" and when this typing is valid for a word but not for a suffix? For Mongolian encoding model, spell checker plays the key role to solve issues of the visual ambiguity. In the above example, we have also checked the control characters MVS, NIRUGU and ZWNJ, and they work flawless. A classic usage of word boundary is the word select functions, which is used to show spelling suggestions by clicking the right mouse button.

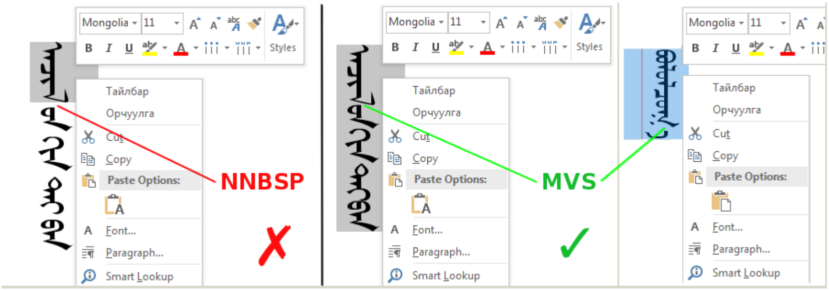


Figure 3: Selection problem for spell suggestion

Due to lack of instructions and incorrect category of NNBSP the word boundary functionality is incorrectly implemented, thus this problem has been classified at level 1 rather than level 4.

Unicode standard should explicitly instruct and specify how applications should handle NNBSP.

Reasons

1. The functionality and objective of NNBSP is not clearly specified. Thus, it is commonly handled same as SPACE character.
2. The category (Zs - Space, Separator) of NNBSP is incorrect for the functionality in Mongol script. Some applications directly replace NNBSP by SPACE.
3. The functionality and name of the character does not match. The inaccurate naming leads to misunderstanding and misconception that developers

handle this character as separator. That is the direct opposite of the actual goal. SPACE is generally understood as separator, but here for Mongolian Suffix Connector we discuss about JOINER control character, although it has space like presentation.

Current workaround

This problem is technical, thus end users took no notice until now. For us as spellchecker developers, this problem is a fatal error.

Solutions

We have proposed L2/17-036 [L2/17-036] to solve this issue but it still has not been approved.

Our preferred solution to this issue is to encode a new character named Mongolian Suffix Connector ASAP. We have carefully reviewed L2/17-036 and decided to reconsider line breaking properties. For more details see section 4.

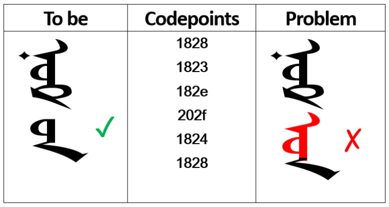
The consequences

NNBSP character will not be used for suffix joining.

2.2 Suffix initial corruption problem

Statement of the problem

The NNBSP does not accurately affect the forms of following suffixes. For example, "un/ün" (suffixes of genitive case) are displayed in initial form (with "titem" or crown, which indicates initial form of letters) instead of expected medial like form.



To be	Codepoints	Problem
✓	1828 1823 182e 202f 1824 1828	✗

Figure 4: Suffix with Titem problem

«ᠨᠣᠮ ᠤᠨ» nom-un ("номын" in Cyrillic, means "of a book" or "book's" in English). In this case, the suffix "un" (ᠤᠨ) is confused with a word "on" (ᠣᠨ) which means a year.

This problem can be called a fatal error and it is one of the technical problems in level 2 - 4.

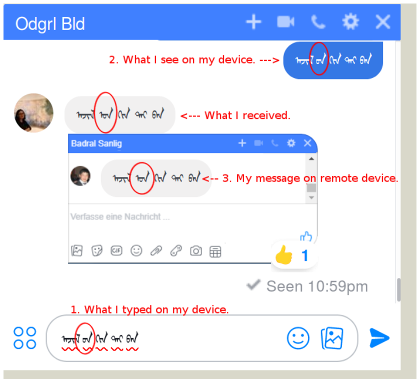
Reasons

1. This problem will occur if the font has no rule for suffixes or the rules implemented are incorrect.
2. Above reason is improbable, because till now every font has implemented the rules for all suffixes as contextual alternative. For example, the rule of suffix ᠤᠨ defined as:

```
sub [NNBSP] [u1824.init]' [u1828.fina] by [u1824.medi];
```

which means replace initial form of U by medial form of U, if it occurs after NNBSP and before final form of NA. The rule is correct. However, the problem still exists? Then the renderer could be at fault.

