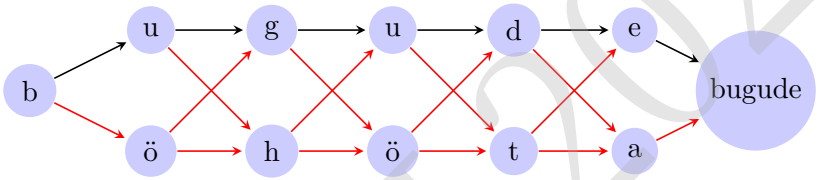
形多字”等自然属性，这种可被映射为相同T的C的组合非常庞大且非常常见（见下文分析）。蒙古文文本中存在的，这种单词输出（屏幕显示或打印）的“变形显现字形”所表现的字形正确，但其存储的“名义字符”序列（文本内码序列）不正确现象，我们称为蒙古文拼写形式多样化现象（Mongolian Spelling Diversity Phenomena），简称为拼写多样化。也有文章称其为同形异码词 (敖敏等, 2011)，是对同一事实的不同层次命名。

正由于蒙古文的拼写多样化现象，蒙古文拼写错误（spelling error） (Islam et al., 2009)有别于其他文种，可细分为“读音错误”和“词形错误”。依据字符序列C转换结果T的唯一性特性可知，当一个单词的词形T错误时对应的C也一定错误，所以词形错误时读音一定错误（详见2.5）。相反，词形正确却无法保证其读音正确。所以拼写多样化也可以叫“字形正确，读音错误”现象。从这个意义上讲，蒙古文的读音错误才是其他文字中所述拼写错误，但从其他文字经验和直觉而言，很容易理解为字形错误是拼写错误。所以我们也可以说蒙古文的拼写错误有“字形正确但读音错误”和“字形读音都错误”两个层次。本文中“词”是指单词名义字符序列，而“词形”是指其变形显现字形序列，所以同形词即指同形异码词。本文生语料统计计算中没有进行字形纠错和归并。

另需要说明的是，MIT编码（因蒙科立成果转化而用户习惯称其为蒙科立编码）） (白双成等, 2013)作为一种“全字符编码”方式，本身就是基于标准编码框架的一种变形显现字形方案。所以，不管是按名义字符形式保存的标准编码还是按变形显现字形形式保存的蒙科立编码，只要是基于“音”的编码形式就必然存在拼写形式多样化现象，本文分析结果通用于所有此类编码方式的文本。文章 (敖敏等, 2011)提到蒙科立编码文本语料库中存在同形异码词，但此文集中于用同形异码字符替换和符合字符拆分、组合方式归并同形词上，与本文目标具有较大差异。蒙古文自动校对 (斯·劳格劳, 2009) (苏传捷等, 2013)等也提及蒙古文同形异码，但都没有对此进行特别深入分析。



2 拼写多样化情况

2.1 简单的拼写多样化案例

我们先看一个简单且容易理解的拼写形式多样化现象。因ö/ü、x/g、d/t、a/e四对字母在相同词内位置和相同阴阳性（ ）条件下经常表现为同形，人们很容易就“发现”录入常用词 (bugude)时可随意替换这几个字母而获得所需词形，并至少可列举出图1所示32种拼写方法（其中只有实线路径是正确拼写）。原本在书面教学和日常使用中就很容易混淆的这些同形异音字母在一个单词中如此高频度组合出现，录入中混淆不足为奇。实际上，这个词形的可拼写形式远不止这32种。

2.2 基于例词的拼写多样化穷举

我们再以常用词 （undusuten）为例，试图穷举其所有可能的拼写形式。这次我们不仅要考虑上例所用同形异音字母之间的替换输入，还要考虑字形之间内含关系（ 被拆分为 ）和切分歧义 可能是 ，也可能是 ），甚至考虑古文用法字形等编码范围内的所有可能录入方式。据此我们可以理论上穷举出如表2的多种拼写方式。

其中音标后上标1和2表示那个字母的变体，就是变体选择符（Mongolian Free Variation Selector）的缩写形式。例如 a2表示字母a的词中变体，也就是古文用的单齿a。 n1表示词首不带点的n辅音（ Consonant）古文写法。 ö1表示ö的词中变体等等。

