TO: UTC

FROM: Biligsaikhan Batjargal et al. (forwarded by Deborah Anderson, Script Encoding Initiative, UC Berkeley)

RE: Mongolian Script Rendering Issues

DATE: 30 July 2010

The attached paper is a draft and reflects current research by the authors. NOTE: It is not to be distributed outside the UTC/L2 because it is intended for eventual publication. However, the authors welcome comments and questions. Please address feedback and questions to: biligsaikhan@gmail.com.

It is hoped that rendering issues can be resolved soon, particularly in light of the recent announcement by the President of Mongolia that requires the use of traditional Mongolian script (alongside Cyrillic) on certain official documents. A link to the announcement (in Cyrillic) is: http://www.president.mn/mongolian/node/903.

An abbreviated, unofficial English translation is provided below (from http://www.montsame.mn/index.php?option=com_news&mt=normal_news&tab=201007&task=news_detail&ne=264):

Мэдээ Мэдээ » English news » Politics

DECREE ISSUED TO PROMOTE MONGOLIAN SCRIPT

2010-07-06 20:12:59 | | Хэвлэх | Найздаа илгээх |

Ulaanbaatar, Mongolia, /MONTSAME/ The President Ts.Elbegdorj issued Tuesday a decree to augment the official usage of the Mongolian script.

In accordance with the decree, issued welcoming the 100th anniversary of the Revolution for National Freedom, official documents and letters addressed to State Head, Speaker, Prime Minister, cabinet members and other dignitaries of foreign countries must be written in the Mongolian script. To these documents are to be attached their translations into any UN official language or a foreign language of that country. ID cards, marriage certificates, educational certificates and diplomas muist be written in Cyrillic alphabet and Mongolian script both.

The second national program on Mongolian script was approved in 2008. The government was ordered to pay an attention to intensifying an implementation of the program and to conclude it within 2011.

The President's decree will come into force from July 11, 2011.

B.Khuder
19.35

A Survey on Rendering Traditional Mongolian Script

Biligsaikhan Batjargal[†], Fuminori Kimura[‡] and Akira Maeda[‡]

[†]Graduate School of Science and Engineering, Ritsumeikan University
[‡]College of Information Science and Engineering, Ritsumeikan University
1-1-1 Noji-Higashi, Kusatsu, Shiga 525-8577, Japan
biligsaikhan@gmail.com, {fkimura, amaeda} @is.ritsumei.ac.jp

Abstract. This paper discusses the rendering issues of complex text layouts, particularly traditional Mongolian script. Solving the rendering issues of complex text layouts is the key fundamental challenge to succeed in the future development of digital libraries. Recently some standards such as Unicode and OpenType format have been implemented and supported widely. Furthermore, traditional Mongolian script has been standardized in Unicode. In this paper, we analyzed existing OpenType fonts and their rendering schemes for traditional Mongolian script. We found some errors, and discovered grammatical rules, which are not documented in international standards. None of the existing OpenType fonts was complete. Lastly, this paper provides some improvements and recommendations for future development.

Keywords: Traditional Mongolian Script, Unicode, Encoding, OpenType, Complex Text Rendering, Digital Library

1   Introduction

In the past decade, much has been initiated in digital library research and development in the Asia-Pacific region, and the importance of digital library systems that preserve cultural heritage have also increased. Internationalization and digital cultural heritage preservation in the Asia-Pacific region require digital library systems to support various ancient or modern complex text scripts. Also writing systems of Asia-Pacific that requires complex transformations between text input and text display for proper rendering are needed. For these complex text scripts the way text is stored is not mapped to the way it is displayed in a straightforward fashion like western scripts. Examples of such writing systems are the Arabic, traditional Mongolian and Brahmic (Indic) family such as Devanagari or Dravidian scripts or the Thai alphabet. Rendering complexities of some scripts is well documented and surveyed, though traditional Mongolian script is not.

The traditional Mongolian script digital library (TMSDL)[1] [1][2][3] was developed to preserve over 800 years of old historical records written in traditional Mongolian script. One of the challenges for the TMSDL is to display and render documents written in traditional Mongolian script correctly. Once, due to poor support for traditional Mongolian script at the operating system level, Garmaabazar et al.[3]

---

[1] http://www.dl.is.ritsumei.ac.jp/tmsdl/

developed a conversion algorithm to display contents in traditional Mongolian script. However, recently in Windows Vista and in later versions, especially in Windows 7, support for traditional Mongolian script and the input locale has been added. Microsoft Windows multilingual text rendering engine –Uniscribe (Unicode script processor) – renders OpenType fonts that can handle the diverse behaviors of all the world's writing systems, including traditional Mongolian script. Thus, this paper surveys the key fundamental component for digital library systems, the rendering of traditional Mongolian script, which is considered to be one of the most complex writing systems in the world.