3. Reason two is also improbable, because we checked two major rendering engines Uniscribe and Harfbuzz, detected that they assign the properties of NNBSP correctly as defined in the Unicode standard. The joining type of NNBSP is defined as U, which means no joining group.[5] While NNBSP is classified in non-joining group, the preceding character takes final form and following character takes initial form. This initial form is substituted by medial form by open type engine as we defined in font rules. If this problem still occurs even when the font has correct lookup rules, and rendering engine works correctly then the culprit could only be the application. We didn't reproduce this problem with Microsoft office and OpenOffice using correct fonts. This problem often occurs in the Internet applications like Facebook messenger or Email applications. We have detected for instance



2. What I see on my device. --->

<--- What I received.

<-- 3. My message on remote device.

1. What I typed on my device.

Figure 5: Conversation in facebook messenger

Facebook messenger replaces NNBSP by SPACE. In this case, the font rules does not executed. (In early stage, some applications also has been replaced MVS with SPACE character.)

Current workaround

Mongolian user groups found their own workaround to avoid this difficulty. They replace NNBSP by a space and write a NIRUGU or back to display a medial form of the first letter of following suffixes. This attempt seems to be a solution in a way, but it causes more problems. Suffixes initiated by NIRUGU is aesthetically incorrect. (**Aesthetic problem**) Replacing NNBSP by SPACE character produces technically catastrophic fault that makes impossible to recognize correct word boundary for higher level processors.

Solutions

[5] ArabicShaping, ftp://ftp.unicode.org/Public/UNIDATA/ArabicShaping.txt

Unicode should specify explicitly for all vendors and developers to hold NNBSP and MVS and should not replace these characters by SPACE. Unicode produced them probably since 2000, but this problem exist still today and many developer are aware of this problem. Thus, we need to take appropriate measures now.

Our ideal solution is that the positional form of first letters of suffixes (or following letters of the NNBSP) must be in the initial form as defined in Arabic shaping model of standard and need to be encoded as initial variants, even if the form looks like medial without "Titem". Those medial forms encoded as initial variants to start some suffixes, then can be regulated by variation selector. Here, first of all positional mismatches [N4884] have to be fixed. We will submit a separate proposal regarding how to fix the positional mismatches. A similar idea was introduced already in [MWG/2-N1]. However, there are two main differences in our solution we are proposing here. First, the NNBSP must not deprecated because we need it to parse Mongolian words correctly. Second, the beginning letter of suffixes must take initial forms, not medial forms due to concern towards the characteristics of NNBSP properties.

The consequences

As a result of our solution, end users will not be obliged to use NNBSP as suffix connector. Because suffix will be shaped correctly in any case either with SPACE or with or without NNBSP. But for it we don't see any corrupted suffix behavior.

2.3 Line breaking problem

Statement of the problem

The problem is, there is a small space in the beginning of the next line when the text is written in certain frame. This line breaking problem is classified as an architectural problem in the level 1.

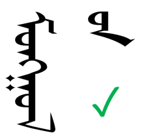
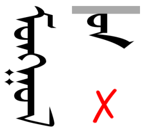
To be	Codepoints	Problem
	182e 1823 1829 182d 1823 182f 202f 1824 1828	

Figure 6: Line breaking problem

Reasons

This problem occurs in the case of writing in a limited space. It is a common behavior mostly in the publishing industry.

Current workaround

If users use SPACE + NIRUGU instead of NNBSP as described in the *Current workaround* of 2.2, then this problem will not occur. But all issues of 2.2 will remain.

Solutions

UTC advised the following to change line-breaking property of NNBSP character as follows. "NNBSP currently has line-break property GL[UE], indicating it prohibits line breaks before and after. However, Mongol script allows a line-break before NNBSP in limited contexts. In addition, UAX #14 and sections §6.2 and §13.5 of TUS state NNBSP is typically displayed with one third the width of a normal space character, but there is an evidence that NNBSP is stretchable (depending up on typographical style) and is not always less than a normal space. Also, it was noted that NNBSP loses its width (i.e., it disappears) over a line-break." [6]

In "COMMENTS ON L2/17-036" [7] the general controversies on the NNBSP issues have been described. In L2/17-052, we can find out that NNBSP is not suitable to be used as Mongolian Suffix Connector.

It is totally ridiculous that the Narrow **NO-BREAK** Space might be **BREAKABLE** and it's main functionality is **JOINING** Mongolian suffixes with the stem.

Our preferred solution is to introduce well specified Mongolian Suffix Connector with newly created line-break property.

The consequences

The Mongolian Suffix Connector requires a characteristic that allows a line break possibility before a character and the width of the character must be assigned to 0 over a line-break. To fulfill this requirement, NNBSP must provide a line break possibility but the name NARROW NO-BREAK SPACE forbid this behavior. Only changing the property yet not changing the name, will result more confusion as the name and the property clashes. Imagine, how tedious it would be to understand and to implement or use correctly the upper level functionality of NARROW NO-BREAK SPACE, which provides line breaking?

