广西科学院学报
Journal of Guangxi Academy of Sciences
2018,34(1):63～67
Vol. 34, No. 1 February 2018

网络优先数字出版时间:2018-02-23 DOI:10.13657/j.cnki.gxkxyxb.20180223.001
网络优先数字出版地址:http://kns.cnki.net/kcms/detail/45.1075.N.20180223.0926.002.html

蒙古文复杂文本布局引擎的标准符合性测试*

Standard Conformance Test of Mongolian Complex Text Layout Engine

呼斯勒[1]**,白双成[1],确精扎布[2]
Huslee[1], BAI Shuangcheng[1], Choijingjab[2]

(1.内蒙古社会科学院,内蒙古呼和浩特 010020;2.内蒙古大学,内蒙古呼和浩特 010019)
(1. Inner Mongolia Academy of Social Science, Hohhot, Inner Mongolia, 010020, China; 2. Inner Mongolia University, Hohhot, Inner Mongolia, 010019, China)

摘要:以蒙古文编码国家标准的研制及其系统实现方面的工作为基础,针对蒙古文复杂文本布局引擎(CTL Engine)及其 OpenType 字库的系统结构,提出蒙古文复杂文本布局引擎的标准符合性测试(Conformance Test for Standards)方案,定义蒙古文复杂文本布局引擎的测试点及其测试实例,并以关键软件系统为依托测试和分析 Uniscribe 和 HarfBuzz 等支持蒙古文的复杂文本布局引擎。

关键词:蒙古文编码 复杂文本布局引擎 标准符合性测试 国家标准
中图分类号:TP391.1 **文献标识码**:A **文章编号**:1002-7378(2018)01-0063-05

Abstract: Based on the research and development of the Mongolian national standard for the encoding and its system implementation, this paper puts forward the Standard Conformance Test Plan of Mongolian CTL Engine in allusion to the Mongolian complex text layout (CTL) engine and the system structure of the Open Type font. The Conformance Test for Standards scheme defines the test points and test examples of the Mongolian CTL Engine, and relies on a key software system to test and analyze complex text layout engines such as Uniscribe and HarfBuzz that support Mongolia.

Key words: Mongolian encoding, complex text layout engine, standard conformance test, the national standard

0 引言

蒙古文信息处理系统的标准符合性测试(Conformance Test for Standards)主要为蒙古文信息处理产品、技术、服务或相关知识产权提供功能性、安全性、实用性方面的测试,这是一种标准化服务,将会广泛应用于蒙古文信息处理领域。蒙古文信息处理系统的标准符合性测试目前还处在起步阶段,但蒙古文复杂文本布局引擎(CTL Engine)的标准符合性测试依据已基本具备[1-4],在蒙古文复杂文本布局引擎的标准符合性测试系统的研究、设计、开发方面也取得了一定进展[5-6]。GB/T 29270.1—2012[4]是蒙古文信息处理系统在标准符合性测试方面的第一个国家标准,也是此项工作起步的标志。该标准的基础是国家标准《GB/T 26226—2010 信息技术

收稿日期:2017-12-14
修回日期:2018-01-08
作者简介:呼斯勒(1977—),男,副研究员,主要从事蒙古文信息处理研究,E-mail:husela@163.com。
*内蒙古自治区蒙古语言文字科研项目(MW-YB-2016026),内蒙古自治区蒙古语言文字信息化专项扶持项目(MW-2017-MGYWXXH-001)和国家社会科学基金西部项目(17XTQ001)资助。
**通信作者。

蒙古文变形显现字符集和控制字符使用规则》[2]，并以此为基础阐述了蒙古文编码字符集的标准符合性测试原理和关键内容。其中，蒙古文复杂文本布局引擎的标准符合性测试是该标准中“屏幕显示”和“打印”等“输出形式”的基础。赵颖霞[5]首次公开蒙古文信息处理系统的标准符合性检测(测试)系统的设计开发工作及其进展，该研究将“蒙古文编码字符集标准符合性检测”视为蒙古文复杂文本布局引擎的标准符合性测试——“字符显示引擎检测”。之后，何正安等[6]公开了另一项蒙古文信息处理系统标准符合性测试系统的设计开发情况，明确指出正确输出蒙古文“涉及到字符引擎(蒙古文复杂文本布局引擎)与蒙古文字库”，并且“先确定字符引擎是否正确，然后使用已经通过检测的字符引擎来检测被测字型”。该研究认识到蒙古文复杂文本布局引擎及其蒙古文 OpenType 字库之间的关系及区别。由此可见，赵颖霞[5]和何正安等[6]对蒙古文复杂文本布局引擎的认识不一致，两个系统的实现细节也存在明显的差别。然而，这两个研究中涉及蒙古文复杂文本布局引擎的标准符合性测试要完成的目标一致，即最后要得出的结果是蒙古文的输出(显示、打印)是否正确。为了更好地说明问题，引用如下两个日常应用中的典型问题。

