蒙古文输入法输入码方案研究*

白双成[123] 张劲松[1] 呼斯勒[23]

([1]北京语言大学 对外汉语研究中心，北京 100190;
[2]内蒙古社会科学院 蒙古语信息处理研究所，呼和浩特 010020;
[3]内蒙古蒙科立软件有限责任公司，呼和浩特 010019)

摘要：科学合理的输入码方案对一个输入法至关重要。通过输入码重码量分布和平均码长等量化指标，综合分析比较了蒙古文读音输入法可使用的三类七种输入码方案，提出了以音节为编码单位的支持模糊输入的输入码方案，应用于项目组新版输入法中获得推广普及。试验结果和推广应用经验表明，新输入码方案顺应人的思维和记忆的同时可保证较高的录入速度。

关键词：蒙古文；输入法；输入码；模糊输入

中图分类号：TP391 **文献标识码**：A

A Comparison Study on Word Coding Methods for Mongolian IME

BAI Shuangcheng[123] ZHANG Jinsong[1] Husile[23]

([1]Beijing Language and Culture University, Beijing 100190, China;
[2]MIT Center, Inner Mongolia Academy of Social Science, Hohhot, 010020, China;
[3]Inner Mongolia Menksoft Software Co., Ltd., Hohhot, 010019, China)

Abstract: Word coding, representing the mapping between a word and a series of keyboard inputs here, is very important for efficient Mongolian IME. Based on the criterions of candidate duplications and average word coding length, this paper made a comparison study on the efficiencies of 7 kinds of coding methods belonging to 3 classes, and proposed a new syllable based Fuzzy input method. Experimental results showed that the method is not only easy for users to memorize, but also very efficient to use.

Key words: Mongolian; IME; composition string; fuzzy input

1 引言

依据蒙古文编码国际标准[1]和国家标准[2][3]，在Windows Vista/7/8等具有基本满足蒙古文特性的复杂文本布局引擎（complex text layout engine）的平台上，利用OpenType技术[4][5]即可实现蒙古文名义字符到变形显现字形的转换，并通过一个简单的键盘映射输入法即可实现蒙古文录入，满足基本的应用需求。为了更好地满足不断发展中的用户需求，再开发一套智能输入法是非常有必要的。对于一个录入速度和准确度要求较高的智能输入法而言，研究制定科学合理的输入码方案至关重要。

我们将输入一个单词所需的输入码的码元个数称为码长[5][6]。就词级录入来说，蒙古文输入法的录入速度取决于词的平均码长，其中含两层含义，一个是对应一个单词的码长越短越好，另一个是输入这个输入码时用户思考寻找的时间越短越好。输入法的码长越长，输入信息越多，对应重码词就越少。反之，码长越短，所含信息越少，对应重码词就越多。单纯地减少码长未必能提高速度，因为寻找一个键的时间可能变得较长[5][7]。所以缩短平均码长的关键在于兼顾两者，输入码方案确定就是在码长与信息量之间寻找一个平衡点的过程。需

*收稿日期：2013.7.15 定稿日期：

基金项目：国家电子发展基金 2010 年度、2011 年度蒙古文专项；国家自然科学基金（61163020）；内蒙古自治区自然科基金项目 2011MS0918 资助项目

作者简介：白双成（1974—），男，博士研究生，研究员，主要研究方向为自然语言处理；张劲松（1968—），男，博士，教授，博士生导师，主要研究方向为基于语音语言处理技术的计算机辅助汉语教学技术的研究工作；呼斯勒（1977—），男，助理研究员，主要研究方向为自然语言处理。

要特别说明的是,好的方案能否落到实处，还要看具体工程实现能力和实际条件。

以往，确定输入法输入码方案时更多的是研发人员从感官认识和个人的主观意识去选择一套方案。为此我们想通过本文较为系统的量化比较各类输入码方案，为下一步的研究和工程实现提供理论依据。

2 部分约定

1) 蒙古文输入也可以用字形、字符编码等信息来输入，但这类输入方式毕竟不是普遍做法，本文讨论的输入法，只限于读音输入（类似于汉字拼音输入），也就是按蒙古文读音为基础的输入法。

