基于 Windows 平台的滇南彝文输入法实现①

胡 刚[1,2], 王嘉梅[1,2], 张建营[1,2], 孙善通[1,2], 汤 雪[1,2], 赵慧云[1,3]

[1](云南民族大学 云南省高校少数民族语言文字信息化处理工程研究中心, 昆明 650500)
[2](云南民族大学 电气信息工程学院, 昆明 650500)
[3](云南民族大学 民族文化学院, 昆明 650500)

摘 要: 彝文古籍数量繁多, 但却在迅速减少, 亟待运用科学手段进行抢救、整理、规范. 为满足考古工作者和出版业界的需要, 从排版系统中对中英彝文混合输入的实用性考虑出发, 采用 Unicode 国际编码标准, 且字库编码区间实现从自造区间到汉字区间的跨越. 就 Windows 操作系统中进行滇南彝文字库设计及输入法实现所涉及到的原理及若干问题进行了阐述, 并在此基础上利用多多输入法 API 进行程序开发, 在克服了操作系统应用环境、字库编码区间及方式、文字混合显示等技术障碍后, 最终开发出"滇南彝文数字化信息处理平台". 无论是从可行性和实用性分析, 该平台的第三套"滇南彝文自由拆分模式的一对多形态编码输入法"在很大程度上解决了当前彝文输入法所普遍存在的易学性问题, 是一款操作稳定、输入快捷、简单易学的彝文输入法, 尤其对彝文信息化的推广具有广泛的使用价值和示范价值.

关键词: 滇南彝文; 形似编码; 编码区间; 字库; 排版系统

Implementation of South Yunnan Yi Language Input Method Based on Windows Platform

HU Gang[1,2], WANG Jia-Mei[1,2], ZHANG Jian-Ying[1,2], SUN Shan-Tong[1,2], TANG Xue[1,2], ZHAO Hui-Yun[1,3]

[1](Colleges and Universities in Yunnan Minority Language Information Processing Research Center, Yunnan Minzu University, Kunming 650500, China)
[2](School of Electrical & Information Engineering, Yunnan Minzu University, Kunming 650500, China)
[3](School of Nationalities & Culture, Yunnan Minzu University, Kunming 650500, China)

Abstract: The number of Yi language ancient books is large, but it's in rapidly reduced, the need to use the scientific method to rescue, organize, specification. To meet the needs of archaeologists and publishing industry, from practicability consideration of Chinese, English, Yi language mixed input in the composing system, using Unicode international coding standard, and realizing the character encoding range from the self-created interval to Chinese interval across. This paper expounded the South Yunnan Yi font design and input method realization involved principles and many issues based on Windows operating systems, and on this basis, use Duoduo-input method API for application development, after overcoming technical obstacles such as the operating system application environments, the font encoding interval and mode, text mixed display, Finally developed the "southern Yunnan Yi digital information processing platform". Whether it is from the analysis of feasibility and practicality on this platform, the third style "Southern Yunnan Yi language free split mode one-to-many shape code input method" to solve learnability problem of the current Yi character input method. it's a stable operation, enter quick, easy learning Yi input method, especially for the promotion of Yi language information technology has widely value in use and application.

Key words: South Yunnan Yi; fractal coding; code interval; fonts; publishing system

① 基金项目:国家自然科学基金(61363085);云南省教育厅科学研究基金重大专项项目(ZD2013013);云南省东南亚南亚西亚研究中心招标课题一般项目(DY2014YB01);国家语委重大科研项目(WT125-61);云南民族大学研究生创新基金科研重点项目(2014YJZ09)

收稿时间:2015-03-28; 收到修改稿时间:2015-05-04

没有信息安全就没有国家安全(习近平). 随着互联网在彝族地区的蓬勃发展, 网络的使用已经成为生活在云南和四川等地区的彝族人民生活中不可缺失的一部分,而这种无形的“信息边疆”安全问题变得日益突出[1]. 彝文信息化技术的跟进是解决当前边疆地区信息安全问题的关键和基础, 其技术难点是在计算机中如何实现彝文编码、输入、显现、存储和传输等问题.

