

On Mongolian punctuation marks

L2/19-132

Author: SHEN Yilei

Date: 2019-04-02

1 Whitespace (non-)incorporation

In Mongolian scripts, punctuation marks are always padded with noticeable whitespaces as wide as interword spaces when adjacent to anything. However, various vendors show discrepancies in treating these whitespaces:

- A majority of vendors incorporate preceding whitespaces into the Mongolian comma, period, and colon only;
- Oyun incorporates preceding and following whitespaces into the Mongolian comma, period, and colon only;
- Abkai doesn't incorporate whitespaces;
- Almas' Mongolian White incorporates preceding and following whitespaces into all dedicated Mongolian punctuation marks;
-

Below is a comparison of whitespace incorporation among major vendors:

Table 1 Comparison of whitespace incorporation among major vendors

	SP	Mong. comma	Mong. period	Mong. colon	Mong. ellipsis	Mong. four dots
Abkai	□	◻	◻◻	◻	◻◻◻◻	◻◻◻◻
Oyun	□	◻	◻◻	◻	◻◻◻◻	◻◻◻◻
Menk	□	◻	◻◻	◻	◻◻◻◻	◻◻◻◻
MonBaiti	□	◻	◻◻	◻	◻◻◻◻	◻◻◻◻
Noto SM	□	◻	◻◻	◻	◻◻◻◻	◻◻◻◻
Huaguang	□	◻	◻◻	◻	◻◻◻◻	◻◻◻◻
Fangzheng	□	◻	◻◻	◻	◻◻◻◻	◻◻◻◻
Orhon	□	◻	◻◻	◻	◻◻◻◻	◻◻◻◻
Almas MW	□	◻	◻◻	◻	◻◻◻◻	◻◻◻◻

Thus, it is vital to standardize the whitespace (non-)incorporation behavior for text prepared with one font to be rendered properly with another font.

There are some quick tests that can help us to decide whether to incorporate whitespaces or not:

- **Alignment test:** Incorporated whitespaces are to be measured for line alignment. CJK punctuation marks are typical examples of incorporated whitespaces, where the virtual em-boxes of the punctuation marks are aligned however large the side-bearings are. However, we expect Mongolian punctuation marks themselves to be aligned at line edges with letters.
- **Stretching/compression test:** Whitespace stretching and compression in line justification become rather complicated with incorporated whitespaces if at all possible, taking whitespace balancing into consideration. This does not impact CJK punctuation marks because we do not frequently encounter adjustment wider than an em, and there is no need to balance two whitespaces in CJK typography.
- **Quotation test:** Incorporated whitespaces cannot be eliminated when enclosed by a pair of quotes or brackets in other scripts. Again, when CJK punctuation marks are cited in other scripts, we do expect the whole em-box to be shown. However, incorporated whitespaces prevent us from quoting Mongolian punctuation marks without them,

and asymmetrical whitespace incorporation even forces us to manually pad a space character on the other side, which is much more awkward.

In conclusion, whitespace incorporation is not a good idea for Mongolian punctuation marks.

The real problem is with those punctuation marks borrowed from CJK blocks, like most of Sibe punctuation marks, where their typical CJK usages do incorporate wide whitespaces. Rendering Mongolian punctuation marks into CJK glyphs can further mislead typists to omit expected space characters, resulting in inconsistent text input. Luckily, there is one way out: we can reasonably resort to script analysis in absence of language tagging information, because:

- An ambiguous character used as a Mongolian punctuation mark is always padded with space characters at least on one side of it in running text, while it is never adjacent to space characters when used as a CJK punctuation mark, even when CJK text is mixed with western text.

Whether independent new characters should be proposed for these punctuation marks needs further discussion beyond the scope of this paper.

2 Character orientation

Let's recall some basic concepts of character orientation: (informal definition)

- A character is displayed *upright* if it keeps the absolute orientation across vertical and horizontal texts, like a CJK ideograph;
- Otherwise it is displayed *sideways* either in vertical text or in horizontal text, like a Latin letter or a Mongolian letter.

