

蒙古文拼写形式多样化现象研究

白双成
内蒙古社会科学院/呼和浩特市
baishuangcheng@qq.com

呼斯勒
内蒙古社会科学院/呼和浩特市
husela@163.com

摘要

蒙古文文本中存在一个有别于多数其他文字的特别现象——看到的单词字形正确但其内码序列不正确，或者说单词“变形显现字形”序列正确但“名义字符”序列不正确的现象，我们称其为蒙古文的拼写形式多样化现象。本文先定义该现象及相关概念，再通过简单图示、例词拼写形式穷举、新闻语料统计分析和基于整篇文章标注统计等多方式、多角度论证这一现象的事实性和严重性，分析导致这一现象的深层原因并指出拼写形式多样化对蒙古文信息处理和应用方面的严重影响，最后提出通过推广普及录入规范和标准提高用户意识、使用智能输入法避免误录、使用校对纠错工具后纠正、基于生语料的统计学习方法为补充等多途径解决方法。本文对蒙古文标准编码的推广普及具有较好的参考价值。

1 引言

蒙古文编码国家标准将蒙古文字符分为“名义字符” (Nominal Character) 集和“变形显现字形” (Presentation Form/Character) 集，并规定前者用于信息存储、传输和计算，有明确的码位，后者（即蒙古文各类纷繁复杂的字母变形形式）仅用于信息的输出（显示和打印），不需要码位 (Unicode, 2020)。把一种文字的字符分为“名义字符”和“变形显现字形”进行编码，并把信息传输与信息输出截然分开来处理，是一个完全不同于英法德等西方文字和汉日韩等东方文字通用方式的较特殊的编码结构与编码方式。原本在其他文字中特别简单的字符 C_i 到字形 T_i 的“一对一”“映射关系”，变成了字符序列 $C_1C_2\dots C_n$ 到字形序列 $T_1T_2\dots T_m$ 的“多对多”的复杂“转换过程”。由于蒙古文的这一转换过程远复杂于阿拉伯文，仅仅依靠词首、词中、词尾等词内位置进行调形的阿拉伯文成功经验已无法满足蒙古文复杂变形需求。幸好目前已有多数操作系统、浏览器及通用基础软件环境具备了这一编码方式的实现条件，技术层面上基本攻克了这一复杂转换过程的技术难题。只待各大厂商严格遵照“用户协定”和转换规则 (确精扎布, 2014) 即有希望实现对于任意一个名义字符序列 $C_1C_2\dots C_n$ 转换为唯一一个变形显现字形序列 $T_1T_2\dots T_m$ ，从而真正实现编码统一。为解决同形词 (Homograph) 完全有可能存在另一个名义字符序列 $C_1'C_2'\dots C_k'$ (k与n不一定相等) 同样被转换为 $T_1T_2\dots T_m$ ，Unicode标准Core Specification中13.4节给出了如表1的一个案例。

字符序列C			词形	词意	字符序列C			词形	词意
1	1824	u		长度	6	1823	o		宫殿
1	1837	r			6	1837	r		
1	1832	t			6	1833	d		
1	1824	u			6	1824	u		

Table 1: Unicode标准提供的蒙古文词形趋同案例

但是， C' 是否是一个完全不合法的字符序列或根本不是用户想要的字符序列呢？这种我们所不希望的事情发生概率如何呢？非常不幸的是，由于蒙古文的“一字一音”、“一字多形”和“一

表 2 穷举 ündjüsitün 一词拼写形式表

字母	ü	ö	j	s	t	n	ü	j	s	t	n
拼写形式	ü	ö	j	s	t	n	ü	j	s	t	n
	ü	ö	j	s	t	n	ü	j	s	t	n
	ü	ö	j	s	t	n	ü	j	s	t	n
	ü	ö	j	s	t	n	ü	j	s	t	n
组合数	2+2*3+3*4*3+3*2	3	2	4	1	4	2	3	3		

2.3 真实语料统计下的拼写多样化

穷举分析毕竟是理论推导，而且有的拼写方式过于复杂，实际出现概率极低。那么实际应用情况是什么样呢？为此我们抓取三个新闻网站的文字性新闻报道页面，并经过行序恢复、HTML标签剔除、编码转换等一系列预处理后形成了测试集，本文称其为MGLNews。其数据情况如表3所示。