As experts mentioned, there exist a proper practice of NNBSP for French. Officially, NNBSP is used to forbid line-breaking of frequently used punctuation signs ";", ":", "!", "?". That means, changing the line-breaking property of NNBSP will lead to a defect of functionality in French.

2.4 Application support problem

Statement of the problem

In practice, there is no single application, which has properly implemented NNBSP. Microsoft Office has the best support but every version of Office suite has different issues and every suite products such as Word and Powerpoint even in the same version illustrate differently. All other remaining issues mentioned in other papers such as Font Fall-Back. (See [L2/17-036])

Reasons

Reasons for improper implementation of NNBSP are to do with misconception, inadequate specification and naming conflict.

[6] L2/18-168, https://www.unicode.org/L2/L2018/18168-script-rec.pdf

[7] L2/17-052, http://www.unicode.org/L2/L2017/17052-mongolian-cmt.pdf

Here, all attempts and suggestions to fix NNBSP are represented and evaluated in the table 3. The problems in last three columns are not explicitly noted above since they have not been encountered in NNBSP.

- **Suffix recognition** is closely related to word boundary problem. It is linguistically an unacceptable mistake, which means a word can not be identified as one word and all separately written parts of the word are recognized as independent words. For example, «ᠠᠵᠢᠯ ᠤᠨ ᠬᠢᠨ ᠲᠠᠢ ᠪᠠᠨ» (Ajil Un Kin Tai Ban) can not be counted and selected as one word but as five different words. In this case, the meaning of the word is completely altered. For example, "Ajil" means work; "Un" is a year; "Kin" is an another name of zither; "Tai" matches to one of three meanings: an encampment of horse relay, a garret and a stage; "Ban" means measure or a plank. Therefore, spell checking is impossible to work accurately. This serious error occurs in all attempts, which use the SPACE character instead of the NNBSP. In addition, removing the NNBSP/MSC means to ignore agglutinative characteristics of Mongolian language and the Mongol script.
- **Consistency** problems break Mongolian text and content. There are plenty of such problems in the Mongol script encoding model. If we use combined control characters for NNBSP/MSC then that will generate more problems. Limited number of reliable users will type two control characters before each suffix no matter how tedious can be. Many users will ignore the second character. The majority of people would ignore the first control character, because the suffix rules will be implemented in font with letters that start suffixes. Depending on contexts, even a same user can use different SPACE characters.
- **Usability** problems are caused by increased usage of invisible characters. Currently, there are enough invisible control characters like MVS, NNBSP/MSC, ZWJ, ZWNJ, FVS1, FVS2 and FVS3. The frequency of Mongolian Suffix Connection is very high and affect 20% of whole Mongolian texts. To apply an additional invisible character like ZWSP for such an excessive use is simply unacceptable to users. The repeatedly used invisible characters are unrecognizable by end users.

3 About Mongolian words and suffixes

A suffix is a letter or group letters that is attached to the end of a root or stem and on its own suffixes cannot express meaning of an individual word. Often, suffixes cause a spelling change to original forms of words. However, there have been some unacceptable comments given by some experts. Some experts consider that there is controversial understanding on words and suffixes. For example, a comment of an expert says "scholars' grammatical analysis considers a suffix is part of word" and "native users' common understanding tends to

No.	**Solution**	**Word boundary**	**Correct suffix form**	**Line breaking**	**Suffix recognition**	**Consistency**	**Usability**
1	NNBSP + Suff	-	-	-	+	+	+
2	SP + NIRUGU + Suff	-	-	+	-	-	+
3	SP + ZWJ + Suff	-	+	+	-	-	-
4	SP + Suff + FVS	-	+	+	-	-	+
5	NNBSP + Suff + FVS	-	+	-	+	-	+
6	NNBSP + ZWSP + Suff	-	+	+	+	-	-
7	**MSC (180F)**	+	+	+	+	+	+

Table 3: Review of previous attempts.

consider a suffix simply another word." [8] Also the same expert continues "it's not even often taught in school which words are suffixes or they shouldn't be separated from the stem word by a new line. Not everything said in grammar or orthography books is grammatical or orthographical requirements - could be just those author's typographical preferences."[9]

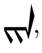
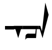
As we know, the Mongol script is a writing system for Mongolian language. Therefore, it follows the Mongolian language grammar. The above mentioned comment undermines this essentiality. The following reference is extracted from a fourth edition of a handbook of Mongolian language published by the Mongolian National University of Education, which has been the leading institute to train Mongol script teachers in Mongolia. This means all fundamental concepts written in this book are in textbooks and they are taught at schools. "Concepts of stem and affixes (prefixes, infix, derivational suffixes and inflectional suffixes or endings) are considered as morphemes. ... Morpheme is a bigger unit than a phoneme and a smaller unit than a word." Moreover the handbook, explains following differences between a morpheme and a word.

