ISO

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION

ORGANISATION INTERNATIONALE DE NORMALISATION

ISO/IEC JTC1/SC2/WG2

Universal Multiple-Octet Coded Character Set (UCS)

ISO/IEC JTC 1/SC 2/WG 2 *N 1808*
1998-04-30

Title:	**Reply to "Proposal WG2 N1734" Raised at the Seattle Meeting Regarding Proposal WG2 N1711**
Source:	**China**
Action:	**Review and Feedback**
Distribution:	**ISO/IEC JTC1/SC2/WG2**

At the WG2 meeting in Seattle in March 1998, Mr. Ken Whistler made some comments and suggestions (N1734) concerning our Mongolian Encoding Proposal N1711. Later we had a special meeting in Hohhot, at which Mr. Whistler's proposal was discussed in detail. We would have submitted this reply of ours to WG2 after we reached common understanding with Standardization Department of Mongolia, if bad communications had not kept us from receiving any feedback from them though we had informed them of our views on April 14. Our reply is as follows:

1. MONGOLIAN SPACE.

Mr. Whistler suggested to use NO-BREAK SPACE instead of MONGOLIAN SPACE and requested us to further justify why both MONGOLIAN SPACE and NO-BREAK SPACE are adopted. In the Mongolian Encoding System, there is need for a unique space called MONGOLIAN SPACE which differs both in form and function from common SPACE (U+0020) and NO-BREAK SPACE. (U+00A0). Such a space has the following distinctive features:

(1) In form, it represents a gap. On the screen there should be a visual representation of a width different from that of SPACE. In print, there should be a regular gap of one third of a full character which differs from that of SPACE.

(2) This space also has the function of a VARIANT SELECTOR to determine the changed forms of the letters preceding and following it. That is, to determine that the word-final character of the given letter preceding it should be used. As for the form of the character that follows it, it involves a lot of special cases and has to be judged according to what suffix is concerned (For detail see Appendix III, 1, in N1711).

(3) It is used to separate a suffix from the letter's word stem, implying that the gap here is not the bound between character strings of the word.

(4) MONGOLIAN SPACE cannot be used to split a word or a line in two.

(5) MONGOLIAN SPACE appears at a very high frequency. Statistics shows that it appears 28117 times, or 28.12%, in a text of 100,000 words.

As for NO-BREAK SPACE, it remains to be used in the encoding of Mongolian word in its original function. Thus, NO-BREAK SPACE indicates how a word is formed, i.e., how several morphemes of a word are separated by it. For example, the Mongolian word ARADCILAL (Democracy) consists of four morphemes ARA-D-CILA-L, which is written as ARA(NBS)D(NBS)CILA(NBS)L in the word formation column in a computer's dictionary or in the language data. The form and function of NO-BREAK SPACE used in such cases differ from those of MONGOLIAN SPACE:

(1) In appearance, NO-BREAK SPACE does not indicate a gap, so it is NO-BREAK SPACE in the full sense of the term.

(2) It does not have the function of a VARIANT SELECTOR that changes the variant forms of a letter preceding or following it. To use NO-BREAK SPACE or not in a sequence of Mongolian letters does not have an effect on the variant forms of any letter in the sequence.

(3) It does not serve as bound between character strings of a word.

(4) Neither a word or a line is split through syllabication wherever NO-BREAK SPACE is used.

In view of the above, our opinion is to preserve the MONGOLIAN SPACE (   ) as described N1711, a space which is different both from SPACE(U+0020) and NO-BREAK SPACE(U+00A0). Reason for preserving it is that in Mongolian language, SPACE(U+0020) and NO-BREAK SPACE(U+00A0) and MONGOLIAN SPACE have their respective uses. For example, DARUG A NAR UN YARIY A (speeches of the leading officers) is to be stored as
(SP)DARU(NBSP)G(MNJ)A(MSP)NAR(MSP)UN(SP)YARI(NBSP)Y(MNJ)A(SP).

2. MONGOLIAN COMBINATORY SYMBOL(?!).

We agree to Mr. Whistler's opinion, i. e., to include this symbol into U+2047 as a separate script. Such treatment is in accordance with 10646 as it is now. U+203C has already a DOUBLE EXCLAMATION MARK (!!) in 10646 which is exactly the same in nature as MONGOLIAN COMBINATORY SYMBOL.

3. MONGOLIAN POSITIONAL FORMAT CONTROL CHARACTERS.

In the Mongolian encoding system, the POSITIONAL FORMAT CONTROL CHARACTER should be used in the following three cases:

(1) Where there is need to show the presentation form of a variant not found in a word, thus, where there is need to show the initial form of the basic script A, we have to use POSITION CONTROL CHARACTER for the initial position; and where there is need to show the medium form of the basic script G, this POSITION CONTROL CHARACTER is to be used.

(2) Where there is need to split a word, e.g., the word SURGAGULI (School) is to be syllabicated into SUR GA GU LI with all syllables linked up, then this POSITION CONTROL CHARACTER should be added to the basic scripts so as to show that R,G,A,G,U are in their medium positions. If the POSITION CONTROL CHARACTER is not added to these scripts, R,A,U and I will be shown in their final positions and G, G and L in their initial positions.