在可获取众多数字资源中，尤其是各种网络资源中选择这三个新闻网站数据作为实验数据的原因主要有

Website	Docs	Sentences	Tokens	Types
中国蒙古语新闻网MNN	86189	4008244	49243265	312902
中国蒙古语广播网CNR	43857	885970	11493090	84458
央视网CNTV	8195	180570	2039115	57216
合计	138241	1214513	62775470	454576

Table 2: MGLNews数据情况汇总

- **可获取性**：选择网络资源的首要原因是它有便利的可获取性，便于其他研究人员也可以获取对照。
- **可靠性**：正规新闻媒体机构主办，稿件经过编辑、审核等多道编审流程发布。具有内容相对可靠、术语相对规范统一、干扰因素相对低等优势。
- **时效性**：作为新闻类网站，具有较强时效性，可基本反映新词术语和蒙古文使用现状。当然，时政类新闻为主的新闻内容词汇量必然比不上文学作品，文风也相对拘谨。
- **代表性**：虽然是正规新闻稿件，但依然存在较为严重的读音错误，也不乏字形拼写错误，具有普遍代表性。
- **可验证性**：因工作便利，可对此三个网站爬取内容进行正确性验证，确保网页爬取、网页模板分析、格式转换、行序回复等工作正确。
- **结构性**：可额外获得结构化数据（Structured Data），便于进行按文档种类各自分类训练，便于进行关键字抽取（Keyword Extraction）、摘要生成（Summary Generation）等有监督学习（Supervised Learning）的后续研究工作。搜索引擎中结构化搜索（Structural Search）就是基于MNN的结构化数据（实际用TREC标记标示）进行了充分训练和验证。
- **可延续性**：这三个新闻网站每日稳定更新。通过前期试验，语料搜集工具成熟后，可以定期更新扩充语料。

2.4 基于完整文章标注统计的拼写多样化

内蒙古日报2014年6月6日第一版的题为 $\text{ᠮᠣᠩᠭᠣᠯ ᠵᠢᠨᠠᠨᠢ ᠵᠢᠨᠠᠨᠢ ᠵᠢᠨᠠᠨᠢ ᠵᠢᠨᠠᠨᠢ}$ 的一篇文章⁰ 共计411词。其中词形正确但读音错误的拼写现象共出现182次，竟达45.25%¹。内蒙古日报社编辑们作为专业文字工作者，非常清楚自己要写的词如何拼读，只是他们没有意识到拼写正确的必要性，受到传统纸质媒体出版的多年影响，一般认为传递的是蒙古文字形，没有必要一定要读音正确。或者说，他们将基于“音”的标准编码还是当成基于“形”的编码来用。这种现象绝不是内蒙古日报独有，它是整个蒙古文信息处理和应用中普遍存在的，也是我们需要认真面对和解决的重要问题。

经认真分析，我们可以看出这篇文件的录入者有几个录入习惯：①不用o、ö，只用u、ü；②不管读音是t，还是d，字形 ᠲ 用t输入，字形 ᠳ 都用d输入；③阴性词里的x、g都用g录入；④词末的 ᠨ 不管阳性、阴性都一律用u录入；⑤不用分写词缀 (ᠶᠢᠨ ᠵᠢᠨ) 前必须用的202F等等。

经我们持续观察，这种录入习惯具有普遍代表性，仅仅是不同人对不同键位具有一定偏好而已，但总体思路基本相似。

3 拼写形式多样化的原因浅析

蒙古文拼写形式多样化原因很多，主要有：

3.1 同形字母混用

如前所述，同形异音字母混用是拼写形式多样化的主要原因。其中尤其以词首o/u、词首ö/ü、词中/词尾的o/u/ö/ü、阴性词中的x/g、所有d/t、词中a/e、词尾a/e/n混用占绝大多数。这种混用更大是源于认识问题，还有一部分是源于确实分不清楚如何正确拼读。