Most Mongolian punctuation marks are displayed sideways, except those listed below, which are displayed upright:

Table 2 List of upright Mongolian punctuation marks

Code point	Glyph	Hudum	Todo	Sibe	Current value	Proposed value
U+203C	!!	1	1	1	U	
U+2047	??	1	1	1	U	
U+2048	?!	1	1	1	U	
U+2049	!?	1	1	1	U	
U+3001	、		Enumeration	Enumeration	Tu	
U+3002	。		Period	Period	Tu	
U+FF01	!	Exclamation	Exclamation	Exclamation	Tu	
U+FF0C	,		Comma	Comma	Tu	
U+FF1A	:			Colon	Tr	Tu
U+FF1B	;	Semicolon	Semicolon	Semicolon	Tr	Tu
U+FF1F	?	Question	Question	Question	Tu	

- U (“Simple upright”): Upright, with no modification in glyphs in general.
- Tu (“Transformed upright”): Upright as a fall back, but involving glyph alternation in general.
- Tr (“Transformed rotated”): Sideways as a fall back, but involving glyph alternation in general.

All these upright punctuation marks are shared with CJK scripts.

It should be noted that Chinese national standards recommend upright Vertical Forms for Mongolian punctuation marks if there is one, even for sideways punctuation marks. Vertical Forms are technical legacies originally only for presentation purposes and are ideally rendered in horizontal layout always with exactly identical glyphs to those in vertical layout. Using of these characters in running text may cause unexpected behavior in various environments and should be deprecated, leaving aside the orientational mismatches for punctuation marks displayed sideways in horizontal layout.

Sideways Mongolian punctuation marks have the Vertical_Orientation value R or Tr, and upright Mongolian punctuation marks generally have the Vertical_Orientation value U or Tu, as expected. The only two exceptions are U+FF1A as the Todo and Sibe semicolon and U+FF1B as the Todo and Sibe colon, displayed in the Mongolian scripts

(and also in Chinese) upright but have the value Tr. The colon’s value can be accounted for by the fact that Japanese uses it sideways in vertical layout, but the semicolon’s lacks such motivation as Japanese uses it predominantly upright. In addition, as Japanese, especially in vertical context, do not use these characters as an integral part of the script as the Mongolian scripts and Chinese do, their Vertical_Orientation values are better changed to Tu.

3 Line breaking

I would like to introduce the Line Breaking map (LB map), a new tool for analyzing line breakability, before going further into this issue. When talking of a punctuation mark’s line breakability, we are primarily concerned with the situation when the punctuation mark is adjacent to alphabetic characters (or ideographic characters for ideographic scripts, which does not concern us for the moment), with or without padding spaces. Based on the previous conclusion of whitespace non-incorporation, all Mongolian punctuation marks are padded with space characters when adjacent to anything. Thus, there are logically four possibilities of break opportunity when a Mongolian punctuation mark is flanked by letters with padding spaces: (break allowed/forbidden before, break allowed/forbidden after). For convenience’s sake, we use short notations like (¬j, j) instead of (break allowed before, break forbidden after), where:

- j (“joined”) stands for no break opportunity despite the presence of spaces;
- ¬j is the opposite of j, i.e., break allowed in presence of one or more spaces.