目前,由于历史的原因,彝文信息处理技术与中文相比还存在较大的差距. 多年来彝文处理系统重复开发, 彝文编码标准和编码区间不统一、字库标准不统一、输入法不统一、操作系统应用环境单一,已严重制约了彝文信息技术的发展[2].

彝文信息化建设目前面临一系列亟待解决的问题, 尤其是基于“标准”、“通用”的彝文处理软件及其便捷、高效的输入法问题,特别是我国的少数民族语言文字, 由于其文字的特殊性, 对于大多数不懂少数民族语言文字的使用者来说仍然是一个关键性的问题, 或是“瓶颈”之所在[3]. 本文研究内容是面向滇南地区的彝文, 将彝文字符集编码纳入国际化标准, 并将彝文字库编码区间限定汉字编码区, 目的旨在实现多应用环境下具有文字排版功能的滇南彝文数字信息化处理平台, 这对规范彝文处理系统的开发标准具有重要的现实意义.

1 平台简介

本文开发的滇南彝文数字化信息处理软件平台由“云南规范彝文数字健笔画拆分模式的形态编码输入法”和“云南规范彝文自由拆分模式的一对多形态编码输入法”及“滇南彝文自由拆分模式的一对多形态编码输入法”三套软件操作系统平台组成.

本课题组以 8223 个古彝文字符为基础建立大型字库, 包含云南规范彝文和滇南彝文两种字体, 其中规范彝文字体采用黑体, 滇南彝文采用手写体, 均基于 Unicode 国际标准编码设计. 基于云南规范彝文的 2 种编码标准输入法操作系统平台均可输入 3749 个(含云南规范彝文 2500 多和四川规范彝文 1200 多)彝文字, 滇南彝文输入法操作系统平台可输入 10212 个彝文字(整合后有补充), 三套彝文输入方式均采用简单高效(规则简单, 不用死记硬背, 十多分钟能学会)快速的自由拆分编码和数字健笔画码组成, 适合不同层次、不懂读音、不懂彝文字的用户使用.

在 Windows 系列的计算机操作系统中只要分别安装上我们编制的三套彝文数字化信息处理软件平台中的任何一套注册软件和输入法安装软件, 以及我们开发的彝文 TrueType 字库软件, 就可分别实现三套彝文输入法.

第三套彝文输入法操作平台中经过多次改进和优化, 具体有如下几点:

①对规范彝文和手写彝文字库中的个别文字细节进行形态修正, 使其更接近文字的古原貌;

②字库量的提升方面, 在整合规范彝文和手写彝文的字库后, 对字库进行了后续补充, 实现彝文字库“万字”的突破, 便于查询和输出显示;

③突破了对操作系统应用环境的局限, 大大提升了用户使用的便捷性. 不仅能安装在 XP 操作系统上, 还可用于 Win7、Win8、Win8.1 操作系统, 理论上还可用于微软即将发布的 Win10 操作系统;

④在输入法安装包上, 提供 32 位和 64 位供用户选择, 尤其 64 位输入法应用于 64 位操作系统, 软件使用的兼容性和运行性能进一步提升. 安装和卸载操作过程不再局限于控制面板下的程序卸载;

⑤个别不易识别形态的彝文字体, 字符编码采用“一对多形态编码”规则, 在输入时提升状态栏的候选字出现频率支持多种词频调整策略, 并可在自动调频开启时, 固定部分字词不参与调频;

⑥ 基于多多输入法 API 进行输入法开发相比于前两款运用 Windows IME 结构生成, 输入法设置更加人性化. 用户可自行修改软件使用界面和属性设置, 且便于后续增添候选字库和修改主码字库;

⑦为保护“滇南彝文数字化信息处理软件平台”的知识产权, 规范本软件使用范围和应用权限, 在输入法的安装时我们设置了使用密钥和安装协议;

⑧在软件的易学性上做了进一步改善, 并制作了对应的使用说明书, 只要首次配合说明书熟悉输入规则, 在后续即可按“读字规则”实现快捷输入;

⑨为进一步加快彝文文字的推广, 其中的第三套滇南彝文输入法操作系统平台, 可应用于中、英、彝混合输入的文字排版系统. 该软件可在彝文古籍整理、文献出版、学术研究、教材编译和建设及学校双语教育等中应用, 还可制作各级政府单位部门及学校名称、门牌、政府公文、信封等彝汉双文书写等.