3.2 分写词缀误录

蒙古文分写词缀视觉上与前导词分写，容易被误解为独立单词。所以让人们普遍理解、接受并规范输入需要一个过程。目前，录入分写附加成分的错误主要集中在不用控制字符202F和使用错误字母两点。例如“所有格”词缀 ᠶᠢᠨ 可以有-iin -ii -iie -jyn -jy -jye-yi -yie 和jin.....jie ji 等十几种错误拼写方。

3.3 控制字符误用

因控制字符是不可见字符，目前操作系统和编辑器又缺少控制字符查重 (Duplication Checking) 或过滤 (Filtering) 功能，乱用、误用情况在所难免。尤其是生僻且写法特殊的外来词，录入者可能反复交替试用几个控制字符后最终获得所需字形，但有可能录入了多余控制字符而浑然不知。即使是常用词也有可能中间插入多余控制字符而表面上看不出来。例如 ᠶ 后放置FVS1 (U+180B) 后变成 ᠶᠦᠰ1 ，之后可加入随意多个控制字符而不变形。所以理论上 ᠶ 可以有A+FVS1和A+FVS1+FVS2等无限多个拼写方式。

3.4 异常同形词使用

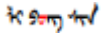
某些人会利用特定字体中形似字符来代替输入。例如，有“方正书版”系统用户利用“方正白体”的特点，将“属格”分写附加成分 ᠶᠢᠨ 直接用词首形辅音 ᠪ 来输入，在一般出版印刷中难以辨别，只有当切换到手写体或进行编码转换时才有可能暴露。而这些用户还会为自己的创新的简易录入而沾沾自喜。

3.5 字体瑕疵的利用

虽然目前各机构都声明自己的OpenType/AAT字体是符合标准，但部分字体显然并未完全执行国家标准、用户协定和转换规则最新版本，导致不同字体转换结果存在差异，破坏了字符序列到字形序列的“唯一性”约定。例如，因对用户协定理解不同或疏忽，有的机构OpenType/AAT字库规则中“辅音+w+辅音”条件下w (u+1838) 被转换为 ᠠ 形，而有的转换为 ᠡ 形。所以， ᠶᠢᠨ 中E被误录为w时，有些字体中显示了期望字形，而有些字体可能显示

⁰原文已转载至<http://www.mgyxw.net/mdls/am/amview.aspx?iid=138925&mid=7273>。

¹本案例数据源自确精扎布教授与作者联名提交的报告，由六月副研究员完成校对审核，因篇幅所限未付标注数据。

为  了。因为非专业人士不可能甄别出这么细致的错误和差异而将错就错地使用，导致使用真正正确的字体时反倒字形出错了。这种字体转换规则在国家标准[8]、用户协定和转换规则执行上的差异所带来的编码混乱危害性不比原来的编码不统一所带来的危害性小。由于这种错误具有一定的隐蔽性和不确定性，只有随着应用深入和数据累计才能暴露，而那时用户将更加迷茫和无措，所以其危害性更大。

4 拼写形式多样化泛滥导致的严重后果

如此泛滥的拼写形式多样化现象具有什么样的后果呢？

4.1 无法检索

无法对这样的文本直接进行各种检索 (Search)，哪怕是文本编辑器中的“查找、替换”等简单功能都是很不确定事情。例如上1.4述这篇文章中一词共出现6次，一致拼错为 ündüsüden，利用此篇文章的人除非恰巧有相同拼错习惯，否则就无法搜索到这个单词。依据本文穷举演算，一词至少有86400种拼写形式，MGLNews中实际出现273种拼写形式，所以不进行任何处理情况下，依靠内码匹配搜索几乎是不可能的事情。

4.2 无法排序

拼写形式多样化泛滥使得无法排序 (Sort)。这里指的排序是指用蒙古文字母表中的字母顺序排序。如果我们将MGLNews不做任何处理而直接排序，总共出现90485次的一词将分别位于273个不同位置。

4.3 无法统计

“检索”和“排序”是计算机最常用、最基本，也是最有实际效能的功能。不能检索和排序的直接后果是不能进行任何形式的统计，哪怕是最基本的字数、词数统计都成为大问题。泛滥的拼写多样化阻碍了成千上万篇文章当做动态资源直接进入“语料库”做进一步研究。这使得语言监测、网络语象等需要大量语料统计支撑的研究工作受阻。

