DATE: 2003-02-13

DOC TYPE: Expert contribution

TITLE: **Use of ZWJ/ZWNJ with Mongolian Variant Selectors and Vowel Separator**

SOURCE: Paul Nelson and Asmus Freytag

STATUS: Proposal

Summary

Display of single characters in combination with variant selectors and vowel separator is needed for books. This topic is not currently well defined. This proposal suggests a clarification of the manner in which the ZWJ/ZWNJ should be used with Mongolian variation selector and vowel separator characters.

We've been in extensive contact with both native users and academicians, as well as implementers and the position put forth here is supported by most members of this community

Background

When the Mongolian block was encoded, the shaping rules were not published at the same time. There is a United Nations University Technical Report (UNU/IIST Report No. 170), which contains the list of variant selection sequences that was published recently in the standard. There is also a Chinese document "Mengguwen Bianma" (Mongolian Encoding) by Quejingzhabu and published by the publishing house of Inner Mongolia University. Mongolia has issued standard MNS 4932:2000 "*Use of Mongolian Character Encoding*." Additionally, "A Users' Agreement Related to the International Standard of Mongolian Encoding" was jointly prepared by China and Mongolia as the two major countries where the Mongolian script is mainly used. These documents together capture the consensus of the WG2 ad-hoc group. However, the ad-hoc group has never completed a single consensus document.

A number of experts and implementers have worked for almost a year on formalizing the rules for Mongolian shaping. It is hoped that the results of this effort can become a Unicode Technical Report when ready.

Use of NARROW NO-BREAK SPACE with Mongolian

The Mongolian script uses the NNBSP between a word and its suffix. When using the NNBSP, the shaping is interrupted, as one would expect, with the suffix beginning a new shaping segment. There are a few exceptions that occur. However, those exceptions should be able to be handled by rules in a smart font or other approach to shaping.

Therefore, Unicode 4.0 should clearly state that supplying a ZWNJ either preceding or following a NNBSP in regular word context would be redundant. Also, the use of the ZWJ is not necessary for the correct display of normal Mongolian text before or after the NNBSP.

Use of Mongolian Vowel Separator
For Mongolian, the MONGOLIAN VOWEL SEPERATOR (MVS) causes the previous word segment to end shaping and begins the following segment in the joined form.

Join controls and variation selectors
Similar to Arabic, the use of the U+200D ZERO WIDTH JOINER (ZWJ) and U+200C ZERO WIDTH NON-JOINER (ZWNJ) with the Mongolian script is required when a particular form of a glyph is wanted in isolation. In other words, the character is not in context of normal text flow.

Unlike Arabic, the Mongolian script also uses free variation selectors to specify specific variant shapes for various forms. The Unicode documentation states, "A free variant selector is encoded after the base character it modifies. Free variant selectors are not productive and are therefore ignored when not immediately preceded by one of their listed base characters." (TUS 3.0, 11.4, p. 291)

This document seeks to resolve an ambiguous situation where variant forms are represented by using a combination of ZWJ/ZWNJ and one of the FREE VARIATION SELECTOR characters. It proposes the order that should be used to enable a standardization of Mongolian text input and processing, but preserves legacy documents.

There are currently two existing implementations of using the ZWJ or ZWNJ in combination with the base character and the FVS characters.

1. At the time the UNU, Chinese and Mongolian documents were written the convention to show the forms in isolation was illustrated as:

   BASE, ZWJ, FVS

   According to Professor Quejingzhabu, this approach "was formulated after repeated examinations and discussions during six to seven years by experts of China, Mongolia, Germany and the Unicode organization." It appears this formulation has been based on the fact that for Mongolian the "base" character for a FVS is a shaped form of a character, not the abstract unshaped form. A literal interpretation of this understanding is seemingly more natural to the writers of these documents and can be found in existing implementations.

2. According to the existing text of the Unicode Standard, the approach 1 above is illegal. Unicode 3.2 states:

"Only the variation sequences specifically defined in the Unicode Character Database in the file StandardizedVariants.html are sanctioned for standard use; in all other cases the variation selector **cannot** change the visual appearance of the preceding base character from what it would have had, in the absence of the variation selector." (emphasis added).

Therefore, the only allowed approach is to place the FVS immediately after the base character and ZWJ/ZWNJ after the FVS.

