L2/23-272

Source:	CheonHyeong Sim (沈天珩 / ᠱᠧᠨ ᠲᠢᠶᠠᠨ ᠾᠧᠩ / ᡧᡝᠨ ᡨᡳᠶᠠᠨ ᡥᡝᠩ / ꡚꡦꡋ ꡉꡞꡗꡋ ꡜꡦꡃ / ཤེན་ཐིའན་ཧེང་)
Title:	Propose to Add Script_Extension for some CJK Punctuations
Date:	2023−11−18
Action:	To be considered by UTC and Script Ad Hoc

Proposal

This proposal requests to add some Script_Extension values for some CJK Punctuations as in the following table:

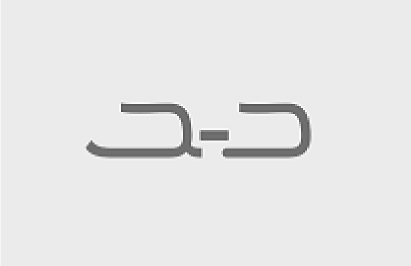
3001	、	IDEOGRAPHIC COMMA	Bopo Hang Hani Hira Kana Mong Yiii
3002	。	IDEOGRAPHIC FULL STOP	Bopo Hang Hani Hira Kana Mong Phag Yiii
3008	〈	LEFT ANGLE BRACKET	Bopo Hang Hani Hira Kana Mong Tibt Yiii
3009	〉	RIGHT ANGLE BRACKET	Bopo Hang Hani Hira Kana Mong Tibt Yiii
300A	《	LEFT DOUBLE ANGLE BRACKET	Bopo Hang Hani Hira Kana Mong Tibt Yiii
300B	》	RIGHT DOUBLE ANGLE BRACKET	Bopo Hang Hani Hira Kana Mong Tibt Yiii

Background

Currently, some CJK punctuations used in Mongolian (Hudum & Sibe writing system), Tibetan and Phags−pa with a Script value “Zyyy” (aka Common) have the Script_Extension values other than “Zyyy”, which may lead to some rendering issues in the plain text.

Although “rich text” environments are becoming more common, including those that support language−tagging and the OpenType “locl” feature, “plain text” environments persist, and are likely to continue to persist for a long time due to their robust nature. In addition, environments that support variation sequences vastly outnumber those that support language−tagging. (Ken Lunde, 2023)

Different fonts are usually used for different scripts, but problems may easily occur with the characters used in multiple scripts when applying font fallback in the plain text. The picture below shows U+0712 U+0640 U+0712 (ܒـܒ) under the default font on Android9:



In the above example, U+0712 is a Dual–joining Syriac character, while U+0640 is tatweel in the Arabic block, with the Script_Extension value "Syrc", which means it can be used not only in Arabic script but also in Syriac script to lengthen the stem. Actually, there is no problem when using the font Noto Sans Syriac alone; but under the default environment, Noto Sans Arabic is called first, so U+0640 is displayed with Noto Sans Arabic, but there is no U+0712 in Noto Sans Arabic, so it falls back to Noto Sans Syriac. Displaying in different fonts leads to the disconnection between the characters, even if the initial form and the final form of U+0712 are rendered correctly. This kind of problem can be found in many scripts.

In order to solve this kind of problem, mixing the characters of multiple scripts into a single font, and using OpenType Layout to do the script–specific (but not language–specific) localizations seems to be a good method.

Before the OpenType Layout features applying to the glyphs, calculating the script for each character is necessary. A character with a Script value "Zyyy" normally inherits the value of the preceding character unless there is no character preceded. In this case, the rendering system can judge the script for a character used in multiple scripts according to the actual context, and applying the script–specific localizations by default easily.

