ISO

INTERNATIONAL ORGANIZATION FOR SANDARDIZATION

ORGANISATION INTERNATIONLE DE NORMALISATION

ISO/IEC JTC1/SC2/WG2

Universal Multiple - Octet Coded Character Set

(UCS)

ISO/IEC JTC1/SC2/WG2 N 1862

Date: 1998-09-17

Title: Revision of WG2 N1711 (Mongolian)
Source: China
Action:
Distribution: WG2 members

After reviewing the document WG2 N1734 (Comments on Mongolian Encoding Proposal) prepared by Mr. Ken Whistler, and Mr. Richard Moore's comments on N1734, Chinese Mongolian experts revised its proposal WG2 N1711 as follows:

1. We agree to remove MONGOLIAN COMBINATION SYMBOL (position xx07). This character could be simply encoded in the General Punctuation Block, its name could be QUESTION EXCLAMATION MARK (position 2047), as it is suggested in WG2 N1734.
2. We agree to remove four MONGOLIAN POSITIONAL FORMAT CONTROL CHARACTERs (position xx1C..xx1F). The functions of these four characters can be realized by two already-encoded characters: ZERO WIDTH NON-JOINER (position 200C) and ZERO WIDTH JOINER (position 200D). If the Mongolian ad-hoc think the four MONGOLIAN POSITIONAL FORMAT CONTROL CHARACTERs better, we agree to preserve them as they are in WG2 N1711.
3. We agree to change the name of MONGOLIAN TODO SOFT HYPHEN (position xx08) to MONGOLIAN TODO HYPHEN as it is suggested in WG2 N1734.
4. We propose to preserve MONGOLIAN SPACE (position xx00) as it is described in WG2 N1711.
5. We propose to preserve the three MONGOLIAN FREE VARIANT SELECTOR CHARACTERs (position xx0D..xx0F), where to put these three characters is left for WG2 to decide.

6. We propose to preserve MONGOLIAN VOWEL SEPARATOR (position xx1B). We agree to change its name to MONGOLIAN VOWEL ZERO WIDTH NON-JOINER as it is suggested in WG2 N1734. Where to put this character is left for WG2 to decide.

These comments have been understood and accepted by Unicode Consortium at its UTC meeting in Redmond, July 1998.

For details, please refer to following documents:

1. WG2 1711 *The Working Meeting on Mongolian Encoding Attended by Representatives of China and Mongolia* (Mongolian encoding proposal)
2. WG2 N1734 *Comments on the Mongolian Encoding Proposal*
3. *Reply to "Proposal WG2 N1734" raised at the Seattle Meeting regarding "Proposal WG2 N1711"* (China) N1808
4. *Feedback on Ken Whistler's Comments on Mongolian Encoding: N1734* (Mongolia + UNU/IIST) N 1833
5. *Comments on Dr. Richard Moore's Feedback* (China)

Comments on Dr.Richard Moore's Feedback*
Dated May 4th,1998

After repeated discussion and correspondence in recent years, all sides concerned have reached a common understanding of most problems concerning Mongolian encoding, for which we feel delighted and express our heartfelt gratitude to all experts in various countries who have made earnest efforts in preparing the proposal for Mongolian encoding. As for the few problems that remain unsolved, we have illustrated different views on them in detail, understood each other's views as fully as possible, and are now approaching unanimity in methods with which to solve most of the problems. On such a basis, the Mongolian Encoding System ought to be finally and quickly settled and we sincerely hope that our Mongolian Encoding Proposal will be adopted at the WG2 Meeting in September, 1998.

To this end, in response to the feedback dated May 4th, 1998, from Mongolia and Dr. Richard Moore, we now put forth the following points for further discussion with you:

1. First of all, we are glad to see that your feedback gives consent to the final suggestions 1,2,4 and 6 of in the N 1734 Proposal. On our part, we agree to 1,2,4 and 6 the N 1734 proposal in their final forms.

2. Mongolian Space

We agree to your suggestion in the feedback ¡°We recommend the retention of the Mongolian space as a separate entity form the NBS.¡± However, we take exception to certain wording in the feedback:

(1) The feedback says, ¡°The letters immediately preceding the Mongolian space are final form variants whereas the letters immediately following it are middle form variants¡±. The actual fact in Mongolian, however, is that letters immediately following it are middle form variants¡±. The actual fact in Mongolian, however, is that letters immediately following Mongolian space may be middle form variants, or final form variants, or peculiar forms that belong to neither of these two. For example, UE in GER(MSP)UEN (of the house) is in its middle form; U in AMAN(MSP)U (of the mouth) is in its final form, and A in MAN(MSP)ACA(from us) is neither middle nor final but a peculiar form.

(2) The feedback says, ¡°Case endings separated from the main stem of the word.¡± In Mongolian, however, what is to be separated by the Mongolian space is not only case endings, but also the separated plurality endings and the endings of reflexiveª²possessive declension.