- Morpheme is a unit that creates words and modifies its meaning. It only occurs in the word structure. While word is indicates certain meaning.
- Morpheme can not be broken down into further elements. In case of break down, it will become meaningless syllables or phonemes on its own. In contrast, words can be broken down into morphemes.
- Morpheme has a lexical or grammatical meaning. While a word is an integration of both of these meanings. Compared to a morpheme, meaning

[8] https://github.com/unicode-org/uli-docs/issues/3
[9] http://www.unicode.org/L2/L2017/17052-mongolian-cmt.pdf

of a word is definite and concrete. Thus, a word can be independently used in a sentence. Meaning of a morpheme is abstract and general. Therefore, it can't independently stand on its own in a sentence.

- In some cases, some words and some morphemes can be visually the same. However, they can be essentially differentiated by their positions, meanings and roles. [MML, P. 102-103]

Following similarity in their forms, if separately written, then they are not suffixes but they are words, that are identical to some words, which indicate something and express certain meanings. Let's say in Mongol scrip suffixes such as , , ,  and not indicate any meaning. But some words have the same writing as some suffixes. For instance, "bar" can be a tiger when it is a word or a form of instrumental case when it is a suffix. This is the only case where suffixes can be confused with words with similar writing, while there is no relationship between the word and the suffix that are the same in its written form. Moreover, according the official training of Mongol script, if one understands suffixes as independent words then he or she is considered to be illiterate. Our main purpose should be about encoding of Mongol script, rather than encouraging every little bit that can lead to misunderstandings.

In the early period of the Mongol script development, separate suffixes were written together words they follow, which is common in Uighur script (See page 167 [Rachewiltz2010]). In the next period, suffixes began to be written separately. Many ancient monuments of the Mongol script reveal these changes. One can see both of these changes in the same monument. For example, küčündür (küčün-dür), qanu (qan-u) and tngri-yin, ulus-un (The Seal of Güyüg khan) [Tumurtogoo 2006, The Seal of Güyüg khan 1246].

A Mongolian linguist Agwaandandar uses a term "to write [suffixes] between words" (bičig-ün jabsar bičikü) to describe orthography for "bar", "iyar", "ta", "te", "nuγud", "nügüd", "luγ-a", "lüge", "daγan", "degen", "taγan", "tegen", "tur", "dur" in his Mongolian grammar work "Monggol usug-un yosun-i sayitur nomlagsan kelen-u cimeg kemegdeku orosibai" (1828). His suggestion considered suffixes not words, but something that should be written between words to indicate certain meanings.

In the Orthography chapter in his Mongolian grammar [Grammar of Written Mongolian, Orphograpy 88], N. Poppe writes that "Case endings are always written separately. If case ending begins with a vowel, it is a middle form. If a case suffix consists of only one vowel, then it is a final form. Most of plural suffixes consist of one or two syllables are also written separately." [MGR, P. 47, 98-120]

The Mongol script contains a total of 227 suffixes and 170 of them are joined directly to the stem but 59 are connected with the stem or the preceding suffix by attaching a space in between. We have researched why some suffixes start without "titem" (crown stroke) and why some are written with "titem"? Since the seventeenth century suffixes started adopting "titem". Almost all suffixes except 17 suffixes out of the 59 use "titem", which are highlighted yellow and gray in table 5. Five of the 17 suffixes are isolated versions and thus highlighted as gray. Those five are ML_A(1820), ML_E(1821), ML_U(2824), ML_UE(1826), ML_I(1822) and ML_YA(1836). All are vowels except YA. All suffixes, in spite of their forms they take (directly connected or not), are logically always connected to the stem word and together they builds certain meaning (form) of the stem word. The purpose of the NNBSP is to express the stem and all succession parts altogether as one word for counting and spell checking. It enables the implementation of spell checker, otherwise we cannot implement it. Many encoding experts agree that the spell checker will play the main role for the Mongol script due to the visual ambiguities. Thus, we have to consider NNBSP seriously in all layers. In L2/17-036, Andrew has suggested that we need to change the property of NNBSP to a category letter instead of a new control character. It is probably most straightforward way, if every software vendor follows it through and if it is handled exactly like a letter. In MWG/2-N1, Jirimutu proposed to use FVSs for the form of the first letter (one of the above six letters) of suffixes. It is a reliable solution for fixing only just to shape the first letters such as to , to etc. This will deprecate NNBSP because every suffix will be displayed correctly with or without NNBSP using other control characters such as variation selectors or ZWJ/ZWNJ. However, the shaping or visual illustration of suffixes is not the only problem, which will be explained in section 2.