BASE, FVS, ZWJ

Since the restriction of having the ZWJ/ZWNJ not be placed between the base and the FVS was added several versions ***after*** the encoding of the FVS, existing data is known and expected to violate this rule.

Proposal 1

We propose to relax the restriction on placement of ZWJ and ZWNJ specifically where the Mongolian FVS characters are concerned (but not for general variation selectors) and replace it by the following rules:

a. When both a free variation selector and ZWJ/ZWNJ character follow a Mongolian base character, the free variation selector should be next to the base letter and the ZWJ/ZWNJ should follow the free variation selector. Some examples follow.

Form	**Character order**
initial variant form	BASE, FVS, ZWJ
medial variant form	ZWJ, BASE, FVS, ZWJ
initial ligated forms	BASE, BASE, FVS, ZWJ BASE, FVS, BASE, ZWJ BASE, FVS, BASE, FVS, ZWJ
medial ligated forms	ZWJ, BASE, BASE, FVS, ZWJ ZWJ, BASE, FVS, BASE, ZWJ ZWJ, BASE, FVS, BASE, FVS, ZWJ

b. If the ZWJ/ZWNJ occurs between a Mongolian base character and the FVS, it may be rendered or processed as if it had followed the FVS. In order to interpret legacy data correctly, it is recommended that implementations attempt to render or otherwise process that combination as if it had been in the preferred order.

The Mongolian Typesetting System 9.1 developed in a cooperative effort under Peking University automatically changes the BASE, ZWJ, FVS to BASE, FVS, ZWJ. This would indicate that the movement to support what is currently defined in Unicode is under way in China. However, documents and standards do recommend a manner different from Unicode.

Adopting this proposal will help to standardize Mongolian input and to better enable flowing text and isolated text to be processed with the same approach. This also allows existing documents to continue to be displayed.

Proposal 2
In the Unicode Standard it is currently implied that the ZWJ/ZWNJ characters cannot be placed between a base character and a combining mark, as the ZWJ or ZWNJ would become the new base character. This should also be true of the FVS and other VARIATION SELECTOR (VS) characters, which have tighter coupling to the base character than does combining marks. Thus, an example in Mongolian script would be:

BASE, FVS, COMB, ZWJ

We request that specific wording be added to the text of the Unicode Standard, Version 4.0 to make this more explicit to users of the ZWJ/ZWNJ. Suggested wording for section 15.2 of Unicode 4.0 is:

"If a join control were to be placed in the middle of a combining sequence, it would interrupt it, and become the base character for the combining characters that follow it. Therefore, a join control must be placed after the last combining character. Similarly, when combining marks are inserted after a base character, they have to be inserted before any following join controls."

Pointers to this text might be placed in chapter 5, as well as sections 8.2, 8.3 and 12.2 as needed.

Additional information

The following L2 or SC 2/WG 2 documents are relevant to this discussion:

1. WG2 N1711
2. WG2 N1734
3. http://www.unicode.org/L2/L1998/98251.txt - Feedback from China on WG2 N1734 (Ken Whistler) with regards to the Chinese proposal on Mongolian WG2 N1711.
4. L2/98-251
5. http://www.unicode.org/L2/L1998/98252-moore.txt - Feedback on Ken Whistler's comments (Richard Moore)

UNU/IIST Technical Report No. 170, *"Traditional Mongolian Script in the ISO/IEC 10646 and Unicode Standards"*, Myatav Erdenechimeg, Richard Moore and Yumbayar Namsrai, August 1999

Mengguwen Bianma (Mongolian Encoding), Quejingzhabu, Publishing House of Inner Mongolia University, 2000.

Use of Mongolian Character Encoding MNS 4932:2000, Mongolia National Standards, 2000

These are the people/organizations involved in reviewing the Mongolian shaping algorithm.

- Paul Nelson [paulnel@microsoft.com]
- Ning Jin-Grisaffi [ningj@microsoft.com]
- Martin Heijdra [mheijdra@Princeton.EDU]
- Andrew C. West [andrewcwest@alumni.Princeton.EDU]
- Tim Partridge [timpart@perdix.demon.co.uk]
- Myatav Erdenechimeg [erdeely@yahoo.com]
- Richard Moore [richard@ifad.dk]
- Asmus Freytag [asmusf@unicode.org]
- Quejingzhabu [qj@mail.imu.edu.cn]
- Chen Zhuang []