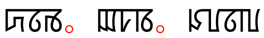
For the characters with a Script value "Zyyy" and a Script_Extension value other than "Zyyy", different rendering systems often have different calculating methods. As a comparison, on Android Systems, the Script_Extension values are completely ignored; but in Google Chrome (on multiple systems), these characters would only inherit the scripts in the Script_Extension. For example, U+202F (NNBSP) has the Script_Extension value "Latn" and "Mong", if it is followed by a character with a Script value "Hani" and preceded by a character with a Script value "Mong" – just imagine there are two Sibe people chatting on an instant messenger software – that is normally a plain text environment, one asking another "what is the meaning of 字" (for ease of typesetting, the Sibe characters below will use the horizontal layout, since the method itself is completely the same between the horizontal layout and the vertical layout):

字 ᡳ ᡤᡡᠨᡳᠨ ᠶᠠᠪᠠ ᠠᡳ ᠰᡝᠮᠪᡳ ?

Here, in the environment of Google Chrome, the NNBSP inherits the value "Mong" of the following character but not the value "Hani" of the preceding character, because "Hani" does not occur in the Script_Extension value of NNBSP. Since the NNBSP got the same

Script value as its following character, a context–based shaping can be applied, and becomes the case particle form automatically without any free variation selectors. Meanwhile, this sequence could **NOT** be rendered as the right glyphs on Android Systems.

It seems that the method Google Chrome uses is better. However, there comes out a new issue. Since the Core Spec says, "Additionally, a small circle punctuation mark is used in some printed Phags–pa texts. This mark can be represented by U+3002 IDEOGRAPHIC FULL STOP, but for Phags–pa the *ideographic full stop* should be centered, not positioned to one side of the column. This follows traditional, historic practice for rendering the ideographic full stop in Chinese text, rather than more modern typography.", it is obvious that U+3002 is needed for Phags–pa with a different required shape to the default one, but the Script_Extension for U+3002 even does not contain "Phag". In such fonts mixed multiple scripts, the script–specific localization would not apply correctly in Google Chrome due to the lack of the Script_Extension "Phag". Since this file is generated by Google Chrome, let us see what will happen if U+3002 follows and precedes Phags–pa characters simutaneously:



Regrettably, U+3002 is not centered automatically, even if the script–specific localization exists in OpenType Layout – because U+3002 here goes by "Zyyy" ("DFLT" in OpenType Layout) instead of "Phag". So the conclusion here is that, adding "Phag" for U+3002 is necessary in order to improve the cross–platform compatibility. It is also the same for Mongolian (Hudum & Sibe writing system) and Tibetan, and the evidences showing Hudum, Sibe and Tibetan need these characters with different shapes to the default one will be given out later in this proposal.

Sibe punctuations

Different from the Hudum writing system, the Manchu writing system, etc., the orthography of Sibe writing system uses the CJK punctuations but not the traditional Mongolian ones.