4 Proposed solution

Here we want to conclude and propose solutions in section 2.

Any property of NNBSP should not be changed. If the line-break property of this character has to be breakable, it is better to introduce a new character, namely Mongolian Suffix Connector and make the property breakable. To encode MSC, we must first define a new line-breaking property, which will enable a possibility for MSC that precedes a character to loose its width (width=0) over a line-break. The line-breaking algorithm should also be implemented and should work properly. Then we define MSC as follows.

4.1 Proposed character

Code point	Proposed character name	Representative glyph
180F	Mongolian Suffix Connector	MSC

Table 4: The proposed character

4.2 UCD properties

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

Line break: `?? # Cf MONGOLIAN SUFFIX CONNECTOR`

?? property has to be defined and implemented in line-breaking algorithms. Joining type: `U`

Script: **Mongolian**

We have proposed L2/17-036 [L2/17-036] to solve all of the problems of the NNBSP, but has not been approved. We strongly encourage the acceptance of the proposal for MSC to replace NNBSP. However, some weakness of the proposal has to be fixed.

The replacement will definitely take some time. The more time passes without an immediate solution, the fewer people use Mongol script, which will then contribute to the elimination of this script.

Therefore, in the meantime, although MVS has a problem to break lines, we propose the solution to use MVS as NNBSP. There will be no modification, neither on NNBSP nor on MVS properties and it fulfills all requirements of Mongolian Suffix Connection except line-breaking defect. We have built prototype fonts using the MVS instead of NNBSP and it works perfect solving all above mentioned problems.

5 Acknowledgments

I express my gratitude to Odgerel Boldbaatar for her careful proofreading of the paper. I am also thankful to Erdenechimeg Myatav for providing detailed information on Technical Report 170 and valuable comments on the history of the Mongolian encoding model.

References

[TR170, P. 10-11] Erdenechimeg, M., Moore, R., Namsrai, Yu. *UNU/IIST Technical Report No. 170 - Traditional Mongolian Script in the ISO/Unicode Standards*, Macau, 1999

[Tumurtogoo 2006, The Seal of Güyüg khan 1246] *Mongolian Monuments in Uighur-Mongolian Script (XIII–XVI Centuries): Introduction, Transcription and Bibliography*, Tumurtogoo, D.; With the Collaboration of G. Cecegdari. Taipei, 2006. (Language and Linguistics Monograph Series A–11).

[MML, P. 102-103] Onorbayan Ts., Tsog-Ochir A., Ariunbold U. *Modern Mongolian Language*, Ulaanbaatar, 2012

[Rachewiltz2010] Igor de Rachewiltz, Volker Rabatzki *Introduction to Altaic Philology*, Leiden, Boston: Brill, 2010.

[Grammar of Written Mongolian, Orphograpy 88] N.N. Poppe. *Grammar of Written Mongolian*. 5th unrevised printing Far Eastern and Russian Institute publications on Asia. Harrassowitz, Wiesbaden, Germany, 2006. LCCN: 56036254

[MGR, P. 47, 98-120] Poppe N. Монгол бичгийн хэлний зүй, Ulaanbaatar, 2012

[MMM, P. 76] Munkh-Amgalan Yu., Shi Kan Орчин цагийн монгол хэлний бүтээвэр судлал, Ulaanbaatar, 2014

[Que 2001] Quejingzhabu, H *Meng Gu Wen Bian Ma - The Mongolian Encoding*, (in Chinese) China, 2001, ISBN: 7810740962

[GB/T 26226-2010] Chinese standard, *Information technology - Mongolian presentation forms character set and use rules of controlling characters*, China, 2011

[MWG/2-N1] Jirimutu, Siqin, Bao Haishan, Burigudu, Menghejiya *The Proposal for deprecation of MSC/NNBSP Mongolian Suffix Form Controlling Behavior*, https://www.unicode.org/~lisa/mongoliandocs/mwg2-1Mongolian-Jirimutu-Proposal.pdf

[L2/17-036] Greg Eck, Andrew West, Badral Sanlig, Siqinbilige, Ou Rileke *Encode Mongolian Suffix Connector (U+180F) To Replace Narrow Non-Breaking Space (U+202F)*, https://www.unicode.org/L2/L2017/17036-mongolian-suffix.pdf

[N4884] Shen Yilei *Positional Mismatches in Mongolian Encoding*, https://www.unicode.org/L2/L2017/17332-mong-pos-matches.pdf 2017