2   The Traditional Mongolian Script

The traditional Mongolian script is written vertically from top to bottom in columns advancing from left to right. This script is the writing system for the Mongolian language and has three derivative scripts - Todo, Manchu, and Sibe (Xibe). The Todo script is used in Oirat and Kalmyks. The Sibe script is used in Xinjiang, in the northwest of China. Similar to Arabic, traditional Mongolian is a contextual script where letters are cursively joined and have initial, medial, and final presentation forms for the same letter. In most cases, the letters join together along a vertical stem, but in the case of certain consonants, which lack a trailing vertical stem, they may form a single ligature with a following vowel. In addition to these cursive and positional forms, many letters also have variant forms used in accordance with spelling and grammatical rules. Thus, the traditional Mongolian script is regarded as complex; encoding complex script features, as well as understanding the layout features and rules exclusively related to the script, are crucial for researchers and developers in the digital library community.

2.1   Grammar and Rules of Traditional Mongolian Script

Some important elements and grammatical rules of traditional Mongolian script are explained below:

2.1.1   Vowel Harmony

Traditional Mongolian script has a characteristic feature of "vowel harmony", whereby a word can only contain either back vowels ('a', 'o', 'u') or front vowels ('e', 'oe', 'ue'), but not both at the same time, with the exception only of a certain limited set of words–the majority of which are foreign words. That is to say, the vowels in a word are either all "masculine" and "neuter" (that is, back vowels plus 'i') or all "feminine" and "neuter" (that is, front vowels plus 'i'). Words that are written with masculine/neuter vowels are considered to be masculine, and words that are written with feminine/neuter vowels are considered to be feminine. The vowel 'i' is considered neutral and can therefore occur in both front and back voweled words, but when 'i' occurs in all syllables the word is considered to be front voweled and behaves as feminine (for example, taking feminine suffixes).

Vowel harmony is an important element of the encoding model, as the gender of a word determines the glyph form of the velar series of consonant letters for traditional Mongolian script. In traditional Mongolian script, the velar letters ('qa' and 'ga') have both masculine and feminine forms. The masculine and feminine forms of these letters have different pronunciations. When one of the velar consonants precedes a vowel, it takes the masculine form before masculine vowels, and the feminine form before feminine or neuter vowels. In the latter case, a ligature of the consonant and vowel is required. When one of these consonants precedes another, or is the final letter in a word, it may take either a masculine or feminine glyph form, depending on its context. Consequently, the rendering system should automatically select the correct gender form for these letters based on the gender of the word [4][5][6]. Vowel harmony in traditional Mongolian script is illustrated in Fig. 1.
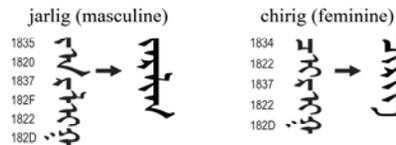
Fig. 1. Traditional Mongolian Script Gender Forms [4].

2.1.2   The 'a' and 'e' in a Word-final Position

In traditional Mongolian script, the letters 'a' and 'e' in a word-final position may take a "forward tail" or "backward tail" form depending on the preceding consonant that they are attached to. In some words, a final letter 'a' or 'e' is separated from the preceding consonant by a narrow gap; in that case the vowel always takes the "forward tail" form, and the 'a' or 'e' is an integral part of the word stem. Whether a final letter 'a' or 'e' is joined or separated is purely lexical and is not a question of varying orthography.

2.1.3   Syllable Closed Consonants

The traditional Mongolian script has another important rule – syllable closed consonants. Consonants which exist at the end of a syllable or a word, and the following letter is not a vowel, are considered to be syllable closed consonants. This rule could be interpreted as "a consonant final syllable rule".

Depending on whether a consonant is syllable closed or not, the glyph form of that consonant varies, taking either a masculine, feminine or variant glyph form. Also, there are two types of syllable closed consonants: "soft" and "hard". Hard syllable closed consonants are the traditional Mongolian letters 'ba', 'qa', 'ga', 'ra', 'sa', 'ta', and 'da'. Letters 'na', 'la', 'ma', 'wa', and 'ang' are soft syllable closed consonants. Syllable closed consonants affect the form of the following suffixes [6]. In this way, the rendering system should automatically select the correct glyph form for syllable closed consonants.