问题一：在 Windows 7/8 系统的记事本中，当用户输入蒙古文词{182A 1822 1834 1822 182D}时其词尾的辅音字母 GA 的字形()不正确。

问题二：在 Windows 7/8/10 系统的记事本中，当用户在蒙古文词(){182 1820 1837 180E 1820}中移动光标跨过蒙古文元音间隔符(MVS)时，发现光标出现原地踏步的“异常”现象。

依据赵颖霞[5]和何正安等[6]的研究，他们研发的两个系统应该都能够准确发现“问题一”，但不能回答其原因。而对于“问题二”两个系统应该都不能发现。这是因为 Windows 7/8/10 的复杂文本布局引擎 Uniscribe 及其蒙古文 OpenType 字库“合伙”欺骗了测试系统。

本课题组多年来在蒙古文复杂文本布局引擎及其 OpenType 字库的设计开发工作中，深入了解蒙古文复杂文本布局引擎及其 OpenType 字库设计开发及运行原理(图 1)，以及基于专用蒙古文 OpenType 字库的简单实验，可以准确理解上述两个问题及其发生原因。“问题一”出现在 Windows 7/8/10 中的蒙古文 OpenType 字库“Mongolian Baiti”中，即该 OpenType 字库在蒙古文词的阴阳性处理方面存在缺陷，而“问题二”出现在 Windows 7/8/10 系统的复杂文本布局引擎(Uniscribe)中，即对蒙古文元音间隔符(MVS)处理存在缺陷。关于复杂文本布局引擎 Uniscribe 和 HarfBuzz 及其蒙古文 OpenType 字库设计开发，可参阅文献[7-8]。

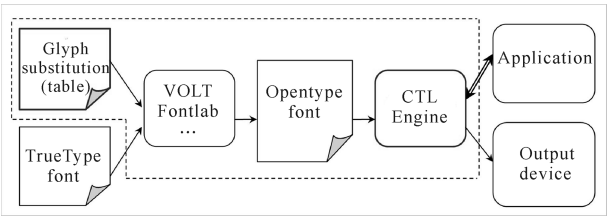

图 1 蒙古文复杂文本布局引擎及其 OpenType 字库

Fig. 1 The Mongolian complex text layout engine and its OpenType fonts

1 测试点及测试样例

1.1 测试点定义

除了须涵盖蒙古文字符集[4]以外，在蒙古文编码国家标准中需要测试的概念、控制字符、复杂文本布局特征较多，有些是蒙古文独有的，有些是同其他文种公用的。这些概念、控制字符、复杂文本布局特征中有些是属于复杂文本布局引擎需要处理的，针对蒙古文复杂文本布局引擎，对需要测试的内容进行定义，并将这些内容称之为“测试点”。

蒙古文的词内位置替换、上下文的替换、(强制和非强制)合体字非常多且直接关系到蒙古文正字法理论、规则和实践方面的内容，不具备蒙古文正字法基础知识则不易理解。其中，传统蒙古文可参考文献[9]，托忒文可参考文献[10]，锡伯文可参考文献[11]，满文可参考文献[12]。

蒙古文复杂文本布局引擎的测试点定义如下：

A. 词内位置替换(Position)

蒙古文字母在词内位置的替换(变形)有独立(Isolate)、词首(Initial)、词中(Medial)和词尾(Final)等 4 个复杂文本布局特征。

B. 上下文替换(Context)

蒙古文字符基于上下文的替换非常丰富。例如：辅音 NA(U+1828)的第一词中形  在元音字母前被替换为其第二词中形  ；依据蒙古文词的阴阳性分析结果正确替换音节末辅音 GA(182D)的字形等等。

C. (强制、非强制)合体字(Ligature)

蒙古文多个字母拼写时，字母字形融合在一起不易切分的合体字，其中圆头辅音和元音的组合是最典型的(强制)合体字。

D.(未定义)

E.蒙古文元音间隔符(Mongolian Vowel Separator,MVS)

用于传统蒙古文词尾的分写元音字母 A/E 与其前面的辅音字母之间。蒙古文元音间隔符不仅是个控制字符,同时也是个空格[1,13-14],然而在蒙古文编码国家标准中未明确定义其空格特性[2-3]。

F.蒙古文自由变体选择符(Free Variant Selector,FVSn)