(3) In very exceptional cases where variant presentation forms have to be compulsorily shown in any sequence without following regular rules. Thus, to show a medium or a final form in the initial position; or to show an initial or a final form in the medium position, or an initial or a medium form in the final position, etc. In order to show such irregular variant forms, this POSITIONAL FORMAT CONTROL CHARACTER is also required.

Based on a comparison between the six designs of CONTROL CHARACTERS N1510,N1515,N1638, N1691,N1711 and N1734 as well as their uses, we are inclined to hold the following views:

(1) We agree to use ZERO WIDTH JOINER(U+200D) and ZERO WIDTH NON-JOINER (U+200C) as POSITION CONTROL CHARACTERS for Mongolian text.

(2)In order to make ZERO WIDTH JOINER (U+200D) and ZERO WIDTH NON-JOINER(U+200C) visible and distinguishable in case of need, a SHOW HIDDEN CHARACTER mode can be used.

4.MONGOLIAN FREE VARIANT SELECTOR CHARACTERS(FVS1,FVS2 and FVS3).

In Proposal N1691, we have considered to use two FREE VARIANT SELECTOR CHARACTERS. The reason why we were inclined to give MONGOLIAN NIRUGU certain function of a CONTROL CHARACTER (i.e., to show one of the four medium forms of MLM.I with MONGOLIAN NIRUGU) and technically treat a few characters (e.g., to treat two of the four medium forms of the ML.QA as final forms) was altogether to remove the FREE VARIANT SELECTOR 3 which is so rarely used. However, in so doing, we gave MONGOLIAN NIRUGU a double function; technically treated certain characters in a way not in line with regular habits for Mongolian writing; such being the case, we began to prefer preserving FREE VARIANT SELECTOR 3. Statistics show that FREE VARIANT SELECTOR 3 ought to be used for the medium form of ML.QA, medium form of ML.GA, medium form of MLM.I, medium form of MLM.KA and final form of MLA.A. That is why we preserved all three FREE VARIANT SELECTORS in Proposal N1711.

In view of the concrete condition of Mongolian texts, our conclusion is as fo
llows:

(1)Three FREE VARIANT SELECTOR CHARACTERS are all needed, viz., MONGOLIAN FREE VARIANT SELECTOR CHARACTER 1 (FVS1), MONGOLIAN FREE VARIANT SELECTOR CHARACTER 2 (FVS2) and MONGOLIAN FREE VARIANT SELECTOR CHARACTER 3 (FVS3).

(2) Where to put these three MONGOLIAN FREE VARIANT SELECTOR CHARACTERS is left for WG2 and Unicode Technical Committee to decide.

5. MONGOLIAN VOWEL SEPARATOR(MVS) .

Mr. Whistler says that technically a sequence like ML.NA+MVS+ML.A can be shown by means of the sequence ML.NA+NON-JOINER+ML.A+FVS2, to which we agree, for it is feasible to make the latter sequence function as a VOWEL SEPARATOR. Then a question arises: in his proposal to use ZERO WIDTH JOINER and ZERO WIDTH NON-JOINER as POSITION CHARACTERS in Mongolian texts, Mr. Whistler says that "-iFf-" can be represented by "-bBNJJb-", we may then ask, if "-mFf-" should be represented by "-bBNJJb-"? If so, it seems that the sequence ML.NA+MVS+ML.A can also be represented by ML.NA+NJ+J+ML.A+FVS1. In such case, will the use of NON-JOINER become a use which is not unified ? No matter how NON-JOINER is treated, its use here involves one or two more characters than if we design a special character. Moreover, in normal writings, this MONGOLIAN VOWEL SELECTOR has a high frequency of appearance, thus, statistics show that it appears 12339 times, or 12.34%, in a text of 100,000 words. But where NON-JOINER is used, at least are required two diacritical marks, NON-JOINER and FVS2, which will naturally result in recording and storing twice as many DIACRITICAL MARKS. One DIACRITICAL MARK will suffice if we use the specially designed VOWEL SELECTOR. This is a problem to be taken into proper consideration in dealing with DIACRITICAL MARKS that appear so frequently in normal writings. In order to lessen recording and storing work, we insist that this special character be preserved. What is more, such treatment will also facilitate Mongolian-Latin transliteration, because in Mongolian studies we usually use a lower dash to represent such a sequence, e.g., N_A.

In view of the above, our opinion is:

(1) to preserve the special character (MNJ).

(2) to change its name to MONGOLIAN VOWEL ZERO WIDTH NON-JOINER, as is proposed by Mr. Whistler.

(3) to leave for WG2 and Unicode Technical Committee to decide where to put this character.

6.MONGOLIAN TODO SOFT HYPHEN(|).

In a Mongolian Todo text, this is the regular hyphen used at the beginning of the next line when a word is syllabicated with a few of its syllables removed there. For example, the word AYIMAGCILAL can be syllabicated like this:

"*** ********* ****** AYIMAG -CILAL ****** ****** *** ******"**

Seeing the above, our opinion is:

(1) to preserve this MONGOLIAN TODO HYPHEN (|).

(2) to call it MONGOLIAN TODO HYPHEN as suggested by Mr. Whistler.