从表1可知，这个单词词形有多达(2+2*3+3*4*3+3*2)*3*2*4*1*4*2*3*3=86400种拼写形式。考虑到这个数字已经非常大，足以说明问题，也为了防止表格过于繁杂，我们没有再列举w辅音的 形等更为偏激的异常拼写方式。如果要考虑连续、交替、重复使用多个控制字符，可以稍微夸张地说每个单词有无穷多种拼写形式。

表 2 穷举ündüsüten一词拼写形式表

字母									
拼写形式	ö ü		a e n	d t	o u ö ü		o u ö ü	d t	a e n	a e n
	o									
	u									
	e a^2 n^1	o u ö ü	i j y^1							
	$ö^1$ $ü^1$									
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、HMTL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。

表 4 拼写形式统计节选数据

拼写种类	所属词形频度	所属词形数	平均频度	所属词例（top 5）
273	90485	1	90485	
179	5954	1	5954	
118	11472	2	5736	
……				
50	122088	29	4209	
49	181698	31	5861	
……				
1	310317	141006	2	

表 5 词 的部分字形相似词

单词	类型	原因	
	原词	名词“要求”	
	形近	动词“要求”	
	拼错	替换	辅音“ ”替换为辅音“ ”
	拼错	替换	辅音“ ”替换为“ ”
	拼错	替换	辅音“ ”替换为元音“ ”
	拼错	替换	辅音“ ”替换为辅音“ ”
	拼错	插入	多输入一个齿“ ”
	拼错	插入	重复输入辅音“ ”
	拼错	插入	多输入一个齿“ ”
	拼错	插入	多输入一个齿“ ”
	拼错	插入	多输入一个齿“ ”
	拼错	插入	多输入一个齿“ ”
	拼错	删除	少输入一个结尾元音 A
	拼错	删除	少输入一个辅音 R
	拼错	删除	首音节少输入一个元音 A

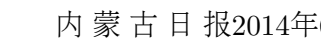
将MGLNews语料所有单词按字形进行归类后获得了所有词形统计数据，表4节选展示了部分典型数据。

表4第一行表示，词形 在语料中共出现90485次，有273种不同拼写方式，是拼写形式最多的词形。第三行表示 和 这两个词形各有118种拼写方式，共出现11472次，每个词形平均出现频度为5736次。最后一行表示141006个词形只出现了一种拼写形式。如前所述，语料中不仅存在读音错误，也有字形错误，表3统计过程中我们并没有进行错误字形纠错。例如，表5展示了MGLNews中单词 的部分词形相似词，其中只有一个是形近字，其余都是拼写错误。

实际数据表明，蒙古文拼写形式混乱程度远超我们之前的估计[3]。这种混乱不仅有理论依据和发生概率，实际使用中人们也确实经常犯这种错误，而且很多用户根本没有意识到自己的错误所在。更何况还有表5所示拼写错误也非常多，这给我们的数据利用形成了非常大的阻碍。

从统计数据分析可知，拼写多样化基本与单词组成字母的同形字母个数及字形组合可能性而定。显然单词越长，拼写形式多样化概率越大。常用词拼写形式个数略显多一些的原因仅仅是因为常用词频度较高，其各种拼写形式出现概率高了而已，相信随着语料量的增大，“大数法则”凸显后，这一现象渐渐平稳并趋于稳定。

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  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  用t输入，字形 都用d输入；③阴性词里的x、g都用g录入；④词末的 不管阳性、阴性都一律用u录入；⑤不用分写词缀( )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 yin可以有-iin –ii -iie –ǰyn –ǰy -ǰye-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 后放置FVS1（U+180B）后变成 ，之后可加入随意多个控制字符而不变形。所以理论上 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 -u直接用词首形辅音 b/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 形，而有的转换为 形。所以， 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的事情。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收入“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研究[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，相铺相成，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研究。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. *JProceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pages ,Singapore, 6-7 August 2009. c2009 ACL and AFNLP*,1241–1249.

Deniz Yuret, Ergun Bicici. 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).