2) 限定在PC普通键盘上，不涉及数字键盘或其他类。组成输入码的字符集(通常称为码元)默认为英文字母{abcdefghijklmnopqrstuvwxyz}，也可以是普通键盘内的其他所有字母，如蒙科立整词输入法[8]中“NG”用“;”输入。码元越多，重码可能性越小，输入就会更快捷，但会加大用户记忆压力。本文虽不涉及数字键盘，但结论对其有一定的借鉴意义。

3) 本文比较的输入码方案中，要求能够分辨蒙古文同形异音字母ᠣ(o或u)、ᠥ(ö或ü)，ᠭ（h或g）ᠳ-ᠲ（d或t）等。此文暂不涉及此类同形异音字母的模糊输入问题。支持同形异音字母的模糊输入后，无非就是定制一套新的输入码比较算法，计算思路不受影响。

4) 因本文比较的输入法都是蒙科立系列产品，文中蒙古文拉丁标注中采用蒙科立输入法键盘布局方案，与国家标准键盘布局略有不同。具体来说用ᠠ-a、ᠡ-e（ᠧ-E)、ᠢ-i、ᠸ-c、ᠸ-v、ᠣ-o、ᠤ-u分别代表七个元音，用ᠨ-n、ᠪ-b、ᠫ-p、ᠬ-h、ᠭ-g、ᠮ-m、ᠯ-l、ᠰ-s、ᠱ-x、ᠲ-t、ᠳ-d、ᠴ-q、ᠵ-j、ᠶ-y、ᠷ-r、ᠸ-w、ᠩ-;代表17个基础辅音，用ᠹ-f、ᠼ-z、ᠽ-Z、ᠺ-H、ᠻ-L、ᠿ-R代表6个扩展辅音。

5) 以词组为单位统计的单词的信息熵保证比以单词为单位的信息熵小，也就是说，以词组为单位进行输入码编码的话，其输入码长度会小很多。如果进一步考虑单词及词组之间的上下文相关性，建立统计语言模型的话信息熵会更小。但本文暂不涉及这些内容。因为我们通过n元模型试验得知，只要单词输入码方案选定合理，直接应用于词组及语言模型，即可获得较好的效果。因篇幅所限，暂不细述。

6) 如果加入单词频率因素，量化比较会更合理些。但碍于目前所能获得的熟语料量严重不足，生语料利用中的部分技术问题还未完成，所以本文暂时无法考虑词频因素。

7) 输入码方案中优先考虑输入的简短快捷和易学,但本文提出的模糊输入特性恰恰是标准音教学和推广中所不能容忍的。我们认为可通过输入法选项来限定,就此不做额外说明。

3 预比较方案简介

按输入码方案的特点，将其分为三类，即方案1为全拼方案，方案2和方案3为整词方案，其余为模糊输入方案。各方案介绍如下：

方案 1：全拼方案

该方案是目前多数读音输入法（或称其为音码输入法）采用的输入码方案。其核心思想是依据拼音（音素）文字特点，为每一个音素安排一个输入键（实际应用中可能会用 n+g 来输入ᠩ（ŋ），用 lh 输入ᠯᠧ（l）等策略）, 通过上下文的分析，尽力解决一字多形，多字同形（同形异音）问题。例如，音素ᠠ（a）的输入键为 a，则不管其词内位置如何，都用这个 a 键输入,如ᠠᠭᠠ的输入码为 agana。为区分同一音素在相同上下文中的多个变体，引入变体选择方案（用词典或规则方式特殊处理除外，此处特指此类特殊问题），如蒙科立音码输入法 2002 版中用 1、2、3 选择不同变形（如ᠪᠠᠨᠠ输入为 banana1），用-a、-e 选择词尾 a、e 变形（ᠬᠠᠷᠠ hara\ᠬᠠᠷ har-a）。后续版中引入穷举候选项和大写输入码（也就是 shift+键值）输入非常用变体字形方式。基于 OpenType 技术实现的键盘映射可归为此类。

方案 2：整词输入法方案

其输入规则是取首音节（词首第一音节）的辅音、元音及第一个半音字母(ᠭᠡᠳᠡᠰᠦ ᠬᠠᠭᠠᠳ, 也称为音节末辅音)，再取其他音节的首字母（多数情况为辅音，只有双词根词中后词为元音开头时为元音）和词末半音字母即可。为了能够区分输入码中的普通辅音和半音字母，采取了表 1 所示半音字母映射输入技巧。如ᠠᠷᠰᠯᠠᠨ（ARS+LAN）为双音节词, 第一音节取首元音 a 和第一个半音字母 r，忽略掉第二个半音字母 s，取第二音节首字母 l 和词末半音字母 n，所以输入码为 arln，为了区分普通辅音和半音字母，将 r 映射为 e，n 映射为 z，最终输入码为 aelz。为此用户需要记住这几个半音字母的输入码(如下表)，其余与全拼一样。

