MWG/3-N12

L2/19-134

Title: Summary of Mongolian updates in the Unicode Standard since MWG/2
Author: Roozbeh Pournader (Internationalization Engineer, WhatsApp/Facebook)
Date: March 28, 2019

Introduction

In April 2018, experts from China, Japan, Mongolia, and the United Kingdom, and the United States met in San Jose, California to discuss various open issues regarding the traditional Mongolian script. A report from the meeting by Lisa Moore is available at https://www.unicode.org/L2/L2018/18108-mwg2-14-meeting-report.pdf. A summary of suggestions and proposals for improvements, edited by the author, is available at https://www.unicode.org/L2/L2018/18107-mwg2-13-summary-improvement.pdf.

Based on the various recommendations and discussions in that meeting (MWG/2), and documents submitted to the Unicode Technical Committee (UTC) for that meeting and since the meeting, as well discussions in the script ad hoc meetings and the UTC meetings, various improvements have been added to the Unicode Standard. Other improvements have been planned. This document lists these improvements.

Clarifications in the Unicode Core Specifications regarding NNBSP

The text of the Unicode Core Specification has been updated to reflect the nuances of how NNBSP works for Mongolian script. Here is the latest text (also available at https://www.unicode.org/versions/Unicode12.0.0/ch13.pdf):

> Because separated suffixes are usually considered an integral part of the word as a whole, a line break opportunity does not normally occur before a separated suffix. The whitespace preceding the suffix is often narrower than an ordinary space, although the width may expand during justification. U+202F NARROW NO-BREAK SPACE (NNBSP) is used to represent this small whitespace; it not only prevents word breaking and line breaking, but also triggers special shaping for the following separated suffix. The resulting shape depends on the particular separated suffix. Note that NNBSP may be preceded by another separated suffix, and NNBSP may also appear between non-Mongolian characters and a separated suffix.
>
> Normally, NNBSP does not provide a line breaking opportunity. However, in situations where a line is broken before a separated suffix, such as in narrow columns, it is important not to disable the special shaping triggered by NNBSP. This behavior may be achieved by placing the break so that the NNBSP character is at the start of the new line. At the beginning of the line, the NNBSP should affect only the shaping of the following Mongolian characters, and should display with no advance width.

Clarification in Line Breaking Algorithm

To match the above changes, the following text has been added to "Unicode Standard Annex #14, Unicode Line Breaking Algorithm":

> NARROW NO-BREAK SPACE has exactly the same line breaking behavior as NO-BREAK SPACE, but with a narrow display width. The MONGOLIAN VOWEL SEPARATOR acts like a NARROW NO-BREAK SPACE in its line breaking behavior. Both of these characters are regularly used in Mongolian text, where they participate in special shaping behavior, as described in *Section 13.5, Mongolian* of [The Unicode Standard, Core Specification].

Property change for NNBSP

To help improve the support for Mongolian script's use of NNBSP, the Script_Extensions property of NNBSP has been expanded to explicitly include Mongolian (see https://www.unicode.org/Public/12.0.0/ucd/ScriptExtensions.txt):

```
# Script_Extensions=Latn Mong

202F          ; Latn Mong # Zs       NARROW NO-BREAK SPACE

# Total code points: 1
```

This would help Unicode implementations that use the Script_Extensions property in script itemization keep NNBSP in the same script run as the Mongolian text that follows and/or precedes it.

Listing the right letters to use in the Core Specification

The list of writing systems using the Mongolian script, as well as terminology to describe it, has been clarified, and a table has been added to guide users of the Unicode Standard on the right characters to use for each writing system. Here is the latest text (also available at https://www.unicode.org/versions/Unicode12.0.0/ch13.pdf):

> The Mongolian block unifies the traditional writing system for the Mongolian language and the three derivative writing systems Todo, Manchu, and Sibe. The traditional writing system is also known as "Hudum Mongolian," and is explicitly referred to as "Hudum" in the following text. Each of the three derivative writing systems shares some common letters with Hudum, and these letters are encoded only once. Each derivative writing system also has a number of modified letterforms or new letters, which are encoded separately. The letters typically required by each writing system's modern usage are encoded as shown in *Table 13-4*.