更何况每个语言的统计计算都会有些个性需求。例如。蒙古文统计中分写词缀与前导词算作一个词更符合语言学要求，NNBSP (U+202F narrow no-break space) 就是用于连接分写词缀和前导词。由于部分独立单词字形与分写词缀同形，自动纠错难度较大，从而，统计准确率很难得到保障。工作量统计中分写词缀算作一个单词，也许有利于编辑人员稿费统计而故意为之。

检索、排序和统计等最基本的应用需求都难于得到满足，我们该怎么办呢？

5 拼写形式多样化问题的解决方案

既然有这么多问题，我们该如何解决呢？大致有两种解决思路。一种是要完全正确录入或同时对被搜索内容和搜索关键字进行拼写纠错，区分同形异音字母和词，达到精准搜索、精准排序、精准统计。另一种是通过额外算法解决同形异音字母的字形模糊匹配，达到模糊搜索和模糊统计，但这是否违反了标准编码制定初衷，纵容用户拼写错误了呢？对此，我们建议：

5.1 推广普及录入规范和标准，提高用户意识

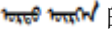
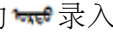

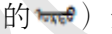
由蒙古文的自然属性、编码特性及拼写多样现象可知，让用户接受蒙古文标准编码的录入规范，形成良好习惯并不是简单的书写媒介更换问题，可以说是一次重大的变革。把握好了，我们可以借此机会将标准音与信息化同时推进。把握不好，将严重阻碍蒙古文信息化进程。为此我们认为，必须同时注重现有使用者培训和未来使用者培养，且后者更重要。未来使用者培养方面，将蒙古文标准编码的录入规范纳入中小学课程，结合标准音推广工作，从入学儿童开始培养起良好的习惯，待他们毕业走上工作岗位时，将按标准录入视为很正常的必然事件。现有使用者培训方面，我们先重点培训编辑、记者、相关蒙古文工作人员，再逐步扩大到普通用户步骤。单从报纸、期刊和图书出版角度考虑，要想让长期习惯于纸质出版做法的相关人员按规范录入具有一定难度。兼顾字形和读音会额外耗费精力，这明显有悖于他们追求绩效考核的初衷。那么如何让他们意识到这样做的好处并愿意付出这份努力也许是个不简单的系统工程问题。将出版资源升级为“语料库”，作为商品授权给研究机构及商业公司，获取经济效益等都不

一定能凑效。如果不能形成长效机制，很难长久维持。所以从他们自身使用便利角度考虑，让他们意识到这样做了可以便利利用以往资料等可能是更合理的方式。

5.2 使用智能输入法避免误录

在标准编码OpenType/AAT字库实现中，一般都会附带一个键盘映射（Keyboard Mapping）输入法。因复杂文本引擎和字体规则担负了名义字符到变形显现字形的映射转换，输入法本身一般只需从键盘字母映射为蒙古文字母，一般不用做额外处理。因蒙古文是个同形异音字符较多的文字，这种键盘映射输入法没有避免同形字符输入错误避免措施，从输出的字形又不易察觉错误，所以，即使用户了解标准编码框架、懂得输入规范，也意识到规范录入的必要性，但指望所有用户都能按标准录入是不现实的。为标准编码的推广普及，不让终端用户陷入迷茫的一个有效途径就是推广智能化程度较高的输入法做预防性处理。鼓励使用完全符合规范和标准的智能输入法，从录入源头避免错误[9]。智能化输入法确保录入字形和读音正确的同时给用户简单易用的用户体验，让用户不再感觉遵循规范是个负担。此处所述输入法不仅局限于全键盘录入，也包括智能终端的虚拟全键盘、数字键盘及OCR识别录入、语音识别录入等所有输入方式。这些输入方式上必须加以监督机制，尽量避免用户录入错误。不管输入法做到什么程度，总是无法避免OOV，而这部分的正确录入只能依靠用户自律或通过网络协同等方式作为弥补。