Appendices

Table of the NNBSP joining suffixes of Mongol script

No.	Mon.	Latin	Cyrillic	Form	Description
1.	ᠶᠢᠨ	-yin	-ын, -ийн, -н	medi	Genitive case, after vowel
2.	ᠤᠨ	-un	-ын	medi	Genitive case, after consonant, masculine word
3.	ᠦᠨ	-ün	-ийн	medi	Genitive case, after consonant, feminine word
4.	ᠤ	-u	-ы, -ны	isol	Genitive case, only after N, masculine word
5.	ᠦ	-ü	-ий, -ний	isol	Genitive case, only after N, feminine word
6.	ᠳᠤ	-du	-д	init	Dative-locative case, after vowel and syllable closing consonant N, M, L, NG, masculine word
7.	ᠳᠦ	-dü	-д	init	Dative-locative case, after vowel and syllable closing consonant N, M, L, NG, feminine word
8.	ᠲᠤ	-tu	-т	init	Dative-locative case, after syllable closing consonant B, GA, GE, R, S, D, masculine word
9.	ᠲᠦ	-tü	-т	init	Dative-locative case, after syllable closing consonant B, GA, GE, R, S, D, feminine word
10.	ᠳᠤᠷ	-dur	-д	init	Dative-locative case, after vowel and syllable closing consonant N, M, L, NG, masculine word

Continued on next page . . .

No.	Mon.	Latin	Cyrillic	Form	Description
11.	ᠳᠦᠷ	-dür	-д	init	Dative-locative case, after vowel and syllable closing consonant N, M, L, NG, feminine word
12.	ᠲᠤᠷ	-tur	-т	init	Dative-locative case, after syllable closing consonant B, GA, GE, R, S, D, masculine word
13.	ᠲᠦᠷ	-tür	-т	init	Dative-locative case, after syllable closing consonant B, GA, GE, R, S, D, feminine word
14.	ᠠ	-a	-а	isol	Locative case, masculine word
15.	ᠡ	-e	-э	isol	Locative case, feminine word
16.	ᠠᠴᠠ	-ača	-аас, -оос	init	Ablative case, masculine word
17.	ᠡᠴᠡ	-eče	-ээс, -өөс	init	Ablative case, feminine word
18.	ᠶᠢ	-yi	-ыг, -ийг, -г	medi	Accusative case after vowel & YA
19.	ᠢ	-i	-ыг, -ийг, -г	isol	Accusative case after consonants
20.	ᠢᠶᠠᠷ	-iyar	-аар, -оор	medi	Instrumental case, after consonant, masculine word
21.	ᠢᠶᠡᠷ	-iyer	-ээр, -өөр	medi	Instrumental case, after consonant, feminine word
22.	ᠪᠠᠷ	-bar	-аар, -оор	init	Instrumental case, after vowel, masculine word
23.	ᠪᠡᠷ	-ber	-ээр, -өөр	init	Instrumental case, after vowel, feminine word
24.	ᠲᠠᠢ	-tai	-тай, -той	init	Comitative case, masculine word
25.	ᠲᠡᠢ	-tei	-тэй	init	Comitative case, feminine word

Continued on next page . . .

No.	Mon.	Latin	Cyrillic	Form	Description
26.	ᠯᠤᠭ᠎ᠠ	-luγa	-лугаа	init	Comitative case, masculine word
27.	ᠯᠦᠭᠡ	-lüge	-лугээ	init	Comitative case, feminine word
28.	ᠤᠳ	-ud	-ууд	medi	Plural suffix, after consonant except N, masculine word
29.	ᠦᠳ	-üd	-үүд	medi$^{??}$	Plural suffix, after consonant except N, feminine word
30.	ᠨᠤᠭᠤᠳ	-nuγud	-нууд	init	Plural suffix, after vowel and N, masculine word
31.	ᠨᠦᠭᠦᠳ	-nügüd	-нүүд	init	Plural suffix, after vowel and N, after vowel and N, YA
32.	ᠨᠠᠷ	nar	нар	init	Plural suffix
33.	ᠢᠶᠠᠨ	-iyan	-аа, -оо	medi	Reflexive-possessive ending, after consonant, masculine word
34.	ᠢᠶᠡᠨ	-iyen	-ээ, -өө	medi	Reflexive-possessive ending, after consonant, feminine word
35.	ᠪᠠᠨ	-ban	-аа, -оо	init	Reflexive-possessive ending, after vowel, masculine word
36.	ᠪᠡᠨ	-ben	-ээ, -өө	init	Reflexive-possessive ending, after vowel, feminine word
37.	ᠶᠤᠭᠠᠨ	-yuγan	-ыгаа, -ыгоо	init	Accusative case + Reflexive-possessive ending,masculine word
38.	ᠶᠦᠭᠡᠨ	-yügen	-ийгээ, -ийгөө	init	Accusative case + Reflexive-possessive ending,feminine word
39.	ᠤᠪᠠᠨ	-uban	-ынхаа, -ынхоо	medi	Genitive case + Reflexive-possessive ending, after N, masculine word

Continued on next page . . .

