ISO

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION

ORGANISATION INTERNATIONALE DE NORMALISATION

---

ISO/IEC JTC1/SC2/WG2

Universal Multiple-Octet Coded Character Set (UCS)

---

ISO/IEC JTC1/SC2/WG2 *N 1833*

**Date:** 1998-08-31

| | |
|---|---|
| Title | Feedback on Ken Whistler's comments (N1734) on Mongolian |
| Source | Richard Moore, United Nations University |
| Date | 4 May 1998 |
| Reference | ISO/IEC JTC1/SC2/WG2 N1734 |

**Dear All,**

**Below are comments, agreed between Mongolia and UNU/IIST, on the questions Ken Whistler raises in his document (WG2 N 1734) regarding the Mongolian encoding proposal N 1711. We apologise that they were not ready in time to provide feedback into the Chinese response which Mike distributed last week.**

**Please note that these comments relate only to document N 1734. We do not specifically comment on N 1711 itself because we have not finished studying it completely yet. We will send our comments on that at a later date.**

**With best regards,**

**richard**

**********

**Feedback on Ken Whistler's Comments on Mongolian Encoding: N1734**

**Mongolia + UNU/IIST**

**********

## 1. Mongolian Space

In Mongolian script, words are often written with case endings separated from the main stem of the word. Further, one stem may have several case endings following it, in which case each separate case ending is written separated from the others. Thus, the form of a single "word" can be:

stem caseA caseB caseC

where the spaces actually appear as white spaces when the word is printed or displayed on a computer screen.

Traditionally, this space separating case endings is called the Mongolian space, and it differs from the normal space mainly in that the letters immediately preceding the Mongolian space are final form variants whereas the letters immediately following it are middle form variants. In addition, the Mongolian space is generally smaller than the normal space (typically one third of the size) and a line of text should not be broken at a Mongolian space.

Many arguments have been put forward relating to the necessity or otherwise of including Mongolian space as a separate character, on the one hand claiming that it is fundamentally different from the existing character NBS (no-break-space) and on the other hand claiming exactly the opposite. However, we do not feel that any of these arguments is particularly convincing one way or the other.

We do tend to agree that much of the functionality of the Mongolian space is either already present in NBS, or could be specifically incorporated into a "Mongolian interpretation of NBS" as Ken Whistler seems to suggest.

However, we can envisage two scenarios in which the NBS might be used in Mongolian which would distinguish it from the Mongolian space.

First, the Mongolian language contains a very large number of "composite words", where a series of words taken together represents a single concept, and the NBS could be used to logically "join" these composite words into a single unit, for example for electronic analysis or searching of documents. In such a use, the space between the elements of a composite word would not only be a normal sized space but it would also have a semantically different meaning from the space linking case endings.

Second, the NBS could be used, e.g. in educational texts, as a separator to show how a word is constructed from syllables or to show how a derivative word is constructed from its components. Admittedly this could also be done using the format control characters and the variant selectors, but these would be much less efficient in this case.

In view of these scenarios, which would be impossible if the Mongolian space were unified with the NBS, we recommend the retention of the Mongolian space as a separate entity from the NBS.

## 2. Mongolian Combination Symbol

We agree that this character should be retained. We do not care what it is called! We are happy for it to be included in the General Punctuation block instead of in the Mongolian section.

## 3. Mongolian Positional Format Control Characters

We accept that the different positional variant forms could be indicated using the existing zero-width joiner and non-joiner characters instead of using specific positional form selectors as proposed in N1711 (and previous proposals).

However, the system based on the joiner and non-joiner requires not only more complicated input and output algorithms than that using the positional form selectors, but also on average significantly longer code strings to generate the equivalent sequence of actual characters. A comparison between the two coding schemes, based on Ken Whistler's table, is given in the following table supplied by Mongolia:

//**********

| DISPLAY | STORE | store (according N1711) |
|---|---|---|
| _O_ | _B_ | _B ISF_ |
| _I_ | _B J_ | _B INF_ |
| _F_ | _J B_ | _B FIF_ |
| _M_ | _J B J_ | _B MEF_ |

The same number of codes is used in two columns.

| | | |
|---|---|---|
| _iO_ | _b J NJ B_ | _b B ISF_ |
| _iI_ | _b J NJ B J_ | _b B INF_ |
| _iF_ | _b B_ | _b B _ |
| _iM_ | _b B J_ | _b B MEF_ |

In Mongolian Script, there is no difference for 'i' between 'iO' and 'iF', but they have to insert J in the iO string to distinguish it from oO. Therefore the difference of numbers of codes, in the two proposals, is -3.

| | | |
|---|---|---|
| _oO_ | _b NJ B_ | _b ISF B ISF_ |
| _oI_ | _b NJ B J_ | _b ISF B INF_ |
| _oF_ | _b NJ J B_ | _b ISF B _ |
| _oM_ | _b NJ J B J_ | _b ISF B MEF_ |

The difference of numbers of codes is -1.

| | | |
|---|---|---|
| _Of_ | _B NJ J b_ | _B ISF b_ |
| _If_ | _B b_ | _B b_ |
| _Ff_ | _J B NJ J b_ | _B FIF b_ |
| _Mf_ | _J B b_ | _B MEF b_ |

There is also no difference for 'f' between 'Of' and 'If', so the difference of numbers of codes is -3.