用于区别在同一条件下出现的同一个字母的不同自由变体,蒙古文编码国家标准中定义了 3 个,即 FVS1、FVS2、FVS3,本文将其统称为 FVSn。

G.窄宽无间断空格(Narrow No-Break Space,NNBSP)

用于蒙古文词干与其分写词缀之间,在蒙古文中应用非常频繁。

H.零宽连接符(Zero Width Joiner,ZWJ)

用于其相邻的蒙古文字符,以连写的显现形式显现。

I.零宽禁连符(Zero Width Non-Joiner,ZWNJ)

用于强行断开相邻两个蒙古文字符的正常连写。

J.FVSn 和 ZWJ 的组合

该测试点是针对某些软件系统中蒙古文字母后面"FVSn 和 ZWJ 的组合"出现时发生的问题定义的。

K.非蒙古文字符和 NNBSP 的组合

该测试点是针对某些软件系统中,蒙古文分写词缀前面的"非蒙古文字符和 NNBSP 的组合"对分写词缀的影响定义的。其中,非蒙古文字符是指非 18XX 编码区字符,包括"N.引用标点符号"的内容。

L.(未定义)

M.自有标点符号

指蒙古文编码国家标准中 18XX 编码区的标点符号[2],例如传统蒙古文省略号(1801)、逗号(1802)、句号(1803)等。

N.引用标点符号

指蒙古文编码国家标准中非 18XX 编码区的标点符号[2],例如:问号(FE15)、叹号(FE16)、括弧(FE35、FE36)、书名号(FE3D、FE3E)等。

在上面的定义中,约定每个测试点的大写拉丁字母序号是固定不变的,并在蒙古文复杂文本布局引擎的标准符合性测试和分析中使用这些序号。

1.2 测试实例

针对上一节中定义的"测试点"设计一份测试样例(图 2),经过多年的实际应用和不断修改调整,该测试实例目前已非常成熟。该测试样例不仅涵盖所有测试点,且文本量很少,已非常适用于复杂文本布局引擎的标准符合性测试及其分析工作。

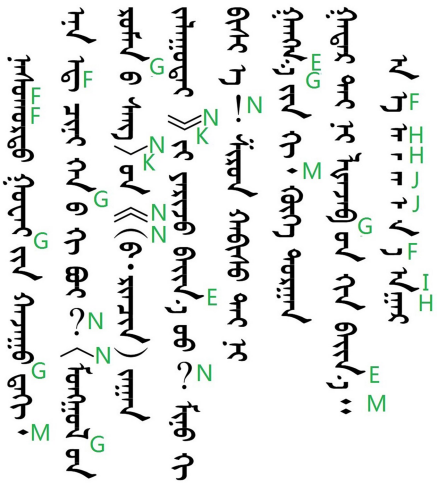

图 2 蒙古文复杂文本布局引擎测试实例

Fig. 2 Testing sample of the Mongolian CTL Engine

图 2 的"测试实例"中只标注了"测试点"的"拉丁字母序号",且只在少数可检查的蒙古文单词及其位置上标注了具有代表性的测试点,没有一一列举所有测试点。图 2 中也未标注测试点 A、B、C 等,这是因为了解蒙古文及其编码系统的人均能识别这些测试点并准确判断其正确与否,而且其出现次数相对较多。

2 测试实践

近几年,实际工作中一直在使用本文提出的蒙古文复杂文本布局引擎的标准符合性测试方案。采用该方案对 Uniscribe、HarfBuzz 等复杂文本布局引擎进行标准符合性测试,获得了良好的效果。

Windows 7/8/10 操作系统、微软 Office 2007/2010/2013/2016 办公软件和 IE 浏览器共享复杂文本布局引擎 Uniscribe。其中,微软 Office 2007/2010/2013/2016 自带复杂文本布局引擎,而 IE 浏览器则依赖操作系统的复杂文本布局引擎。此外,据了解,最新的 Linux、Android 操作系统和 Chrome、Firefox、Opera 等主流浏览器都应用复杂文本布局引擎 HarfBuzz。开源项目 Chromium 及其衍生浏览器在 Mac OS X 和 iOS 系统中依赖操作系统的复杂文本布局引擎,Chrome 浏览器 29 以下的版本在 Windows、Android 系统中也依赖操作系

统的复杂文本布局引擎。

2.1 Uniscribe

为了测试在蒙古文中应用广泛的 Uniscribe，选择 Windows 7/8/10 操作系统的记事本、IE 浏览器和 Microsoft Office 2007/2010/2013 等具体的软件系统。