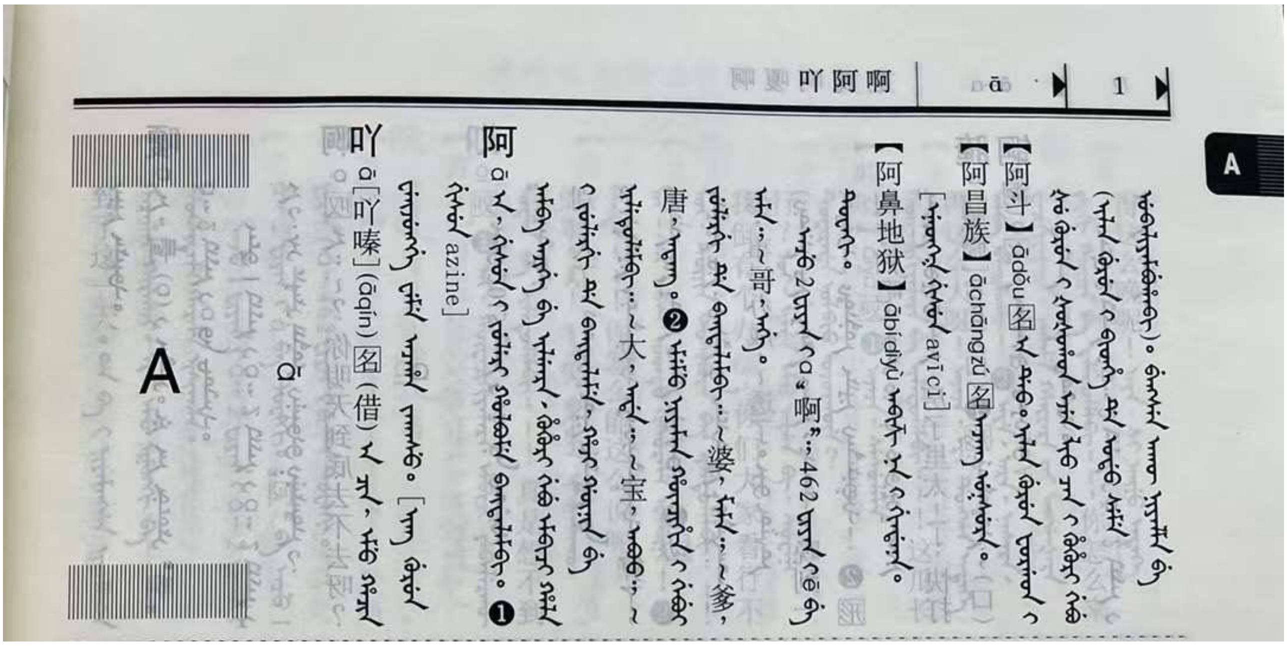


Fig.1 《汉锡词典》P1

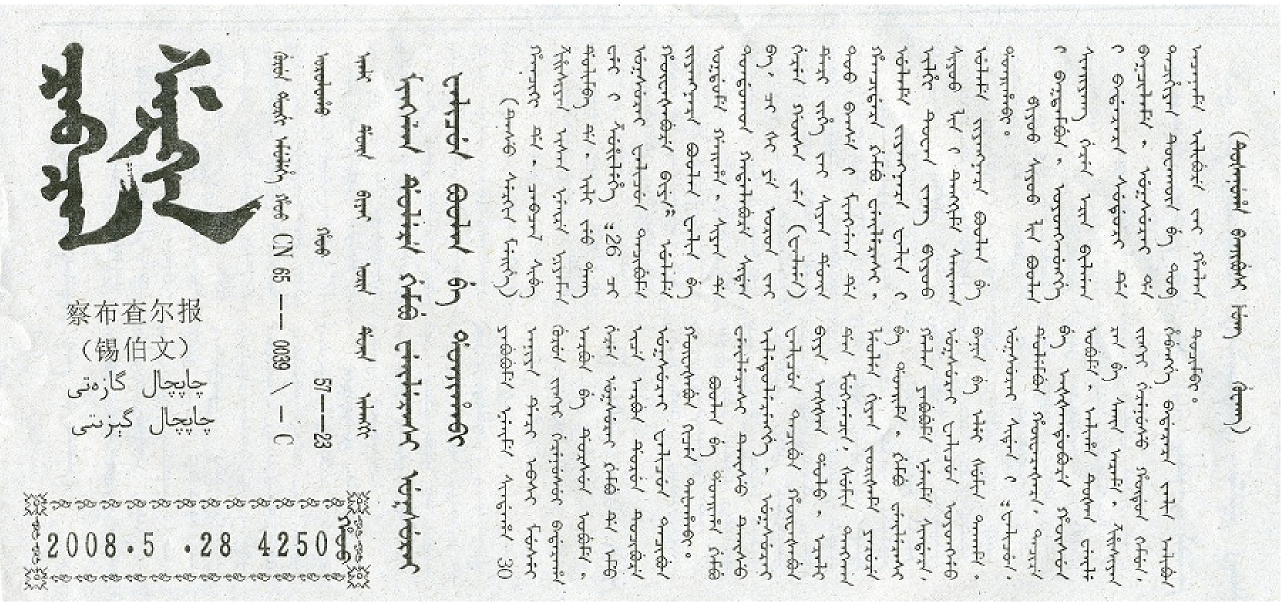


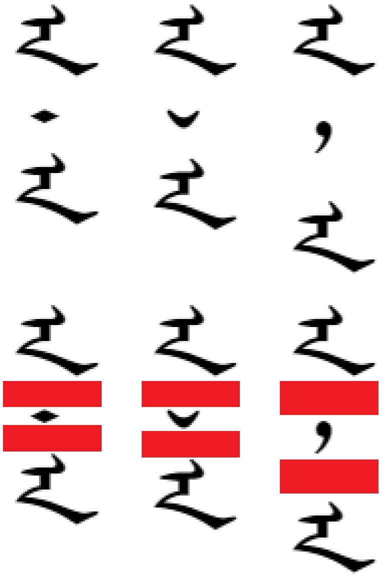
Fig.2 《察布查尔报》(锡伯文） 2008/05/28

From the multiple sources of Sibe texts above, we can easily see that at least these CJK punctuations are used:

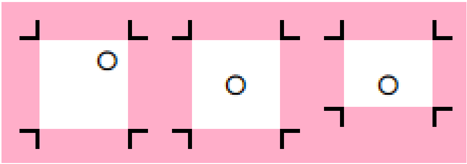
3001	、	IDEOGRAPHIC COMMA
3002	。	IDEOGRAPHIC FULL STOP
FF0C	，	FULLWIDTH COMMA
FF1A	：	FULLWIDTH COLON
FF1B	；	FULLWIDTH SEMICOLON

For U+FF0C, U+FF1A and U+FF1B, their Script_Extension values are “Zyyy”, in other words, they can inherit the Script value of the preceding character unconstrainedly, so no action should be taken. Actually, the exclamation mark (U+FF01) and the question mark (U+FF1F) are also used in Sibe orthography, but they are inconsequential since their Script_Extension values are also “Zyyy”, no action should be taken either.