2　设计及开发

以下选取平台中的第三套滇南彝文输入法软件操作平台阐述其技术原理及其实现过程，具体流程如下:

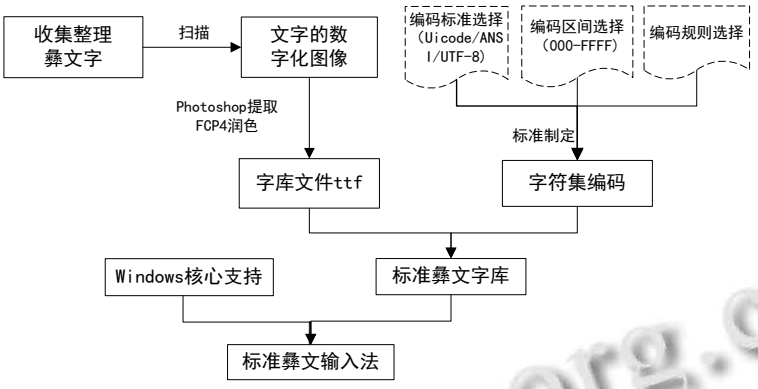

图 1　输入法制作流程图

2.1 字库制作

(1)字库制作: 字库的设计和制作是输入法开发中最基础的工作，而核心关键是语料字样选取、字符集编码设计，这两项是决定输入法性能的主观因素，而软件应用环境和输入法开发方式是决定输入法性能的客观因素. 如何利用计算机制作字库一般要经过设计字稿、扫描输入、程序拟合、人工修饰、质量检查、组织字库、测试字库、安装使用等几个步骤，彝文字库也不例外.

绝大多数的字库都源自设计字稿，字稿中的字体要求清晰、光滑、视觉效果良好，全部的字稿必须采用统一的规范才能够达到制作字库的要求. 利用扫描仪或其他输入设备将字稿图片输入电脑，要求套框取字一丝不差，而且要根据不同尺寸和清晰度的字稿灵活地调整扫描分辨率和其他相关参数，以达到不失真、反映出原字稿的全貌. 运用 Photoshop 作图工具提取扫描后的数字化图像，使用 FCP4 进行文字图像边缘化润色，通过修改字体、写入字体版权、控制字体属性等，最终形成 ttf 后缀格式的字库文件. 或者在 FCP4 下做字,之后将原有字库模板的情况下，将我们的字库粘贴进去之后进行修改大小位置(高: 10—210 宽: 10—230). 下图示 2 是为完成后的字模图像

(2)字库编码区间: 在编码区间选择上，多年来，计算机普遍采用 ASCII 码来表示字符(字母、数字、标点符号和控制符). 但 ASCII 编码字符集是一个高度受限的编码字符集,用这种编码来表示英文字符是不成问题的，但对于不同语言的所有字符(如中文和日文)，不能表示科学符号,更不能表示古代文字(神秘符号和象形文字)和音乐符号[4]. Unicode 编码是 ASCII 字符码后的一种新字符编码,基于 Unicode 的系统允许自行使用 65000 个不同的字符，足以覆盖各种语言的所有字母，外加数千种符号此外,还保留了约 18000 个未用的编码值以供将来使用. 根据编码字符集所依托的体系结构及原有规范彝文内码设定情况(彝族文字区 A000～A4CF 内、收容中国西南彝族文字和字根),同时考虑到系统的可扩充性，标准彝文字库选择 Unicode 字符集[5]. Unicode 编码空间分配区[6](16 位)如图 3 所示:

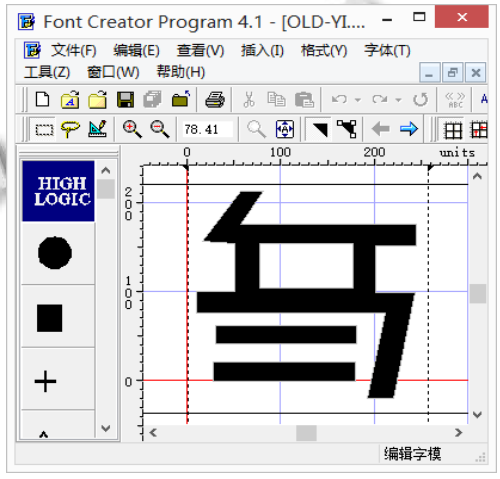

图 2　完成的字模图像

彝文字库若是采用用户自造字编码区，不会因占用已定编码空间而覆盖其他文字，仅仅临时用来测试字库和输入法的安装和使用. 但是存在的问题是在排版系统为保证对中、彝、英文的支持，必须将彝文字库编码区限定在汉字编码区内. 在字库中添加彝文字还