2.1.4 Case Suffix

In traditional Mongolian script, case suffixes are separated from the stem of a word or from other suffixes by a narrow gap. Suffixes have masculine and feminine pairs (for example, –dur/-tur and –dür/-tür), and a stem may take receive suffixes.

Any attached suffixes are considered to be an integral part of the word as a whole. A suffix affects the form of the preceding letters. What`s more, the final letter of the stem or suffix preceding the particular suffix takes the final positional form, whereas the first letter of the particular suffix usually takes a medial form or a final form (single letter suffixes), depending on the particular suffix.

Some special cases where the first letter of the suffix takes a normal initial form and a variant initial form are listed in Table 1.

Table 1. Some special cases' suffixes of traditional Mongolian [5][6].

Case	Case suffixes		Shape of the Initial glyph	Special attention
Dative-Locative Case		tur/tür	the initial form of 'ta'	Added to words ending in vowels and soft syllable closed consonants
		tu/tü		
		dur/dür	the initial form of 'da'	Added to words ending in hard syllable closed consonants
		du/dü		
Ablative Case		ača/eče	the initial form of 'e'	Pronunciation and encoding varies according to vowel harmony
Comitative Case		luγa	the initial form of 'la'	Added to masculine words
		lüge		Added to feminine words
		tai/tei	the initial form of 'da'	Pronunciation and encoding varies according to vowel harmony

Any rendering system should consider all the above rules and also must select the correct glyph form for a letter according to the grammatical rules of traditional Mongolian.

2.2   Traditional Mongolian Script in the Unicode Standard

From the Unicode Standard version 3.1 to the latest Unicode Standard version 5.2.0, traditional Mongolian script has been standardized in Unicode and isolated form for the vowels, and the initial form for the consonants are encoded at the range of U+1800-U+18AF [4]. Other encoding standards were surveyed by Garmaabazar et al. [3]. Traditional Mongolian script has also been standardized in the Chinese standard– GB 18030. However, some implementations that follow the GB 18030 standard, such as Menksoft Mongolian IME[2] use the Private Use Areas (PUA) at the range of U+E234-U+E34F, instead of using basic characters (U+1800-U+18AF) of Unicode for storing the traditional Mongolia text.

---

[2] http://en.wikipedia.org/wiki/Menksoft_Mongolian_IME

2.2.1 Representative Glyphs

The encoded characters in the Unicode range at U+1800-U+18AF are the isolated forms for the vowels and the initial forms for the consonants. Letters that share the same glyph forms are distinguished by using different positional forms for the Mongolian code range. For example, the representative glyph for U+1823 (Mongolian letter 'o') is in the isolated form, whereas the representative glyph for U+1824 (Mongolian letter) 'u' is in the initial form. The various positional and variant glyph forms of a letter are considered as presentation forms. It is the responsibility of the rendering system to select the correct glyph form for a letter according to its context. Thus, having a robust rendering algorithm is vital for displaying traditional Mongolian script correctly.

2.2.2 Variant forms: Free Variation Selectors

Free variation selectors are encoded in Unicode for traditional Mongolian script, when a glyph form cannot be predicted algorithmically by the rendering system. Those are:

- U+180B Mongolian free variation selector one (FVS1);
- U+180C Mongolian free variation selector two (FVS2); and
- U+180D Mongolian free variation selector three (FVS3).

The user needs to append an appropriate variation selector to the letter to indicate to the rendering system which glyph form is required. These format characters normally have no visual appearance. When required, a free variation selector immediately follows the base character it modifies. This combination of base character and variation selector is known as a standardized variant[3] [4].

2.2.3 Narrow No-Break Space

The narrow no-break space (NNBSP)–U+202F is encoded to define traditional Mongolian suffixes as an integral part of the word as a whole. Basically, a line break opportunity does not occur before a suffix, and whitespace is represented when using NNBSP [4].

2.2.4 Mongolian Vowel Separator

The Mongolian vowel separator (MVS)–U+180E is used to represent the whitespace that separates a final letter 'a' or 'e' from the rest of the word. MVS is very similar in function to NNBSP, as it divides a word with a narrow non-breaking whitespace. Whereas NNBSP marks off a grammatical suffix, the 'a' or 'e' following MVS is not a suffix but an integral part of the word stem. For example, the word 'qana' without a gap before the final letter 'a' means, "the outer casing of a vein," whereas the word 'qana' with a gap (MVS) before the final letter 'a' means, "the wall of a tent", as shown in Fig 2. The words 'qana' are encoded "U+182C, U+1820, U+1828, U+1820" and "U+182C, U+1820, U+1828, U+180E, U+1820" respectively.