5.3 使用校对纠错工具后纠正

虽然通过前两项可以解决今后的问题，但对于历史数据或字形扫描、手写录入等场合，我们需要依赖自动校对和自动纠错。目前“词典+规则”是实现蒙古文文本校对常用方法[6]。不管使用不确定有限状态自动机（NFSA）数据结构获取较高计算效率、使用词干/词缀和生成规则来节省存储空间或是使用最一般的字符串匹配的库结构，其本质无非都是依赖词库，词库中有的词认为是正确词，词典中没有词（OOV）就认为是错别词。再进一步用搭配库或规则对部分同形异音词（例如  的  录入为  的 ）进行甄别。从公开资料来看，未登录词、同形多音词处理还不够成熟，句法和语义层面错误基本未能触及，甚至词法层面的形态变化分析[10]还有待提高，校对和纠错效果基本取决于词典词汇量。字形拼写错误的纠错也不尽如人意，所以有待进一步完善和改进校对和纠错。

5.4 探索基于生语料的统计学习方法

我们有了确保单词读音正确的熟语料，可以顺利开展一些统计建模的科研工作[11]，但目前我们所能获得量还难以支撑实际应用需求，更无法满足需要大数据支撑的个别模型。虽说可以采取各种手段缓解数据稀疏（Data Sparse），但归根结底还得需要足够量数据支撑统计建模[12]。很显然，深度机器学习（Deep Machine Learning）使用的词向量表示（Word Vector Presentation）来说，语料量越大，低维空间（Low Dimensional Space）上的词向量越趋于精准[13]。更何况熟语料没有读音错误只是一种假设，而我们日常产生的原始数据又有如此严重的拼写多样化，我们所能采取的防范措施又不能解决所有问题，所以我们不能单纯等待和依赖加工足够量的熟语料后再开展相关研究工作。另一方面，新词术语研究、语言动态监测、舆情分析等工作总不能还要依赖加工的熟语料。直接利用生语料的研究工作是熟语料建设的必要补充和回旋途径，相辅相成，互为补充，也是解决拼写多样化的一个重要途径。

6 总结

综上所述，蒙古文文本中大量存在拼写多样化现象，严重影响着蒙古文文本的日常应用及科研工作。各种解决方式都无法独自满足需求，需加以综合利用。各项工作的开展势必依赖语料库建设、知识库建设及相应大数据、机器学习等方面的突破，所以我们的研究工作还任重道远。

参考文献

- Aminul Islam, Diana Inkpen. 2009. Real-Word Spelling Correction using GoogleWeb 1T 3-grams. *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1241-1249, Singapore, 6-7 August 2009. c2009 ACL and AFNLP.

- Deniz Yuret, Ergun Bici. 2009. Modeling Morphologically Rich Languages Using Split Words and Unstructured Dependencies. *ACL-IJCNLP 2009 Conference*.
- Daniel Jurafsky, James Martin. 2010. *Speech and Language Processing*. 人民邮电出版社.
- Jacob Devlin, Rabih Zbib. 2014. Fast and Robust Neural Network Joint Models for Statistical Machine Translation. *ACL2014*1370—1380.
- The Unicode Consortium[EB]. <http://www.Unicode.org>.
- 敖敏,熊子瑜,呼和. 2011. 基于蒙科立输入法的蒙古语同形异码词研究. 第十一届全国人机语音通讯学术会议.
- 白双成, 张劲松, 呼斯勒. 2013. 蒙古文输入法输入码方案研究. *中文信息学报*2013(06):169-174.
- 确精扎布. 2014. 确精扎布蒙古文信息处理专辑. 内蒙古教育出版社.
- 国家质量监督检验检疫总局, 国家标准化管理委员会. 2011. GB 25914-2010.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. 中国标准出版社.
- 斯·劳格劳. 2013. 基于不确定有限自动机的蒙古文校对算法. *中文信息学报*, 2009,(06).
- 苏传捷,侯宏旭,杨萍等. 2013. 基于统计翻译框架的蒙古文自动拼写校对方法. *中文信息学报*, 2013,(06).
- S·苏雅拉图. 2001. 蒙古文整词计算机生成理论研究. *中文信息学报*,2001(04):59-65..
- 赵伟,侯宏旭,从伟,宋美娜. 2010. 基于条件随机场的蒙古语词切分研究. *中文信息学报*. 2010, 24(5).