L2/98-268R

Title: Analysis and UTC Position Regarding Mongolian Encoding Issues
Source: Ken Whistler and the Unicode Technical Committee
Date: July 30, 1998
Status: Expert Contribution and UTC Position Paper
Distribution: UTC and L2 Members
References: WG2 N 1711, L2/98-104, L2/98-251, L2/98-252, L2/98-272

Note: The Introduction and Summary sections of this document represent analysis of the Mongolian encoding issues, presented as an expert contribution by Ken Whistler. The second part of this document, labeled "UTC Positions" represents the consensus position of the UTC regarding these issues, as agreed to by the UTC after discussion of the various Mongolian documents.

Introduction

In L2/98-104 I raised a number of issues regarding the latest full proposal for encoding Mongolian, WG2 N 1711.

L2/98-251 is the response to those questions by the Chinese national body, authored by Prof. Choijinzhab. L2/98-252 is the response to those questions from Mongolia and UNU/IIST, authored by Richard Moore; L2/98-252 also provides feedback on the Chinese position. L2/98-272 is the Chinese response to the issues raised by Richard Moore in L2/98-252.

In this position paper, I provide a brief summary of what I understand to be the Chinese and Mongolian stands on the controversial issues. I note where there seems to be consensus and where there is a difference of opinion. To keep things straight, I refer to the issues by the same numbers (1-6) that I used in L2/98-104, since China also used those numbers for reference. Then, for discussion purposes, I suggest what I think would be the best UTC position to take on these issues (and on the overall Mongolian encoding proposal), given this feedback from China and from Mongolia.

Summary of Chinese and Mongolian Responses

1. Mongolian Space

L2/98-251 presented evidence that the Mongolian space could contrast in usage in Mongolian with a NBSP. From the way NBSP is discussed in L2/98-251, I think Prof. Choijinzhab may have NBSP mixed up with ZWSP, however.

L2/98-252 presented other examples, which sound valid to me, where Mongolian space would contrast with NBSP in Mongolian text.

L2/98-272 argues against designation of particular usage of NBSP in Mongolian, but concludes still that NBSP could be used in Mongolian and is different from Mongolian space.

Summary: Both China and Mongolia want to retain a distinct character for Mongolian space, not unified with the NBSP.

2. ?! Punctuation Character

L2/98-251 concurs with L2/98-104, that this character can be encoded at, say U+2047, in the punctuation block, rather than as a specifically Mongolian punctuation mark.

L2/98-252 concurs with the Chinese position.

3. Positional Format Controls

L2/98-251 agrees that the analysis I presented in L2/98-104 is valid, and that the already encoded ZWJ and ZWNJ can be used instead of the 4 positional format controls suggested in WG2 N 1711. So China was agreeing that those 4 characters were not needed in the Mongolian proposal.

L2/98-252 agrees that the ZWJ and ZWNJ could be used instead, but does a reanalysis of the possible cases to show the number of characters required for each, and argues that the four positional variant controls of WG2 N 1711 would accomplish the task more "efficiently and elegantly".

L2/98-272 presents statistical evidence from a large Mongolian corpus to show that the only two kinds of cases that have any significant occurrence in text (showing a single positional form in isolation, or showing syllabification of Mongolian words with spaces between forms) do not require any difference in total number of characters required, under either proposal. It further claims that all the other theoretically possible positional combinations, while of conceivable use, do not show up even once in the large (multi-million word) corpus. Thus it states, correctly, in my opinion, that number of required characters for the joiner/non-joiner analysis versus the 4 positional format controls analysis is not criterial for practical use in real text. China restates its willingness to make use of the ZWJ and ZWNJ.

Summary: China concurs with the analysis of L2/98-104 and agrees to use ZWJ and ZWNJ instead of 4 new Mongolian-specific positional format controls. Mongolia admits the ZWJ and ZWNJ analysis would work, but restates a preference for the positional format controls.

4. Free Variant Selectors

L2/98-251 provides some justification and restates the need for 3 of these for Mongolian. China expresses no preference for whether these are Mongolian-specific or are encoded outside of the Mongolian block for general usage.

L2/98-252 concurs with China and restates the need for 3 free variant selectors.

5. Mongolian Vowel Separator

L2/98-251 argues the case for retaining this character, stating that it represents a more parsimonious encoding of a rather commonly occurring situation in Mongolian text. It points out that the MVS also maps directly onto existing expectations about keyboard input practice, and that it directly reflects Latin transcriptional practice for Mongolian as well. China states that it doesn't care what the character is named, and agrees that MONGOLIAN ZERO WIDTH NON-JOINER would serve just as well.

L2/98-252 suggests that the MVS is redundant, and that its effects could be predicted by special-casing the rules for positional variant formation by specifying the set of consonants after which the -a/-e vowel takes the separated form (consonant before takes a final form, followed by a short gap and then the final form of the -a/-e).