表 1 整词输入法半音字母映射关系表

半元音字母	d	r	s	l	m	b	g	n		u	i
对应输入码	a	e	c	v	o	p	k	z		u	i
例词	ᠳᠠ	ᠡᠷ	ᠡᠰ	ᠡᠯ	ᠡᠮ	ᠠᠪ	ᠡᠭ	ᠡᠨ		ᠤᠭᠤ	ᠡᠴ
读音	ed	ar	as	al	am	an	ag	cn		huu	ai
输入码	ea	ae	ac	av	vo	ap	ak	cz		huu	ai

这一输入码方案的好处是输入码中音节结构清晰，基本不存在切分歧义。每个单词只有一种输入码，反过来给定一个输入码, 其音节结构基本只有一种。例如ᠠᠷᠰᠯᠠᠨ输入码为 agn，ᠪᠠᠨᠨ 的输入码为 bann，反过来 agn 只可能是 a、g、n 开头的三音节单词的输入码，bann 只可能是 ba、n、n 开头的三音节单词的输入码。其最大问题在于用户需要记忆上表所述对应关系，类似汉字的五笔字型输入一样，虽然缩短了输入码长度，但录入过程中用户要切换思路，降低了击键速度和提高了入门门槛。

方案 3：在方案 2 的基础上要求补充所有半音字母。提出这一方案的目的仅仅是想在数据上比较一下方案 2 以外的半音字母对重码过滤的贡献度。实际应用中录入复杂度已接近方案 1，还要记忆半音字母对应关系，无实际应用价值。

方案 4：只取各音节的首字母，忽略所有其他字母。此方案将在词组输入和基于统计语言模型输入中具有重要作用。

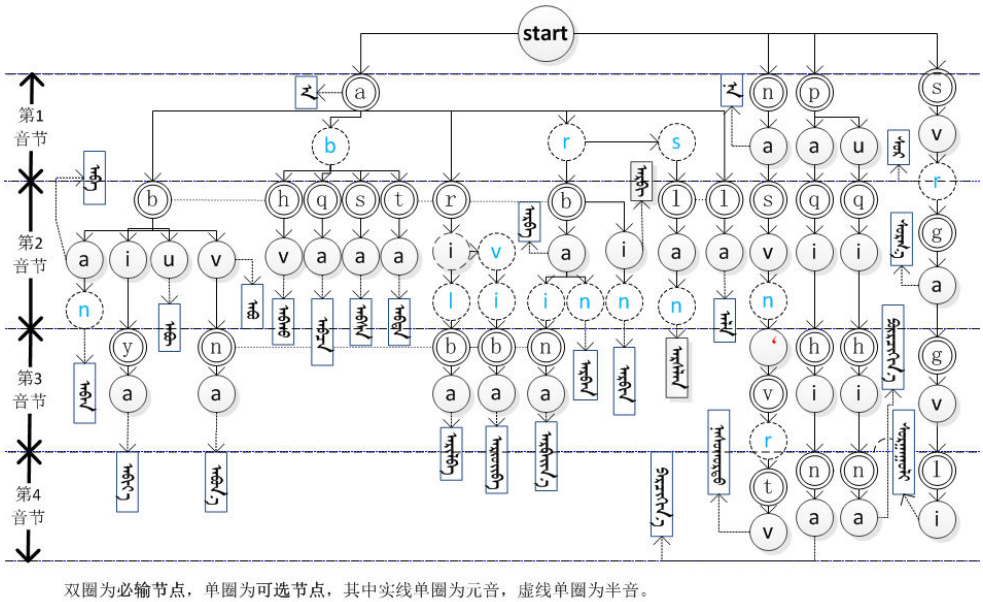

图 1 模糊输入方案示意图

方案 5：取各音节首字母，再取首音节元音。此方案主要观察首音节元音贡献度。

方案 6：取各音节首字母，再取词尾半音字母。此方案主要观察词尾半音字母贡献度。

方案 7：方案 5、6 之和，与方案 2 类似，只是不再对半音字母进行映射。也就是说，辅音和半音字母统一按其读音相似取相同的输入码，输入规则也是从首音节取首辅音+元音+第一个半音字母，其他音节取首字母，再取词尾半音字母。如  的输入码直接取 arln。与方案 2 相比，显然会减轻用户负担，但其输入码中不再含有清晰的音节结构信息，导致输入码歧义增多，候选也将随之增多。按照此方案规则，r 和 n 已经分不清楚是普通辅音还是半音字母，为此 arln 的音节切分可能为 a+r+l+n/ar+l+n/ar+ln 三种，较方案 2 相比，候选项中将多出 a+r+l+n/ar+l+n 两种音节切分对应候选项。