Another important thing is, these punctuations used in Sibe are **NOT** full width. At the character level, a punctuation usually follows a word closely, but precedes a word with an extra space (U+0020), which is the same as Latin punctuations, Mongolian punctuations, etc., in order to break lines and justify texts correctly; however, at the glyph level, normally for both Mongolian punctuations and the CJK punctuations used in Sibe, the actual space above them and below them are the same:



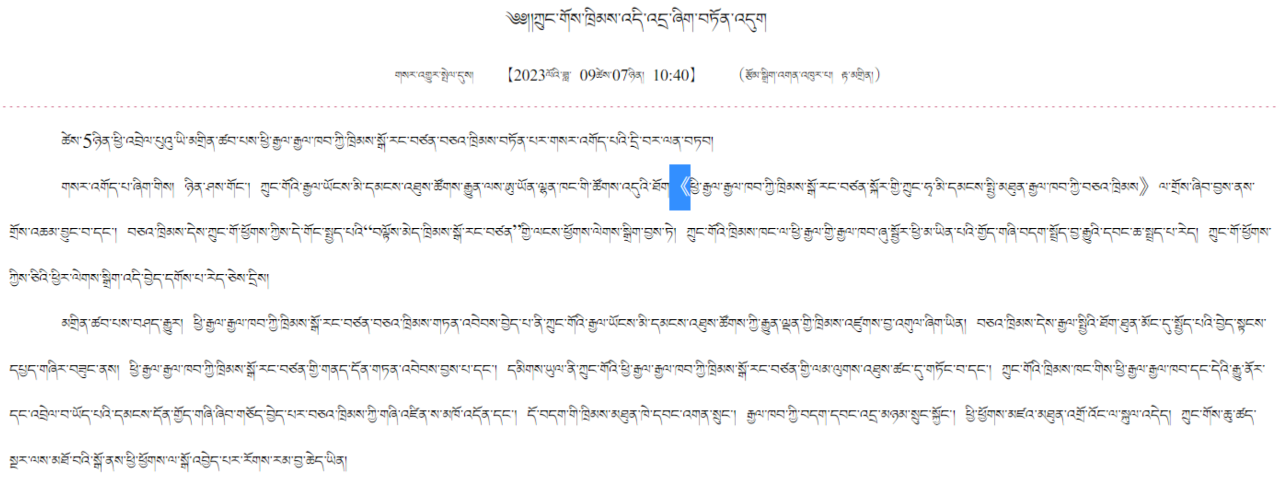
In order to effect this, the above space should be together with the punctuation in a single glyph, while the below space is just a single character U+0020. So, as a conclusion, the picture below contrasts the glyph of U+3002 in different environments (in vertical layout):