Mongolian punctuation marks are then classified into four types, as shown below in a simplified version of the LB map, with predefined values of Unicode character property Line_Break, as specified in UAX14, located therein:

Table 3 Simplified LB map for Mongolian punctuation marks

Before\After	¬j “Break after”	j “No break after”
¬j “Break before”	(¬j, ¬j) “Break before & after” (All other classes)	(¬j, j) “Break before only” OP
j “No break before”	(j, ¬j) “Break after only” EX CL SY ZW IS CP WJ	(j, j) “No break” (No member)

Most of Mongolian punctuation marks fall into one of the two major types, (¬j, j) (“break before only”) and (j, ¬j) (“break after only”):

- Mongolian punctuation marks of type (¬j, j) (“break before only”):

Table 4 List of Mongolian punctuation marks of type (¬j, j) (“break before only”)

Code point	Glyph	Hudum	Todo	Sibe	Manchu	Current value	Proposed value
U+1800	᠎	Siddham	Siddham			AL	OP
U+3008	᠎	Quotation	Title	Title		OP	
U+300A	᠎	Quotation	Title	Title		OP	
U+300C	᠎		Quotation	Quotation		OP	
U+300E	᠎		Quotation	Quotation		OP	
U+FF08	(Parenthesis	Parenthesis	Parenthesis		OP	
U+FF3B	[Bracket	Bracket	Bracket		OP	
U+11660...1166C	᠎ ... ᠎	Siddham	Siddham			BB	OP

- OP (“Open Punctuation”): Prohibit line breaks after.
- AL (“Alphabetic”): Are alphabetic characters or symbols that are used with alphabetic characters.
- BB (“Break before”): Generally provide a line break opportunity before the character.

- Mongolian punctuation marks of type (j, ¬j) (“break after only”):

Table 5 List of Mongolian punctuation marks of type (j, -j) (“break after only”)

Code point	Glyph	Hudum	Todo	Sibe	Manchu	Current value	Proposed value	
U+1801	…	Ellipsis				AL	EX	
U+1802	·	Comma	Old comma			EX		
U+1803	·	Period	Old period			EX		
U+1804	:	Colon	Colon			BA	EX	
U+1805	⋮	Four dots	Four dots			BA	EX	
U+1808	᠎					Comma	EX	
U+1809	᠎					Period	EX	
U+203C	᠎	1	1	1		NS	EX	
U+203D	᠎	1	1	1		NS	EX	
U+2047	᠎	1	1	1		NS	EX	
U+2048	᠎	1	1	1		NS	EX	
U+2049	᠎	1	1	1		NS	EX	
U+3001	᠎	Enumeration		Enumeration		CL		
U+3002	᠎	Period		Period		CL		
U+3009	᠎	Quotation	Title	Title		CL		
U+300B	᠎	Quotation	Title	Title		CL		
U+300D	᠎	Quotation		Quotation		CL		
U+300F	᠎	Quotation		Quotation		CL		
U+FF01	!	Exclamation	Exclamation	Exclamation		EX		
U+FF09)	Parenthesis	Parenthesis	Parenthesis		CL		
U+FF0C	,	Comma		Comma		CL		
U+FF1A	:					Colon	NS	EX
U+FF1B	;	Semicolon	Semicolon	Semicolon		NS	EX	
U+FF1F	?	Question	Question	Question		EX		
U+FF3D	᠎	Bracket	Bracket	Bracket		CL		

- CL (“Close Punctuation”): Prohibit line breaks before.
- EX (“Exclamation/Interrogation”): Prohibit line breaks before.
- AL (“Alphabetic”): Are alphabetic characters or symbols that are used with alphabetic characters.
- BA (“Break After”): Generally provide a line break opportunity after the character.
- NS (“Nonstarter”): Allow only indirect line breaks before.

The present values of line breaking property of some characters contradict their expected line breaking behavior as Mongolian punctuation marks, as shown in red in the tables above. Further inspections show that no uses of these characters in other scripts contradict the newly proposed values. Therefore, the line breaking property values of these characters should be changed as specified above. See Appendix A for a more general analysis.

4 Summary

- Characters should not incorporate significant whitespaces when used as Mongolian punctuation marks. These spaces should be explicitly typed out as space characters.
- The Vertical_Orientation values of U+FF1A and U+FF1B should be changed to Tu.
- The Line_Break values of U+1800 and U+11660...1166C should be changed to OP; the Line_Break values of U+1801, U+1804, U+1805, U+203C, U+203D, U+2047, U+2048, U+2049, U+FF1A, and U+FF1B should be changed to EX.