(3) The feedback also proposes to use NBS to mark (a) composite words, (b) syllable construction and (c) component construction. We admit that in Mongolian encoding, Mongolian space and NBS are used on different occasions. But at the same time, we object to using NBS in three totally different places as put forward in the feedback. Thus, in the old Mongolian script, UYILEDBURILEL-UN HUCUN (productive force) is a composite word and according to the view (1) in the feedback, is to be marked: UYILEDBURILEL-UN (NBS) HUCUN, which involves the collocational function of a word. The word UYILEDBURILEL-UN HUCUN should be syllabicated into UYIª²LEDª²BUª²RIª²LELª²UN HUª²CUN, which is to be marked according to the view (2) in the feedback:
UYI(NBS)LED(NBS)BU(NBS)RI(NBS)LEL(NBS)UN HU(NBS)CUN, which involves the phonetic structure of a word. Or, when the same word UYILEDBURILEL-UN HUCUN is divided into its morphemes UYILEª²Dª²BURIª²LEª²Lª²UN HUCUN, it should be marked, according to the view (3) in the
feedback, like: UYILE(NBS)D(NBS)BURI(NBS)LE(NBS)L(NBS)UN HUCUN, which involves the morphemic structure of a word.

What shall we do in case there is need to indicate all three kinds of information in the electronic dictionary? To give NBS three totally different functions of different levels and different meanings will involve unnecessary troubles in treating words and sentences.

Our opinion is:

(1) to retain the Mongolian space;

(2) to affirm that Mongolian space and NBS have different functions in Mongolian encoding. As for concrete uses for NBS, there seems to be no need to lay down any rigid rules.

3. Mongolian Positional Format Control Characters

In the feedback, the positional format controls of N 1711 and N 1734 are carefully compared and a conclusion is reached that the procedure of N 1734 is more complicated and requires more characters than the procedure of N 1711. On the surface, this conclusion seems to be wellª²founded and hence irreproachable. However, a mere analysis of the examples cited in the feedback will lead to reconsidering this conclusion. We all know that in Mongolian encoding, the control character should be used to indicate the position within a word in the following three cases:

(1) When there is need to show a separate representation variant form not within a word, e.g., the positional format control character for initial form should be used to show the initial form of the basic character A; and that for the middle form should be used to show the middle form of the basic character G.

(2) When there is need to split a connected word, e.g., the word SURGAGULI(school) is syllabicated into SUR GA GU LI with its form connected, we should add to the basic characters positional format control characters for the middle forms of R,G,A,G,U and L. If no positional format control characters are added to these letters, then we should show the final forms of R,A,U and I and the initial forms of G,G&L.

(3) In extremely exceptional case, where there is need to show representation form variants in a compulsory way not following regular wordª²form variation rules in any sequence, e.g., to show a middle or final form in its initial position; or to show the initial or final form in the middle position; or to show the initial or middle form in the final position. A position format control character is also required to indicate such irregular representation form variants.

Of the 36 examples listed on pp.2&3 of the feedback, 4 belong to the first kind, viz., -O-, -I-,-F-, and-M-, to show which N 1711 and N 1734 use nearly the same number of characters; 5 belong to the second kind, viz., -iF-,-iM-,-IF-,-M f-,-iMf-, to show which N 1711 and N 1734 use exactly the same number of characters. A comparison between these two kinds leads to no conclusion as to whether one kind is any better than the other. As many as 25 examples listed in the feedback belong to the third kind. Judging by a comparison between these 25 examples, N 1711 seems to economize much more characters than N 1734,a phenomenon which requires further analysis.

We must note that these three positional format control characters differ greatly in their uses. Firstly, uses of the first & the second kinds belong to the regular and correct spelling category; whereas uses of the third kind belong to the irregular and incorrect spelling category. Secondly, these three kinds of uses differ greatly, too, in their frequency of appearance. Therefore, in designing the kinds and numbers of positional format control characters, we must take into consideration their frequency of appearance in verbal material. Statistics shows that in verbal material of 1,000,000 words, the first kind appears 1622 times, or 0.162% of the total number of words; the second kind appears 181 times in 66 words, or 0.006% of the total number of words; whereas the third kind, which is extremely exceptional in ordinary Mongolian texts, appears at such a low frequency as to become almost negligible. It does not appear even once in the verbal material of one million words in our statistics, nor does it appear in verbal material of 5 million words. In other words, its frequency of appearance is lower than 1/5 million. So no matter how many examples are given for it, this third kind cannot serve as basis for comparing the different positional format control characters. And in designing positional format control characters, we must not attach equal importance to all three kinds; instead, we should take into prior consideration the first and the second kinds that belong to the regular and correct spelling categories, yet not ignoring the third kind. In a word, the explanation of the procedure of the third kind, which is a matter of little account, is a bit too trivial. At present, views vary most greatly of all in Mongolian positional format control characters, especially in the third kind which appears at a very low frequency, for we do not differ much in the first and the second kinds that belong to the regular and correct spelling categories.