---

[3] http://www.unicode.org/Public/UNIDATA/StandardizedVariants.html

The MVS always selects the forward tail form of a following vowel 'a' or 'e'. Also, it may affect the form of the preceding letter. The particular form that is taken by a letter preceding an MVS depends on the particular letter and in some cases on whether traditional or modern orthography is being used [4].
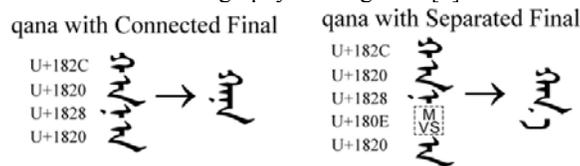

Fig. 2. Mongolian Vowel Separator [4].

2.3   OpenType Format

The OpenType format is a cross-platform compatible font format developed jointly by Adobe and Microsoft. The OpenType supports widely expanded multilingual character sets and layout features, which provides richer linguistic support and advanced typographic control, such as ligatures, glyph substitution, swash variants, kerning, and more. OpenType fonts allow embedding the traditional Mongolian script rules in a single file. Microsoft developed the guidelines for creating and supporting OpenType fonts for traditional Mongolian Script [7].

Recently, several OpenType fonts – Code2000, Simsun-18030, Daicing fonts, Manchu Font 2005, Mongolian Baiti, MongolUsug and MongolianScript were developed for traditional Mongolian script. We surveyed possible candidate fonts Mongolian Baiti, MongolUsug, and MongolianScript for testing to render traditional Mongolian script correctly.

2.3.1   Mongolian Baiti

This font was developed by the Founder Corporation, Peking University. It is important to note that Microsoft distributed Mongolian Baiti with MS Windows Vista and Windows 7. The latest version is 5.01. Isolated forms for the vowels and the initial form for the consonants are encoded at the range of U+1800–U+18AF. The rest of the variant forms of glyphs and presentation forms are stored within the font file and indexed by Glyph Index (GID).

2.3.2   MongolianScript

This font was developed by Erdenechimeg Myatav, the task force leader of the proposal "Traditional Mongolian Script in the ISO/IEC 10646 and Unicode Standards"[4]. The latest version is 2.0. Variant forms of glyphs or presentation forms are encoded at the range of U+F300–U+F3B0. The ligature set is encoded at the range of U+F400–U+F4C1.

---

[4] http://www.iist.unu.edu/newrh/III/1/docs/techreports/report170a.tgz

2.3.3 MongolUsug

This font was distributed with traditional Mongolian script editor–VertNote[5]. The latest version is 2.37. Variant forms of glyphs, presentation forms, and ligatures are encoded in the PUA at the range of U+E000–U+E811.

2.4   Rendering System

Recently, rendering traditional Mongolian script via Unicode and OpenType font has begun to be supported in most Windows applications (including Microsoft Office Publisher, most Adobe applications, and Microsoft Office, from version 2003), Windows Vista, Windows 7 and many Mac OS X applications, including Apple's own, such as TextEdit, Pages and Keynote. In older versions of Windows, OpenType support for traditional Mongolian script can be added by updating the Uniscribe driver to the latest version. In Unix-like systems, there are active developments such as Pango[6]. The XenoType Technologies is working to release a Mongolian Language Kit[7] for Mac OS X.

However, no attempt has been carried out to check the rendering systems of traditional Mongolian script and shaping algorithms of the OpenType fonts.

3   Experiments

We conducted a preliminary experiment to survey the rendering algorithm of OpenType fonts as well as to check that traditional Mongolian script contents are displayed correctly. Rendering algorithms of the three Mongolian OpenType fonts: Mongolian Baiti, MongolianScript, and MongolUsug, are reviewed. In addition to checking the basic rules of traditional Mongolian script, we tested the following complex grammatical rules:

- Vowel harmony, feminine and masculine words;
- Syllable closed consonants;
- Case suffixes; and
- Usage of free variant selectors.

We rendered text and HTML files with the same words of traditional Mongolian script in the different OpenType fonts, and compared the rendered results with the correct forms. We selected over a hundred feminine words with syllable closed consonants or suffixes. We tested all 43 standardized variants of traditional Mongolian script. The experimental setup is for the English version of Windows 7, build 7600.16385 with Internet Explorer version 8.0 and Uniscribe – Unicode script processor version 1.626.