No.	Mon.	Latin	Cyrillic	Form	Description
40.	ᠦᠪᠡᠨ	-üben	-ийнхээ, -ийнхөө	medi	Genitive case + Reflexive-possessive ending, after N, feminine word
41.	ᠳᠠᠭᠠᠨ	-daγan	-даа, доо	init	Dative-locative case + Reflexive-possessive ending, after vowel and syllable closing consonant N, M, L, NG, masculine word
42.	ᠳᠡᠭᠡᠨ	-degen	-дээ, дөө	init	Dative-locative case + Reflexive-possessive ending, after vowel and syllable closing consonant N, M, L, NG,feminine word
43.	ᠳᠤᠷᠢᠶᠠᠨ	duriyan	-даа, -доо	init	Dative-locative case + Reflexive-possessive ending, after vowel and syllable closing consonant N, M, L, NG, masculine word
44.	ᠳᠦᠷᠢᠶᠡᠨ	düriyen	-дээ, -дөө	init	Dative-locative case + Reflexive-possessive ending, after vowel and syllable closing consonant N, M, L, NG, feminine word
45.	ᠲᠠᠭᠠᠨ	-taγan	-таа, тоо	init	Dative-locative case + Reflexive-possessive ending, after syllable closing consonant B, GA, GE, R, S, D, masculine word
46.	ᠲᠡᠭᠡᠨ	-tegen	-тээ, төө	init	Dative-locative case + Reflexive-possessive ending, after syllable closing consonant B, GA, GE, R, S, D, feminine word
47.	ᠲᠠᠶᠢᠭᠠᠨ	-tayiγan	-тайгаа, -тойгоо	init	Comitative case + Reflexive-possessive ending, masculine word
48.	ᠲᠡᠶᠢᠭᠡᠨ	-teyigen	-тэйгээ, -тэйгөө	init	Comitative case + Reflexive-possessive ending, feminine word
49.	ᠠᠴᠠᠭᠠᠨ	-ačaγan	-аасаа, оосоо	init	Ablative case + Reflexive-possessive ending, masculine word

Continued on next page . . .

No.	Mon.	Latin	Cyrillic	Form	Description
50.	ᠡᠴᠡᠭᠡᠨ	-ečegen	-ээсээ, өөсөө	init	Ablative case + Reflexive-possessive ending, feminine word
51.	ᠦᠭᠡᠢ	ügüi	үгүй	init	Negation
52.	ᠤᠷᠤᠭᠤ	uruγu	-руу, -рүү, -луу, -лүү	init	Directive case
53.	-ᠮᠢᠨᠢ	mini	минь	init	Personal Possessive
54.	-ᠮᠠᠨᠢ	mani	маань	init	Personal Possessive
55.	-ᠴᠢᠨᠢ	čini	чинь	init	Personal Possessive
56.	-ᠲᠠᠨᠢ	tani	тань	init	Personal Possessive
57.	-ᠨᠢ	ni	нь	init	Personal Possessive
58.	-ᠠᠨᠤ	anu	ану	init	Personal Possessive
59.	-ᠢᠨᠦ	inü	инү	init	Personal Possessive

Table 5: NNBSP joining suffixes

ISO/IEC JTC 1/SC 2/WG 2
PROPOSAL SUMMARY FORM TO ACCOMPANY SUBMISSIONS
FOR ADDITIONS TO THE REPERTOIRE OF ISO/IEC 10646[1]
Please fill all the sections A, B and C below.
Please read Principles and Procedures Document (P & P) from http://std.dkuug.dk/JTC1/SC2/WG2/docs/principles.html for guidelines and details before filling this form.
Please ensure you are using the latest Form from http://std.dkuug.dk/JTC1/SC2/WG2/docs/summaryform.html.
See also http://std.dkuug.dk/JTC1/SC2/WG2/docs/roadmaps.html for latest *Roadmaps*.