4 量化比较

试验中使用的是 1215799 条词。

表 2 各方案重码分布表

方案 1：全拼方案										
重码	1	2	3	4	5	6	7	8	9	9+
频率	998668	72599	23858	60	0	20	0	0	0	0
比例	**91.1855**	6.6288	2.1784	0.0055	0.0000	0.0018	0.0000	0.0000	0.0000	0.0000
方案 2：蒙古文整词输入法输入码方案										
频率	791699	130927	33138	8336	1169	2553	283	211	282	185
比例	81.7210	13.5146	3.4206	0.8605	0.1207	0.2635	0.0292	0.0218	0.0291	0.0191
方案 3：方案 2+补齐其他半辅音										
频率	935763	95704	24743	2021	122	824	15	10	48	12
比例	**88.3410**	9.0350	2.3359	0.1908	0.0115	0.0778	0.0014	0.0009	0.0045	0.0011
方案 4：只取音节首音										
频率	144348	77461	37743	27555	10675	13745	3775	6217	3788	22757
比例	41.4717	22.2548	10.8437	7.9166	3.0670	3.9490	1.0846	1.7862	1.0883	6.5382
方案 5：音节首音+首音节元音										
频率	262747	124096	56578	36510	12050	13908	2399	5173	3567	10529
比例	49.8045	23.5228	10.7245	6.9206	2.2841	2.6363	0.4547	0.9806	0.6761	1.9958
方案 6：音节首音+末尾半辅音										
频率	333108	107836	44186	24104	12475	11666	5727	4621	3163	13116
比例	59.4834	19.2564	7.8903	4.3043	2.2277	2.0832	1.0227	0.8252	0.5648	2.3421
方案 7：音节首音+首音节元音+末尾半辅音										
频率	553409	145392	50190	20926	6612	7309	1460	1509	1433	1983
比例	**70.0320**	18.3989	6.3514	2.6481	0.8367	0.9249	0.1848	0.1910	0.1813	0.2509

从表中分析可知

全拼方案类-方案 1:

1) 只需了解键盘布局即可会拼写，易学性突出；
2) 拼写所有可能组合，没有拼不出来的字，是其他方案的必要补充；
3) 高达 91.1855%的无重码率，重码集中前 3 项，便于盲打；
4) 音节切分没有歧义；
5) 只需规则推导，不用词表也可以；
6) 输入码长，效率低；
7) 没有词库支持错误率较高；

整词方案类-方案 2 至方案 3

1) 平均码长比方案 1 短很多，但无重码率还是高达 81.7210，重码分布与方案 1 接近；
2) 音节切分基本没有歧义；
3) 需记忆对应关系，按人类学理论，易中断录入者的思维过程，不符合人的自然行为；
4) 录入未登录词时，需要在方案 1 与方案 2 之间进行切换；
5) 必须有词表支持；

模糊方案类-方案 4 至方案 8

1) 平均码长接近方案 2，但与其相比，输入更自然流畅，符合自然行为；
2) 补充信息越多重码率越低，输入灵活，便于读音不确定词的录入；
3) 与方案 1 结合录入未登录词更为自然，无需切换；
4) 利于用更为简短输入码用于基于统计模型的连续输入；
5) 失去音节结构，重码率升高；
6) 输入灵活，但同时输入随意性增大，盲打难度提高；