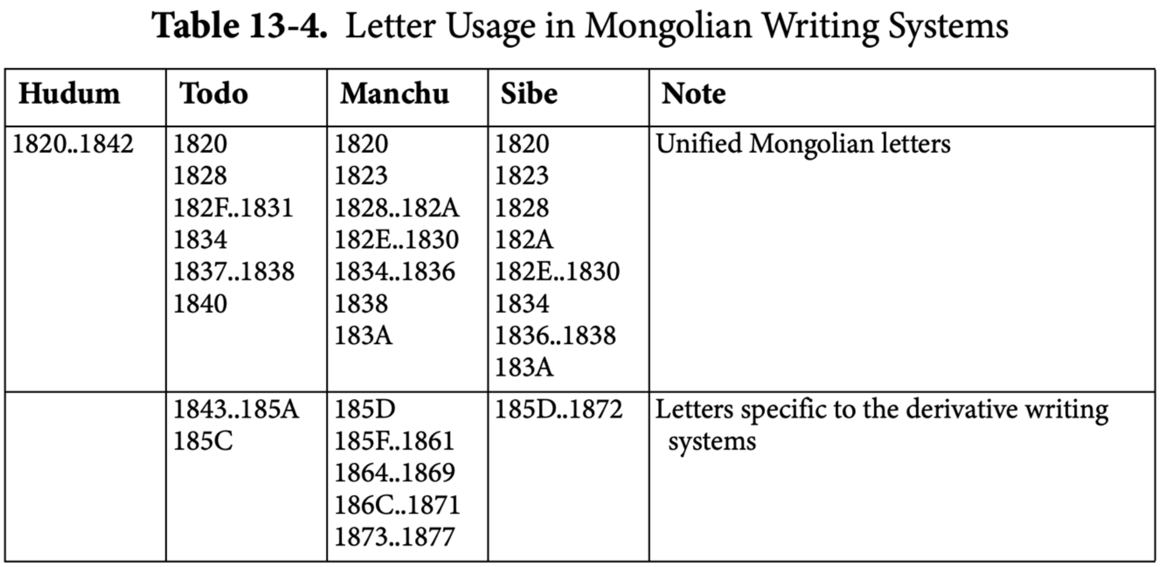

Table 13-4. Letter Usage in Mongolian Writing Systems

Hudum	Todo	Manchu	Sibe	Note
1820..1842	1820 1828 182F..1831 1834 1837..1838 1840	1820 1823 1828..182A 182E..1830 1834..1836 1838 183A	1820 1823 1828 182A 182E..1830 1834 1836..1838 183A	Unified Mongolian letters
	1843..185A 185C	185D 185F..1861 1864..1869 186C..1871 1873..1877	185D..1872	Letters specific to the derivative writing systems

Clarification on use of Mongolian soft hyphen

The Core Specification has been updated to mention that U+1806 MONGOLIAN TODO SOFT HYPHEN is also used in Mongolian language and is not restricted to Todo. The latest text reads:

> In writing Mongolian and Todo, U+1806 MONGOLIAN TODO SOFT HYPHEN is used at the beginning of the second line to indicate resumption of a broken word. It functions like U+2010 HYPHEN, except that U+1806 appears at the beginning of a line rather than at the end.

Exclusion from domain names

MWG/2 recommended that the present encoding of traditional Mongolian script is not suitable for use in internationalized domain names (IDN). UTC accepted that recommendation in UTC meeting #155, in May 2018, and modified “Unicode Standard Annex #31, Unicode Identifier and Pattern Syntax”, to add the Mongolian script to its Table 4, Candidate Characters for Exclusion from Identifiers. It also added language specific to Mongolian to UAX #31, explaining the reasoning for that exclusion: “Some scripts also have unresolved architectural issues that make them currently unsuitable for identifiers.” The related data in “Unicode Technical Standard #39, Unicode Security Mechanisms”, has also been updated to reflect this information.

This change helps Unicode implementations to avoid using the Mongolian script in domain names and similar applications where the confusability of the present encoding could lead to security attacks. ICANN has also been notified of this change.

If the encoding model is changed or extended in a way that the security concerns no longer exist, the Mongolian script can be moved back to the group of scripts that could be considered for domain names.

Moving the script specification from the code chart to separate document

The Mongolian code chart is not the best vehicle to document the complexities of the traditional Mongolian script. Its format is quite insufficient and has caused confusion for many of its users. At the same time, the complementary specification for the Mongolian script in the Unicode Core Specification needs much more details than currently provided. Both the code chart and the Core Specification also are hard to develop, since any changes to them require the approval of the both the UTC and ISO/IEC JTC1/SC2, which is time consuming.

In that light, UTC attendees have recommended reducing the reliance on the Mongolian code chart to specify the nuances of the Mongolian script, and instead creating a Unicode Technical Note (UTN) to describe the Mongolian script much more clearly and thoroughly. Early drafts of the UTN are already under development, and once ready, the Mongolian code chart will be simplified and a pointer to the UTN will be added.

After stabilization, the Mongolian UTN can advance to become a Unicode Technical Report (UTR), and later a Unicode Technical Standard (UTS) or a Unicode Standard Annex (UAX). This would result in Unicode implementations for the Mongolian script that match user requirements better and are more interoperable.