A A more general analysis with the LB map

More generally, breakability on either side of a character when adjacent to alphabetic characters (AL) or ideographic characters (ID), with or without padding spaces, can be described with a quaternary attribute of {b, f, v, j}, where:

- b (“breakable”) stands for break opportunity in absence of any space;
- f (“fluid”) stands for breaking behavior sensitive to the AL/ID distinction: it behaves like type s when adjacent to AL characters, and like type b when adjacent to ID characters;
- s (“space-conditioned”) stands for break opportunity only in presence of one or more spaces;
- j (“joined”) stands for no break opportunity despite the presence of spaces.

The relations of the four attribute values are better presented with a table:

Table 6 Comparison of {b, f, s, j} regarding break opportunities in various contexts

b	f	s	j	
✓	✓	✓	✗	SP __; __ SP Mostly in alphabetic scripts; especially in Mongolian
✓	✓	✗	✗	ID __; __ ID Mostly in ideographic scripts
✓	✗	✗	✗	AL __; __ AL Mostly in alphabetic scripts

Thus, a 4×4 LB map can be drawn to classify punctuation marks in general:

Table 7 The 4×4 LB map

	b	f	s	j
b	(b, b) ID... B2 RI	(b, f)	(b, s) BB	(b, j)
f	(f, b)	(f, f) AL HL NU	(f, s) PR	(f, j) OP
s	(s, b) NS BA HY IN	(s, f) PO	(s, s) ZWJ GL QU	(s, j)
j	(j, b) ZW EX CL SY	(j, f) IS CP	(j, s) WJ	(j, j)

For ideographic scripts, only the border between f and s is significant, because punctuation marks in these scripts are rarely adjacent to letters or spaces. However, for alphabetic scripts in general, it is the other way around: the b–f and the s–j borders may be important while the f–s border doesn’t count. (For Mongolian, which is a special case of the alphabetic script, only the distinction of j and b/s/f (= ¬j) is significant, as discussed above.)

Table 8 Simplified versions of the LB map for ideographic and alphabetic scripts

(a) Ideographic scripts					(b) Alphabetic scripts				
	b	f	s	j		b	f	s	j
b	(b/f, b/f)		(b/f, s/j)		b	(b, b)	(b, f/s)		(b, j)
f					f	(f/s, b)		(f/s, f/s)	(f/s, j)
s	(s/j, b/f)		(s/j, s/j)		s				
j					j	(j, b)	(j, f/s)		(j, j)

Now let's examine if the newly proposed Line_Break values for shared punctuation marks are also appropriate values for their non-Mongolian usages. As an illustrative case, the interrobang (‡, U+203D) is also used in CJK scripts and European scripts. When used in CJK scripts, it prohibits a break opportunity with an immediately preceding ideographic character but allows a break opportunity with the immediately following ideographic character, which constrains its value to (s/j, b/f). Likewise, when used in European scripts, it prohibits a break opportunity with an immediately preceding alphabetic character but allows a break opportunity with the following space, which constrains its value to (f/s/j, b/f/s). As the conclusion we have come to above, its Mongolian behavior constrains its value to (j, b/f/s). A conjunction of the three constraints confines the Line_Break value of the interrobang to (j, b/f). Therefore, EX as a member of (j, b/f) is an appropriate Line_Break value for the interrobang.

Table 9 Illustration of constraint conjunction in the LB map

	b	f	s	j	
b	(b, b)	(b, f)	(b, s)	(b, j)	Red: CJK (s/j, b/f); Green: European (f/s/j, b/f/s); Blue: Mongolian (j, b/f/s); Yellow: Conjunction (j, b/f)
f	(f, b)	(f, f)	(f, s)	(f, j)	
s	(s, b)	(s, f)	(s, s)	(s, j)	
j	(j, b)	(j, f)	(j, s)	(j, j)	

(End of document)