L2/98-272 presents a fairly complex counterargument to the Mongolian position on this, pointing out that the default for a specified set of consonants in Mongolian does not apply to that same set of consonants in the other 4 languages written with the Mongolian script. Rather than make the special, gapped forms be the default rule-governed behavior and then have to override it in the other languages, L2/98-272 argues that the special behavior in Mongolian proper can best be handled by introducing the MVS, which additionally matches input and transliterational practice.

Summary: China argues to retain the MVS, while not caring about its name. Mongolia argues that it is redundant, and can be predicted by rule.

6. Todo Soft Hyphen

L2/98-251 presents a little more evidence of usage of this character, argues that it is still needed, but agrees that it could be called TODO HYPHEN instead.

L2/98-252 abstains on this issue, for lack of expertise on Todo.

==========

UTC Positions on Mongolian Encoding

Except when otherwise noted, these positions are organized by the numbers used in document WG2 N 1734 (= L2/98-104), and are stated in terms of agreement or disagreement with the opinions expressed by the Chinese national body in documents L2/98-251 and L2/98-272 or by Mongolia and UNU/IIST expressed in document L2/98-252, authored by Richard Moore.

1. Mongolian Space

The UTC accepts the Chinese and Mongolian requirements for encoding a separate Mongolian space. There is a reasonable case for the common usage of a non-breaking space of Mongolian-specific layout width that can be used for Mongolian-specific (common) purposes and which could meaningfully contrast with a regular NBSP used in the same text.

The UTC suggests that this character be named NARROW NO BREAK SPACE and that it be encoded in the General Punctuation block at U+202F. The concept of the Mongolian space (a non-breaking space, narrower than a normal non-breaking space, and contrasting with it in usage)

could be of use in other scripts as well; therefore it is better to make this a general use punctuation character, rather than limiting it to the Mongolian script.

2. ?! Punctuation Character

The UTC agrees with the Chinese and Mongolian positions. The xx07 MONGOLIAN COMBINATION SYMBOL ("?!") should be encoded in the General Punctuation block, perhaps at U+2047 or U+2048. The preferred name for this symbol would be QUESTION EXCLAMATION MARK, since it is of general applicability and would not be limited to Mongolian usage only.

3. Positional Format Controls

The UTC agrees strongly with the Chinese position regarding positional format controls. China has concurred with the analysis that use of the already encoded ZWJ and ZWNJ, in the same way they are used for positional format control in the Arabic and Syriac scripts, would meet the requirements for encoding such behavior in the Mongolian script. The Chinese argument regarding the lack of a need for the alternative proposal involving 4 new positional format controls is cogent.

Accordingly, the UTC urges that the 4 Mongolian-specific positional format controls shown at xx1C..xx1F in document WG2 N 1711 be removed from the proposal before the preparation of a PDAM for ballotting on the Mongolian encoding.

4. Free Variant Selectors

Both China and Mongolia have reiterated the need for 3 free variant selector characters. The UTC notes this consensus requirement, agrees with the Chinese and Mongolian positions, and has no objection to the retention of these 3 free variant selector characters in the proposed Mongolian encoding, as shown in document WG2 N 1711.

The documentation for these characters should clearly state that the free variant selector characters should follow the basic letter they show variation for. This is to ensure compatibility with other types of variant marks which may be encoded for other scripts in the future.

5. Mongolian Vowel Separator

The UTC agrees with the Chinese position regarding the Mongolian vowel separator. The case is convincing that the MVS would simplify mapping of existing input practice to the encoding and also accords with transcriptional practice for Mongolian. UTC supports the encoding of the MVS. Contra the suggestion in L2/98-104, the UTC sees no particular advantage in changing the name from MONGOLIAN VOWEL SEPARATOR to MONGOLIAN ZERO WIDTH NON-JOINER. The name MONGOLIAN VOWEL SEPARATOR, as shown in document WG2 N 1711 should be retained.

The documentation of this character should clearly detail the effect it has on the positional form selection of the consonant it follows and the vowel it precedes in Mongolian proper. Also,

differences between Mongolian and the other languages that use the Mongolian script should be detailed as a guide to the proper use (or non-use) of this character for particular languages.

6. Todo Soft Hyphen

The UTC accepts the requirement for this character, as expressed by the Chinese national body.

The UTC would like to see more evidence regarding the positions where the Todo Soft Hyphen can occur. If, as suggested in L2/98-251, the *only* position that this character can occur is at the beginning of a line (to mark a work broken with some syllables on each line), then TODO SOFT HYPHEN is probably the best name for it. If this hyphen can ever be displayed internally in a line, then TODO HYPHEN is the better name.

Overall recommendation

The UTC provisionally accepts the repertoire, ordering, and suggested names for the Mongolian script from document WG2 N 1711, with the exceptions as noted above regarding issues 1-6.