The first glyph shows a default glyph of U+3002 in vertical layout, the second one shows a language−specific glyph of Traditional Chinese (Hong Kong, Macao, Taiwan), the third one shows a script−specific glyph of Mongolian and Phags−pa script. Since these punctuations used in Sibe are **NOT** full width, simply adding VS2 to change U+3001 or U+3002 into the centered form is still not enough, so adding the Script_Extension values is necessary.

Hudum & Tibetan punctuations

In the traditional orthography, there are no book title marks — the book title marks are characteristics in East Asia — they do not even occur in the Western scripts. However, in the modern orthography, Hudum and Tibetan borrow the book title marks from Chinese. Many Mongolian and Tibetan news websites are now widely using the book title marks, for example, the picture below is a screenshot from http://tibet.people.com.cn/16016973.html:



༄༅།།ཀྲུང་གོས་ཁྲིམས་འདི་འདྲ་ཞིག་བཏོན་འདུག

གསར་འགྱུར་སྤེལ་དུས། 【2023ལོའི་ཟླ 09ཚེས་07ཉིན། 10:40】 （རྩོམ་སྒྲིག་འགན་འཁུར་པ། ཧྲ་མཁྱེན།）

ཚེས་5ཉིན་ཕྱི་འབྲེལ་ཕྱུའུ་ཡི་མཁྱེན་ཚབ་པས་ཕྱི་རྒྱལ་རྒྱལ་ཁབ་ཀྱི་ཁྲིམས་སྒྲོམ་རང་བཙན་བཅའ་ཁྲིམས་བཏོན་པར་གསར་འགོད་པའི་དྲི་བར་ལན་བཏབ།

གསར་འགོད་པ་ཞིག་གིས། ཉིན་ཤས་གོང་། ཀྲུང་གོའི་རྒྱལ་ཡོངས་མི་དམངས་འཐུས་ཚོགས་རྒྱུན་ལས་ཁུ་ཡོན་ལྷན་ཁང་གི་ཚོགས་འདུའི་ཐོག《ཕྱི་རྒྱལ་རྒྱལ་ཁབ་ཀྱི་ཁྲིམས་སྒྲོམ་རང་བཙན་སྐོར་གྱི་ཀྲུང་ཧྭ་མི་དམངས་སྤྱི་མཐུན་རྒྱལ་ཁབ་ཀྱི་བཅའ་ཁྲིམས》 ལ་གྲོས་ཞིབ་བྱས་ནས་གྲོས་འཆམ་བྱུང་བ་དང་། བཅའ་ཁྲིམས་དེས་ཀྲུང་གོ་ཕྱོགས་ཀྱིས་དེ་གོང་སྤྱད་པའི“བསྒྲོས་མེད་ཁྲིམས་སྒྲོམ་རང་བཙན་”གྱི་ལངས་ཕྱོགས་ལེགས་སྒྲིག་བྱས་ཏེ། ཀྲུང་གོའི་ཁྲིམས་ཁང་ལ་ཕྱི་རྒྱལ་གྱི་རྒྱལ་ཁབ་ཞུ་སྐོར་ཕྱི་མ་ཡིན་པའི་གྱོད་གཞི་བདག་སྐྱོང་བྱ་རྒྱུའི་དབང་ཆ་སྤྲད་པ་རེད། ཀྲུང་གོ་ཕྱོགས་ཀྱིས་ཅིའི་ཕྱིར་ལེགས་སྒྲིག་འདི་བྱེད་དགོས་པ་རེད་ཅེས་དྲིས།