5 平均码长

为进一步用一个单值直观比较输入码方案的特性，在此借用平均码长概念。长度为 len 的输入码有 n 个重码时，我们还需要用一个空格或数字键选择，所以每个词需击键 len+1 次。我们认为按空格选择第一候选项和按数字键选择其他候选项具有决然不同的消耗，以至于放大其影响而忽略空格输入时，平均击键数可公式化为（len*n+n-1）/n=len+1-1/n。例如同样是码长为 2 的输入码，如果重码为 2，则平均码长为 2+1-1/2=2.5,如果重码为 3 时平均码长为 2+1-1/3=2.66。所以总平均码长公式为：

$$\sum_{i=1}^{N}(len(i)+1-\frac{1}{C(i)})$$

其中 len(i)是第 i 个输入码长度，C(i)为输入码 i 的重码词个数。

需要说明的是当 n 大于阈值时，还需要翻页。但我们比较的几种输入码方案中 10 以上重码量较少，为简化公式，没有对其进行细化。

表 3 平均码长对照表

平均码长 \ 输入码方案	1	2	3	4	5	6	7
13.3241648633558532	+						
9.1610102127707780			+				
8.1045034123279134							+
7.8463315321564586		+					
7.6303673915450299						+	
7.5787450455590584					+		
7.1276000965339712				+			

从表中可以看到，我们方案 6 和方案 7 之间可以找到一个近似方案 2 的平均码长。

需要说明的是，表 3 所示数据中，没有考虑单词实际出现频率，加上计算单词中涵盖大量蒙古文形态变化派生的单词，所以平均码长要比现实文字录入的平均码长长很多。

6 总结

基于本文试验,我们在新输入法中采用了融合方案 4 至方案 7 的改进型混合输入码方案，即，用户必须输入各音节首字母，其余字母都可以随意忽略。通过本试验我们明确了所采取

输入码方案的优缺点。为进一步研究与工程实现奠定了基础。

除上述量化比较中凸显的优点外，此方案还有如下优点值得阐述：

1） 输入中可避开读音易混淆字母

如输入hohe时，如果不确定第一音节元音到底是第六元音 o 还是第七元音 u 时，省略掉这个元音，补充 e 元音输入为 hhe 即可。又如，torhu 这个词中有既有 o 也有 u，无法确认时可以省略掉这两个元音，补充第一音节末尾的 r，输入为 trh 即可让此词出现在首屏，此时依据输入提示输入下个字母 u 让其成为首选项。

2) 有助于输入外来词和特殊词拼写

如amErika 输入为 amrk，ewurupe 输入为 ewrp，prcgram 输入为 prrm 等。

值得说明的是，输入码中的各成分的权值不一样，更为自然的方式是测定各信息的熵[9][10]，通过熵来确定单词的最短输入码更为合理，但这种方式推导出来的输入码仅仅是满足最小化输入码的需求，会增加人们输入码记忆难度，实际意义需要研究。如果与上述分析类似，将输入码分为音节首字母、元音、半音字母等，通过计算熵来验证上述结果，对词级编码方法的研究可能有意义。

更进一步，以上所有分析源于一个思想，认为蒙古文读音输入法的编码方案的不同，实质上是从音素、音节、单词、词组等不同层次进行编码的结果。从音素层次，对每个音素进行编码，就是全拼方案。而其余都是从音节层次（更准确说是抽象音节结构角度考虑，而不是穷举所有音节结构）进行编码。基于单词及以上单位的编码方式目前不太适合蒙古文。这个需要合适的语言模型及语料库的支持。

参考文献

[1]. The Unicode Consortium[EB]. http://www.Unicode.org.

[2]. 确精扎布. 蒙古文编码[M] . 呼和浩特: 内蒙古大学出版社, 2000.

[3]. 国家质量监督检验检疫总局，国家标准化管理委员会.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[S]. 北京.中国标准出版社，2011.11.

[4]. OpenType specification[EB]. http://www.microsoft.com/typography/otspec/ .

[5]. 姚延栋, 吴健, 孙玉芳，呼斯勒. 传统蒙古文变形显示机制研究与实现[J].中文信息学报,2005, 18(5):84-89.

[6]. 朱巧明、李培锋等.中文信息处理技术教程[M].北京.清华大学出版社，2005.9.

[7]. 吴军.数学之美[M].北京.人民邮电出版社.2012.6:185-196.

[8]. S·苏雅拉图.蒙古文整词计算机生成理论研究[J].中文信息学报,2001,(04):59-65.

[9]. 那日松,淑琴.蒙古文信息熵和拉丁转写研究[A]. 中国计算技术与语言问题研究——第七届中文信息处理国际会议论文集[C], 2007:782-785.

[10]. Thomas M.Cover, Joy A. Elements of information theory[M]. New York. Wiley.1991.