Our opinions are like this:

(1) Both N 1711 and N 1734 are feasible with only slight difference in the treatment of details of little importance. It is quite indifferent to us what method is eventually adopted.

(2) The final settlement of Mongolian encoding must by no means postponed for five years on the ground that we still differ in views on such a problem. If that should happen, the loss would greatly overweigh the gain, especially for China where a Mongolian population of more than 4 million are using the old Mongolian script only.

4. Mongolian Vowel Separator:

First of all, we agree to the view expressed in the feedback, ¡°The proposal to use the nonª²joiner in place of the proposed Mongolian vowel separator...¡± Views in the feedback can be classified under three groups: (1) The Mongolian vowel separator is no longer needed where the positional format control characters and the variant selector are used; (2) The Mongolian vowel separator is not necessary either only if we can determine the form of the consonant preceding the final separate vowel, for then we can solve the problem with algorithm; (3) Since the final separate form of the vowel is the most commonly used variant, so the variant selector may not be used.

Our views are as follows:

(1) We may use the positional format control character and the variant selector, or the joiner and the nonª²joiner to indicate a separate vowel and the consonant preceding it. But the number of characters used is different: For example:

Ken's method: some letters + ML.NA + NJ = ML.A + FVS2 (4 characters being used)

Moore's method: some letters + ML.NA + MEF + ML.A + FIF + FVS2 (5 characters used)
or: some letters +ML.NA + MEF + ML.A + FVS2 (4 characters used)

1711 method: some letters + ML.NA + MVS + ML.A (3 characters used)

(2) We can also mark the separate vowel and the consonant preceding it with the algorithm provided we determine the form of the consonant preceding the final separate vowel. For example:

some letters + ML.NA + ML.A + FIF + FVS2 or:

some letters + ML.NA + ML.A + FVS2

(3) The feedback says that since the most commonly used variant is the final separate vowel form. so the variant selector may not be used. We have figured out the frequency of appearance of the separate variant and connected variant 1 of the final vowel A/E based on the old Mongolian script corpus of one million words and the result shows:

Separate variants of the vowel A/E appear (after the 9 consonants N. H. G. M. L. J. Y. R & W) 12195 times, or 1.15%

Connected variants 1 of the vowel A/E (after the 9 consonants N, H, G, M, L, J, Y, R & W) appear 3652 times, or 0.34%

Connected variant 1 of the vowel A/E appear (after the 10 consonants S, SH, T, D, CH, TS, Z, H, ZR, LH) 5756 times, or 0.54%

Connected variant 1 of the vowel A/E appear altogether 9408 times, or 0.88%

We could take the separate variant as the variant 1 and do without the variant selector provided we consider only the frequency of appearance of the separate A/E and connected variant 1 following the 9 consonants N, H, G, M, L, J, Y, R & W in the old Mongolian script. However, in the old Mongolian script, the connected variant 1 of the vowel A/E appears not only after the 9 consonants N, H, G, M, L, J, Y, R & W, but also after the 10 consonants S, SH, T, D, CH, TS, Z, H, ZR, LH. Apart from this, there is the matter which we cannot but take into consideration, i.e., the connected variant 1 of the vowel A also appears after some 40 consonants in Todo (Q,G,M,T,D,CH,J,TS,Y,W,H,JI& NI), Sibe (K, G, H, SH, T, D, J, TS, Z, RA, CH & ZH), Manchu (K, R, ZH) and Manchu AG (GH, NG, C, JH, TT, DDH, T, DH, SS, CY, ZH & Z) scripts, while in the old Mongolian script, the separate variant of the vowel A/E appears only after those particular consonants. It is inappropriate to regard as final A/E variant 1 the separate variant of the vowel A/E that appears after 9 consonants in a single language and at the same time to regard as the final A/E variant 2 the connected variant of the A/E vowel that appears after 59 consonants belonging to 5 languages. Besides, there is still the tradition of converting the old Mongolian script into Latin letters among Mongolian scholars. In that tradition, the separate variant of the vowel A/E in a Mongolian text is transcribed as ¡°-A/-E¡±, whereas the connected variant of the vowel A/E is transcribed as ¡°A/E¡± without any bar preceding it. Thus:

Some letters + ML.NA + separate A¡± is transcribed as ¡°some letters + N_A/ N-A¡±

Some letters + ML.NA + connected A¡± is transcribed as ¡°some letters + NA¡±

Some letters + ML.YA + separate E¡± is transcribed as ¡°some letters + some letters + Y_E/ Y-E¡±

Some letters + ML.YA + connected E¡± is transcribed as ¡°some letters + YE¡±

Therefore, if we make a regulation that the store form of the separate variant of the vowel A/E be A/E with a certain control character preceding it, then it will not only distinguish itself from the connected variant, but also link up with Latinisation, a tradition for the Mongolian scholars.

Our opinion is to retain the Mongolian vowel separator.

We sincerely hope that experts in various countries will return to us their valuable views on our comments in good time so that we can bring to completion our work on Mongolian encoding as soon as possible.

Chinese Delegation

July 7, 1998