མཁྱེན་ཚབ་པས་བཤད་རྒྱུར། ཕྱི་རྒྱལ་རྒྱལ་ཁབ་ཀྱི་ཁྲིམས་སྒྲོམ་རང་བཙན་བཅའ་ཁྲིམས་གཏན་འབེབས་བྱེད་པ་ནི་ཀྲུང་གོའི་རྒྱལ་ཡོངས་མི་དམངས་འཐུས་ཚོགས་ཀྱི་རྒྱུན་ལྡན་གྱི་ཁྲིམས་འཛུགས་བྱ་འགུལ་ཞིག་ཡིན། བཅའ་ཁྲིམས་དེས་རྒྱལ་སྤྱིའི་ཐོག་ཁྱབ་མོང་དུ་སྤྱོད་པའི་བྱེད་སྟངས་དཔྱད་གཞིར་བཟུང་ནས། ཕྱི་རྒྱལ་རྒྱལ་ཁབ་ཀྱི་ཁྲིམས་སྒྲོམ་རང་བཙན་གྱི་གནད་དོན་གཏན་འབེབས་བྱས་པ་དང་། དམིགས་ཡུལ་ནི་ཀྲུང་གོའི་ཕྱི་རྒྱལ་རྒྱལ་ཁབ་ཀྱི་ཁྲིམས་སྒྲོམ་རང་བཙན་གྱི་ལམ་ལུགས་འཐུས་ཚང་དུ་གཏོང་བ་དང་། ཀྲུང་གོའི་ཁྲིམས་ཁང་གིས་ཕྱི་རྒྱལ་རྒྱལ་ཁབ་དང་དེའི་རྒྱུ་ནོར་དང་འབྲེལ་བ་ཡོད་པའི་དམངས་དོན་གྱོད་གཞི་ཞིབ་གཅོད་བྱེད་པར་བཅའ་ཁྲིམས་ཀྱི་གཞི་འཛིན་ས་མཐོ་འདོན་དང་། དོ་བདག་གི་ཁྲིམས་མཐུན་ཁེ་དབང་འགན་སྲུང་། རྒྱལ་ཁབ་ཀྱི་བདག་དབང་འདྲ་མཉམ་སྲུང་སྐྱོང་། ཕྱི་ཕྱོགས་མཛའ་མཐུན་འགྲོ་འོང་ལ་སྐུལ་འདེད། ཀྲུང་གོས་རྒྱ་ཚད་སྤར་ལས་མཐོ་བའི་སྒོ་ནས་ཕྱི་ཕྱོགས་ལ་སྒོ་འབྱེད་པར་རོགས་རམ་བྱ་ཆེད་ཡིན།

Fig.3 人民网藏文版 2023/09/07

Also, see http://mgyxw.cn/am/2023/10/15/9376_657481.html for reference of the Mongolian (Hudum writing system) example. In the separate Mongolian fonts or Tibetan fonts, the book title marks are made to nearly half width — that is, no extra space either at the left side or at the right side, which is completely different from the glyphs in Chinese style. Although my font has the Opentype Layout feature “kern” and “dist” under Tibetan script, as we can see in the above picture, the glyphs are still in Chinese style, nothing happens to them because the characters cannot inherit the Script value from the Tibetan characters. It looks very uncomfortable with the redundant space. So adding the Script_Extension values is necessary.

Conclusion

To sum up, U+3001 and U+3002 are needed for Mongolian (Sibe writing system), U+3002 is also needed for Phags—pa, and the book title marks (U+3008..U+300B) are needed for both Mongolian (Hudum writing system) and Tibetan.

Note that, even if there is no rendering issue occurs, the additions are **still needed** theoretically, because they are indeed used in the corresponding scripts. The rendering issues are only the straw that breaks the camel’s back.

(End of document)