| | | |
|---|---|---|
| _Oo_ | _B NJ b_ | _B ISF b ISF_ |
| _Io_ | _B J NJ b_ | _B b ISF_ |
| _Fo_ | _J B NJ b_ | _B FIF b ISF_ |
| _Mo_ | _J B J NJ b_ | _B MEF b ISF_ |

The difference of numbers is -1.

| | | |
|---|---|---|
| _iOf_ | _b J NJ B NJ J b_ | _b B ISF b_ |
| _iIf_ | _b J NJ B b_ | _b B INF b_ |
| _iFf_ | _b B NJ J b_ | _b B FIF b_ |
| _iMf_ | _b B b_ | _b B b_ |

The difference is -5.

| | | |
|---|---|---|
| _oOf_ | _b NJ B NJ J b_ | _b ISF B ISF b_ |
| _oIf_ | _b NJ J B b_ | _b ISF B INF b_ |
| _oFf_ | _b NJ J B NJ J b_ | _b ISF B FIF b_ |
| _oMf_ | _b NJ J B b_ | _b ISF B b_ |

The difference is -4.

| | | |
|---|---|---|
| _iOo_ | _b J NJ B NJ b_ | _b B ISF b ISF_ |
| _iIo_ | _b J NJ B J NJ b_ | _b B INF b ISF_ |
| _iFo_ | _b B NJ b_ | _b B FIF b ISF_ |
| _iMo_ | _b B J NJ b_ | _b B b ISF_ |

The difference is -3.

| | | |
|---|---|---|
| _oOo_ | _b NJ B NJ b_ | _b ISF B ISF b ISF_ |
| _oIo_ | _b NJ B J NJ b_ | _b ISF B INF b ISF_ |
| _oFo_ | _b NJ J B NJ b_ | _b ISF B FIF b ISF_ |
| _oMo_ | _b NJ J B J NJ b_ | _b ISF B b ISF_ |

The difference is -1.

The total difference is -17 codes in this part, for example.
//**********

This latter point implies that documents would require significantly greater storage space and would take significantly longer to transmit electronically. This is of particular concern to Mongolia because the level of computing and communications technology available to normal users is relatively low.

In view of this, we would prefer to retain the positional format control characters despite the fact that they provide functionality which can be mimicked by the joiner and non-joiner because we feel that they provide this functionality in a much more efficient and logical way.

We would further suggest that, since it is likely that a number of Arabic speaking countries suffer the same lack of state-of-the-art technology as Mongolia, these positional format control characters would additionally offer a more efficient and logical alternative for coding variant forms in Arabic which could similarly benefit these countries.

With regards to the Positional Indicator Character (xx1C in document N1691):

In document N1691 (and various predecessors) this character, or ones like it, were included in the proposals as a suggested means of generating positional forms (isolated, initial, medial, final) of characters. But as we have pointed out a number of times, beginning with document N1497 which we submitted to and which was discussed at the Singapore WG2 meeting in January 1997, the use of his (and similar) character(s) in these proposals is logically flawed because strings containing it are ambiguous.

More specifically, in N1691 it is stated that:

| | |
|---|---|
| (PIC)X | means X is final form |
| X(PIC) | means X is initial form |
| (PIC)X(PIC) | means X is middle form |

With this scheme, the string

AB(PIC)C(PIC)

has two possible interpretations:

1) B and C are both initial forms
2) C is middle form

and there is no way of distinguishing these alternatives.

This character thus appears to serve no useful purpose (its intended functionality now being provided correctly by the positional format control characters and/or by the joiner/non-joiner) and is logically unsound. We therefore repeat our recommendation that it should be removed.

4. Mongolian Free Variant Selector Characters
---

Since the maximum number of possible variants of any single positional form appears to be four, three free variant selectors are both necessary and sufficient.

We have no preference regarding whether they are considered as Mongolian "characters" or as something more general.

5. Mongolian Vowel Separator
---

The proposal to use the non-joiner in place of the proposed Mongolian vowel separator, as in the example some letters + ML.NA + NJ + ML.A + FVS2 does not work if the non-joiner is also used to distinguish positional form: the above string would give the final form of ML.NA but the second variant **isolated** form of ML.A. (No, there isn't one! We assume that in this case you'd just get the default variant of the isolated form.)

However, the Mongolian Vowel Separator is in any case entirely redundant -- the separated final forms of the ML.A and ML.E characters are available in the character set as variants, so the required string can be generated using only the positional format characters and the variant selectors (We guess this is what Ken meant, but he just got the details slightly mixed).

Actually, one could perhaps go further.

The letter preceding the separated vowel form is always final form or middle form, and this form is determined by the actual letter (i.e. it is not a matter of choice). So this could perhaps be incorporated into the rules for calculating the default form of a character: e.g. a letter defaults to final form if 1) it is followed by a separator or 2) it is followed by a separated vowel and is one of some particular set of letters (i.e. the ones which are final form not middle form before a separated final vowel) or ....

Further, we believe that the separated form is actually the most commonly used final form variant, in which case this should perhaps be the default final form, thereby removing the necessity to use the variant selector FVS2 to obtain the separated form.

## 6. Mongolian Todo Soft Hyphen

We are not sufficiently familiar with the Todo script to offer any comments on this issue.