---

[5] http://www.uukhai.com/archives/1152

[6] http://www.pango.org/

[7] http://www.xenotypetech.com/osxMongolian.html

3.1 Vowel Harmony, Feminine and Masculine Words and Syllable Closed Consonants

The rendered results of our experiment, which surveys the rules of traditional Mongolian script (vowel harmony and syllable closed consonants were unusual as illustrated in Table 2), resulted in irregular for all fonts. For instance, the syllable closed consonant 'ga'-U+182D of the feminine word has rendered correctly in some words, but not in all. MongolUsug in particular has failed to render the syllable closed consonants. Errors which occurred are highlighted and explained in the last column.

3.2 Suffixes

MongolianScript has failed to render almost all suffixes, though Mongolian Baiti and MongolUsug have failed on comitative case suffixes. The rendered results of the grammatical suffixes for traditional Mongolian script are illustrated in Table 3.

Table 2. Rendering the traditional Mongolian script in various OTFs.

3.3 Usage of Free Variant Selectors

Mongolian Baiti has failed to render 6 variants, MongolianScript 24 and MongolUsug 12 variants of all 43 forms. The rendered results of the selected standardized variant [4] for traditional Mongolian script are illustrated in Table 4. Errors which occurred are highlighted and explained briefly in the last column.

Table 3. Rendering the grammatical suffixes of traditional Mongolian script in various OTFs.

4 Conclusion and Recommendation

In this paper, we surveyed the rendering issues of complex text layouts especially in regard to traditional Mongolian script. We analyzed existing OpenType fonts and their rendering schemes for traditional Mongolian script. The study produced some errors, and revealed grammatical rules, which are not documented in international standards.

OpenType fonts that we surveyed did not include certain rules of traditional Mongolian script such as vowel harmony, feminine and masculine words, syllable closed consonants and case suffixes. We realized that some rules such as syllable closed consonants and the usage of case suffixes, have not been well documented in Unicode standard.

All fonts failed to display the variant glyphs with free variant selectors correctly, which were already standardized in Unicode. In general, Mongolian Baiti was better

than others with few shortcomings on rendering variant forms, suffixes and feminine words.

In addition, all fonts need some improvements in their rendering algorithms and all grammatical rules need to be standardized in international use such as in Unicode. The guidelines for creating and supporting OpenType fonts for traditional Mongolian Script [7] need to be updated as well.

Finally, we believe this survey will make some contribution to developers of digital libraries that utilize complex scripts such as traditional Mongolian script.

Table 4. Rendering the variant glyphs of traditional Mongolian script in various OTFs.

Unicode Basic Glyph	Position	Description of variant appearance	Character Sequence	Correct Variant Glyph	Mongolian Baiti	Mongolian Script	Mongol Usug
	medial	second form of "a"	1820 180B				
		third form of "a"	1820 180C				
	initial	second form of "e"	1821 180B				
	final						
		second form of "i"	1822 180B				
	medial	second form of "o"	1823 180B				
		third form of "ue"	1826 180C				
≈							≈
	isolate	feminine second form of "qa"	182C 180B				
	initial	second form of "qa"	182C 180B				
	medial	third form of "qa"	182C 180C				
		fourth form of "qa"	182C 180D				
	medial	second form of "ga"	182D 180B				
	final	feminine form of "ga"	182D 180B				
≈							≈
	medial	feminine form of "ga"	182D 180D				
	initial	second form of "da"	1833 180B				
	medial						
		second form of "ja"	1835 180B				
	medial	second form of "ya"	1836 180B				
		third form of "ya"	1836 180C				
	final	second form of "wa"	1838 180B				

References

1. Garmaabazar, K., Maeda, A.: Retrieval Technique with the Modern Mongolian Query on Traditional Mongolian Text. In: 9th International Conference on Asian Digital Libraries, LNCS, vol. 4128, pp. 478--481. Springer (2006)
2. Garmaabazar, K., Maeda, A.: Building a Digital Library of Traditional Mongolian Historical Documents. In: 7th ACM/IEEE-CS Joint Conference on Digital Libraries, p. 483. ACM (2007)
3. Garmaabazar, K., Maeda, A.: Developing a Traditional Mongolian Script Digital Library. In: 11th International Conference on Asia-Pacific Digital Libraries: Universal and Ubiquitous Access to Information, LNCS, vol. 5362, pp. 41--50. Springer (2008)
4. The Unicode Consortium: The Unicode Standard 5.0. Addison-Wesley (2007)
5. Pugh, ERE., Mongolian Grammar Reference, http://www.linguamongolia.com/vhar1.html
6. Choimaa, Sh.: Mongol bicigiin zov bicix dŭrmiin xuraangui (in Mongolian), The Mongolia Society special papers, Bloomington, (1991)

7. Creating and Supporting OpenType Fonts for the Mongolian Script, http://www.microsoft.com/typography/otfntdev/mongolot/