A. Administrative

1. **Title:** Proposal for solution of NNBSP issues
2. Requester's name: Badral Sanlig, Munkh-Uchral Enkhtur
3. Requester type (Member body/Liaison/Individual contribution): Individual contribution
4. Submission date: 2018-09-10
5. Requester's reference (if applicable):
6. Choose one of the following:
   This is a complete proposal: Yes
   (or) More information will be provided later:

B. Technical – General

1. Choose one of the following:
   a. This proposal is for a new script (set of characters): No
      Proposed name of script:
   b. The proposal is for addition of character(s) to an existing block: Yes
      Name of the existing block: Mongolian
2. Number of characters in proposal: 1
3. Proposed category (select one from below - see section 2.2 of P&P document):
   A-Contemporary ___ B.1-Specialized (small collection) X B.2-Specialized (large collection) ___
   C-Major extinct ___ D-Attested extinct ___ E-Minor extinct ___
   F-Archaic Hieroglyphic or Ideographic ___ G-Obscure or questionable usage symbols ___
4. Is a repertoire including character names provided? Yes
   a. If YES, are the names in accordance with the “character naming guidelines” in Annex L of P&P document? Yes
   b. Are the character shapes attached in a legible form suitable for review? Yes
5. Fonts related:
   a. Who will provide the appropriate computerized font to the Project Editor of 10646 for publishing the standard?
   Badral Sanlig
   b. Identify the party granting a license for use of the font by the editors (include address, e-mail, ftp-site, etc.):
   MongolianScript, SIL licensed, open font
6. References:
   a. Are references (to other character sets, dictionaries, descriptive texts etc.) provided? Yes
   b. Are published examples of use (such as samples from newspapers, magazines, or other sources) of proposed characters attached? Yes
7. Special encoding issues:
   Does the proposal address other aspects of character data processing (if applicable) such as input, presentation, sorting, searching, indexing, transliteration etc. (if yes please enclose information)? Yes
   See above proposal

8. Additional Information:

Submitters are invited to provide any additional information about Properties of the proposed Character(s) or Script that will assist in correct understanding of and correct linguistic processing of the proposed character(s) or script. Examples of such properties are: Casing information, Numeric information, Currency information, Display behaviour information such as line breaks, widths etc., Combining behaviour, Spacing behaviour, Directional behaviour, Default Collation behaviour, relevance in Mark Up contexts, Compatibility equivalence and other Unicode normalization related information. See the Unicode standard at http://www.unicode.org for such information on other scripts. Also see Unicode Character Database (http://www.unicode.org/reports/tr44/) and associated Unicode Technical Reports for information needed for consideration by the Unicode Technical Committee for inclusion in the Unicode Standard.

[1] Form number: N4502-F (Original 1994-10-14; Revised 1995-01, 1995-04, 1996-04, 1996-08, 1999-03, 2001-05, 2001-09, 2003-11, 2005-01, 2005-09, 2005-10, 2007-03, 2008-05, 2009-11, 2011-03, 2012-01)

C. Technical - Justification

1. Has this proposal for addition of character(s) been submitted before? Yes
   If YES explain See: L2/17-36 http://www.unicode.org/L2/L2017/17036-mongolian-suffix.pdf
2. Has contact been made to members of the user community (for example: National Body, user groups of the script or characters, other experts, etc.)? Yes
   If YES, with whom? W3, MWG/2 group members
   If YES, available relevant documents:
3. Information on the user community for the proposed characters (for example: size, demographics, information technology use, or publishing use) is included? No
   Reference:
4. The context of use for the proposed characters (type of use; common or rare) Common
   Reference:
5. Are the proposed characters in current use by the user community? No
   If YES, where? Reference:
6. After giving due considerations to the principles in the P&P document must the proposed characters be entirely in the BMP? Yes
   If YES, is a rationale provided?
   If YES, reference:
7. Should the proposed characters be kept together in a contiguous range (rather than being scattered)? Yes
8. Can any of the proposed characters be considered a presentation form of an existing character or character sequence? No
   If YES, is a rationale for its inclusion provided?
   If YES, reference:
9. Can any of the proposed characters be encoded using a composed character sequence of either existing characters or other proposed characters? No
   If YES, is a rationale for its inclusion provided?
   If YES, reference:
10. Can any of the proposed character(s) be considered to be similar (in appearance or function) to, or could be confused with, an existing character? Yes
    If YES, is a rationale for its inclusion provided? Yes
    If YES, reference: See above proposal
11. Does the proposal include use of combining characters and/or use of composite sequences? No
    If YES, is a rationale for such use provided?
    If YES, reference:
    Is a list of composite sequences and their corresponding glyph images (graphic symbols) provided?
    If YES, reference:
12. Does the proposal contain characters with any special properties such as control function or similar semantics? Yes
    If YES, describe in detail (include attachment if necessary)
    See above proposal
13. Does the proposal contain any Ideographic compatibility characters? No
    If YES, are the equivalent corresponding unified ideographic characters identified?
    If YES, reference: