

## COMMENTS ON L2/17-036 (MONGOLIAN SUFFIX CONNECTOR)

**Liang Hai** (梁海)

[lianghai@gmail.com](mailto:lianghai@gmail.com)

25 January 2017

This document is a summary of comments on L2/17-036 *Encode Mongolian Suffix Connector (U+180F) To Replace Narrow Non-Breaking Space (U+202F)* from the author and the following experts (ordered alphabetically) – but it should *not* be considered to be a consensus:

**Chimbai** (ᠴᠢᠮᠪᠠᠢ 青柏), **Hasutai** (ᠬᠠᠰᠤᠲᠠᠢ 哈苏台), **Ma Xudong** (马旭东), **Shen Yilei** (沈逸磊),  
**Tengis** (ᠲᠡᠩᠭᠢᠰ 天吉思), **Wang Yihua** (王奕桦).

### 1. General controversies on the NNBSF issue

We've been discussing NNBSF issues for so long that often fail to see those "expected" behaviors we take for granted are actually controversial, either based on controversial theories or not applicable to all the languages that use the Mongolian script.

**1.1.** It's based on controversial grammatical analysis whether a certain suffix-alike structure is an NNBSF-applicable suffix.

1. Encoding should not be controlled by controversial grammatical analysis, especially since NNBSF doesn't always have visible effect it's purely grammatical to decide whether NNBSF should be used.

**1.2.** The space before suffixes is often okay for line breaking.

1. Non-breaking is a typographical choice under certain grammatical analysis, but not an orthographical correct-or-wrong shaping behavior, and is not always preferred. Such behavior shouldn't be forced by the encoding.
2. In printed materials it's common to see a suffix at the beginning of a line, which can't be always considered simply wrong. It's even more common to see a suffix occupying its line alone when text is set as one word (or a few words) per line. Even in Quejingzhabu's *Rules draft 9* (25 Dec 2012) it's said NNBS should allow line break in special cases.
3. Native users are more aware of visual words than the grammatical words defined by certain grammatical analysis. 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.
4. NNBS's current behavior makes it impossible to force a line break before the suffix without breaking the suffix's special shaping (hard line break after NNBS) or moving the NNBS to the next line (hard line break before NNBS, which takes space at the beginning of next line).
5. Encoding should not enforce a non-breaking behavior according to controversial prescriptive grammar. Encoding must allow users to force a line break without breaking orthographical shaping.
6. Therefore it's better to let users control whether the space should be non-breaking.

**1.3.** The space before suffixes is not often preferred to extend the word (as in word counting or cursor jumping or whole word search).

1. Manchu and Sibe consider suffixes as separate words. See [L2/17-008 Proposal to encode one Manchu format character](#).
2. The prescriptive grammar rule of not considering suffixes as separate words doesn't exactly reflect how native users feel.
3. For such a controversial topic, the encoding should choose a simple model and leave the grammatical analysis to natural language processing at higher levels.

**1.4.** It might be a myth ever since the introduction of NNBSB's Mongolian script usage to say it's 1/3 ordinary space.

1. Manchu and Sibe require the space to be visually the same as an ordinary space – both width and stretchability. See [L2/17-008 Proposal to encode one Manchu format character](#).
2. Although suggested as "1/3 the width of an ordinary space" by the Unicode Core Specification, the space before suffixes often just has the width of an ordinary space. It's not as narrow as 1/3 ordinary space even in latest fonts (although often narrower than the ordinary space).
3. According to Quejingzhabu's *Rules draft 9* (25 Dec 2012), NNBSB should be 1/3 *em* (meant to be a bit narrower than an 1/2 *em* ordinary space – a common practice in legacy CJK fonts) instead of 1/3 *ordinary space*. (Also in this work though, MVS is defined as 1/6 *em* which doesn't really make sense because, MVS is removed during shaping and the spacing at MVS's place should be handled by the glyphs before and after it.)
4. Bottom line, it could be the same wide as an ordinary space or a bit narrower, but this is a typographical preference, not orthographical requirement. Encoding should not enforce certain width or stretchability.
5. Therefore it's better to let users control the space's width and stretchability.

## 2. Alternative solutions

**2.1.** We can replace NNBSB with a sequence `<*SP, x>` (instead of a single character), where `*SP` is any space (of any width, breakable or not, up to users' choice) and `x` is a format character prefixed *only* to suffixes that do have an irregular shape.

1. So it can be simply told from text appearance whether the format character is needed. Independent from grammatical analysis.
2. User controllable non-breaking behavior.
3. User controllable width and stretchability.
4. Higher levels' natural language processing should tailor word boundaries.

**2.2.** Shen Yilei suggested `<NNBSP, ZWSP>` for allowing line breaking if we don't want to replace NNBSP with something else.

1. If the existence of ZWSP is taken into considerable by font developers, Can we get expected correct shaping under current OpenType shaping model when the suffix after ZWSP is moved to the next line?

**2.3.** A silly question: Why can't we use FVS for irregular suffixes?

### **3. On the proposal itself**

#### **3.1. Identical to MVS?**

Page 1, 1st paragraph; page 4, 1st paragraph:

*... The definition of this new character can follow that of the U+180E Mongolian Vowel Separator ...*

Page 3, last paragraph:

*... The requirements for the tsatslag connector and the requirements of the suffix connector are identical ...*

1. MVS is a `cf` (Other, format) character and is removed from glyph sequence (a non-spacing invisible character from a user's view), but MSC is meant to be spacing and kept.
2. Can MSC be `Default_Ignorable_Code_Point` like MVS?

#### **3.2. Width myth?**

Page 2, section "1. Incorrect Shaping", 3rd paragraph:

*... The space is not a full-width space, but instead about 1/3 of a regular U+0020 space ...*

See section 1.4 in this document for discussion.

#### **3.3. "Word" definition?**

Page 2, section "1. Incorrect Shaping", 3rd paragraph.

1. It's said, "the definition of a word is very important" – then the definition of a "word" must be clarified here. You can't just say "the space is actually a part of the word" without providing a theory.

### **3.4. Enforced non-breaking behavior and extended word-boundary?**

Page 3, section "2. Incorrect breaking behavior". Page 6-8, Appendix III.

See section 1.2 and 1.3 in this document for discussion.

### **3.5. Applicability?**

Page 11, "Directive Case"; page 12, "Negation", "Regular Action":

*... may or may not use NNBSF ...*

Should be clarified why. See also section 1.1 in this document.