记事本中对测试点 E 的处理上存在缺陷，未能体现其“空格”特性。用户可以通过记事本右键菜单中的“显示 Unicode 控制字符(S)”功能实现其可视特性，也可以通过蒙古文 OpenType 字库中的技巧来支持其空格特性。

IE 浏览器中，除了对测试点 E 的处理存在与记事本一样的缺陷，对测试点 K 和 N 的处理也存在问题，这个正在严重影响蒙古文编码国家标准在 IE 浏览器中的应用。

Microsoft Office 中，除了对测试点 E、N 的处理存在与 IE 浏览器一样的缺陷，该系统中的用户编辑操作总是莫名其妙地影响蒙古文的正确输出(显示、打印)。这些问题导致普通用户对 Microsoft Office 又爱又恨的现状。

2.2 HarfBuzz

本研究以 Chrome 浏览器作为 HarfBuzz 复杂文本布局引擎的代表进行标准符合性测试和分析。表 1 展示了 3 年内 Chrome 浏览器主要版本的测试和分析结果。其中，字母 E、F、G、H 表示 Chrome 浏览器对这些测试点的处理存在缺陷。需要注意的是，对测试点 E 的处理总是存在缺陷。

表 1 Chrome 浏览器的测试

Table 1 Testing of Chrome browser

Chrome 版本 Chrome version	Operating System OS					
	NT 5.X	NT 6.X	Linux	Android	Mac OS X	iOS
27	✗	✗	⊗	EFGH	✗	✗
28	✗	✗	EFGH	EFGH	✗	✗
29～30	✗	E	EFGH	EFGH	✗	✗
31	✗	E	⊗	EFGH	✗	✗
32	✗	E	EFGH	EFGH	✗	✗
33	EFGH	EFGH	FGH	EFGH	✗	✗
34～37	EF	EF	EF	EF	✗	✗
38	E	E	E	E	✗	✗
更高 Higer	E	E	E	E	✗	✗

注：NT 5.X 是指内核为 NT 5 的系列 Windows 操作系统；NT 6.X 是指内核为 NT 6 的系列 Windows 操作系统；✗表示依赖操作系统；⊗ 表示未测试或测试数据遗失

Note: NT 5.X indicates that the Windows OS core is NT 5; NT 6.X indicates that the Windows OS core is NT 6; ✗indicates depending on the operating system; ⊗ indicates a later or missing test data; Higer indicates a later version of Chrome

在 NT 5.X 系列 Windows 操作系统中，Chrome 浏览器从 33 版本开始引入了独立的复杂文本布局引擎，而在 NT 6.X 系列 Windows 操作系统中，Chrome 浏览器从 29 版本开始引入了独立的复杂文本布局引擎，即开源项目 HarfBuzz，从而对蒙古文的处理不依赖操作系统的复杂文本布局引擎。在 Linux 桌面操作系统和 Android 操作系统中，Chrome 浏览器 27 及更高版本中一直在应用独立的复杂文本布局引擎，所以对蒙古文的处理不依赖操作系统的复杂文本布局引擎。然而，测试发现 Chrome 浏览器在 Mac OS X 和 iOS 操作系统中处理方式与其它操作系统中的处理方式不一致，也未查阅到准确的官方信息，估计须应用 AAT 字体技术。

3 结束语

为突出重点，在这里只讨论最新 Uniscribe 和 HarfBuzz 的共性问题，即测试点 E 的处理缺陷。这也是已有的标准符合性测试系统不容易发现的问题。进一步对蒙古文元音间隔符(MVS)出现的这一问题及其根源进行比较广泛的研究和分析，发现蒙古文元音间隔符(MVS)在“Unicode Character Database (UCD)[14]”中的“General Category”和“Bidirectional Class”两个属性值(Property Value)存在问题。

MVS Cf;0;BN; ;;;;N;;;;;

NNBSP Zs;0;CS;<NB>;;;;N;;;;;

从属性的对比中可以发现，MVS 的“General

Category"和"Bidirectional Class"属性值分别为 Cf 和 BN,而 NNBSP 对应的属性值为 Zs 和 CS。可以理解为 MVS 是纯粹的控制符,且无"空格"特征,所以 Uniscribe 和 HarfBuzz 等复杂文本布局引擎未能正确处理 MVS。实际上,MVS 和 NNBSP 拥有非常相似的字符特征[14]。

此外,本文未对苹果系统(Mac OS X 和 iOS)的复杂文本布局引擎及其 AAT 字库的设计开发进行研究,这是我们目前正在进行的重点工作之一。

参考文献:

[1] 确精扎布.蒙古文编码[M].呼和浩特:内蒙古大学出版社,2000.
CHOIJINGJAB. Mongolian encoding[M]. Hohhot: Inner Mongolia University Press,2000.

[2] 确精扎布,陈壮,何正安,等.信息技术 蒙古文变形显现字符集和控制字符使用规则:GB/T 26226—2010[S].北京:中国标准出版社,2011.
CHOIJINGJAB,CHEN Z,HE Z A,et al. Information technology—Mongolian presentation forms character set and use rules of controlling character: GB/T 26226—2010[S]. Beijing:China Standard Press,2011.

[3] 确精扎布,何正安,达胡白乙拉,等.信息技术 传统蒙古文名义字符、变形显现字符和控制字符使用规则:GB 25914—2010[S].北京:中国标准出版社,2011.
CHOIJINGJAB,HE Z A,DAHUBAIYILA,et al. Information technology—Traditional Mongolian nominal characters, presentation characters and use rules of controlling characters: GB 25914—2010[S]. Beijing: China Standard Press,2011.

[4] 王欣,何正安,达胡白乙拉,等.信息技术 编码字符集测试规范 第1部分:蒙古文:GB/T 29270.1—2012[S].北京:中国标准出版社,2012.
WANG X,HE Z A,DAHUBAIYILA,et al. Information technology—Specification for the testing of coded character sets—Part 1: Mongolian: GB/T 29270.1—2012[S]. Beijing:China Standard Press,2012.

[5] 赵颖霞.蒙古文信息处理产品标准符合性检测系统的设计与实现[D].呼和浩特:内蒙古大学,2013.
ZHAO Y X. The design and implementation of Mongolian information processing products standards compliance testing system[D]. Hohhot:Inner Mongolia University,2013.

[6] 何正安,王欣,陈海.蒙古文软件标准符合性检测的研究与设计[J].信息技术与标准化,2015(1/2):47-49,53.
HE Z A,WANG X,CHEN H. Research and design of standard conformance test of Mongolian software[J]. Information Technology & Standardization, 2015(1/2):47-49,53.

[7] Microsoft. Typography[EB/OL].[2017-10-11]. http://www.microsoft.com/Typography.

[8] Freedesktop.org. HarfBuzz[EB/OL].[2017-10-11]. http://www.freedesktop.org/wiki/Software/HarfBuzz/.

[9] 确精扎布.传统蒙古文名义字符到变形显现字符的转换规则:第九稿[M]//确精扎布.确精扎布蒙古文信息处理专辑.呼和浩特:内蒙古教育出版社,2014:380-403.
CHOIJINGJAB. The traditional Mongolian nominal characters to presentation character conversion rules:9 edition[M]// CHOIJINGJAB. Choijingjab's Mongolian information processing album. Hohhot:Inner Mongolia Education Press,2014:380-403.

[10] 确精扎布.托忒文名义字符到变形显现字符的转换规则:第五稿[M]// 确精扎布.确精扎布蒙古文信息处理专辑.呼和浩特:内蒙古教育出版社,2014:404-416.
CHOIJINGJAB. The Todo nominal characters to presentation character conversion rules: 5 edition[M]//CHOIJINGJAB. Choijingjab's Mongolian information processing album. Hohhot:Inner Mongolia Education Press,2014:404-416.

[11] 确精扎布.锡伯文名义字符到变形显现字符的转换规则:第二稿[M].确精扎布.确精扎布蒙古文信息处理专辑.呼和浩特:内蒙古教育出版社,2014:417-426.
CHOIJINGJAB. The Sibe nominal characters to presentation character conversion rules: 2 edition[M]// CHOIJINGJAB. Choijingjab's Mongolian information processing album. Hohhot:Inner Mongolia Education Press,2014:417-426.

[12] 确精扎布.满文名义字符到变形显现字符的转换规则:第四稿[M]// 确精扎布.确精扎布蒙古文信息处理专辑.呼和浩特:内蒙古教育出版社,2014:427-440.
CHOIJINGJAB. The Manchu nominal characters to presentation character conversion rules: 4 edition[M]//CHOIJINGJAB. Choijingjab's Mongolian information processing album. Hohhot:Inner Mongolia Education Press,2014:427-440.

[13] The Unicode Standard[S/OL].[2017-10-15]. http://www.unicode.org/.

[14] 中华人民共和国国家质量监督检验检疫总局,中国国家标准化管理委员会.信息技术 通用多八位编码字符集(UCS):GB 13000—2010[S].北京:中国标准出版社,2011.
General Administration of Quality Supervision Inspection and Quarantine of the People's Republic of China,Standardization Administration of the People's Republic of China. Information technology—Universal multiple-octet coded character set (UCS): GB 13000—2010[S]. Beijing: China Standard Press, 2011.

(责任编辑:陆 雁)