有一种方法,就是统一到汉字编码区添加彝文字,即使覆盖占用汉字编码空间, 只要切换不同的输入法可混合键入中、彝、英字, 选择对应的字体, 目的文档就可显示[7]. 故本课题组使用的是汉字编码区, 选择自造编码区的彝文字库对 Windows 自带楷体字库(默认汉字编码区)进行覆盖修改,即可转换成汉字编码区内的彝文字库. 以下图 4-6 鉴于篇幅考虑, 省略中间部分字库.

拼音文字编码区	0000……33FF
汉字扩充编码区	3400……4DB5
汉字编码区	4E00……9FA5
规范彝文编码区	A000……A4C6
韩文（古彝文）编码区	AC00……D7A3
UTF-16 代理区	D800……DFFF
用户自造字编码区	E000……F8FF
限制使用区	F900……FFFF

图 3 Unicode 编码空间分配区

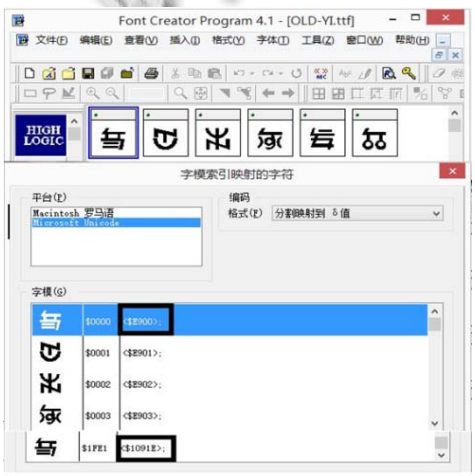

图 4 非标准 ttf 字库(自造编码区间 E900-1091E)

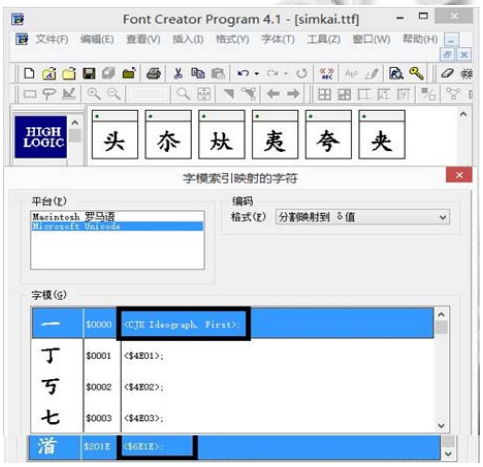

图 5 自带楷体 ttf 字库(修改编码区间 4E00-6E1E)

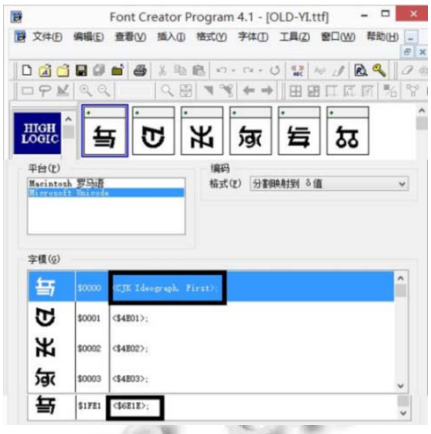

图 6 标准 ttf 字库(汉字编码区间 4E00-6E1E)

(3)字库导出: 将彝文标准字库 OLD-YI.ttf 拷贝到地址为: C:\Windows\Fonts 即系统盘字体目录下, 后续工作都将支持彝文的显示. 目前还没有合适的软件实现从 ttf 字库文件到 txt 文档的批量导字, 课题组使用 UltraEdit 软件进行单字一一反向编码方式. (例如彝文字符 的编码为 4e00,导出时的输入为 004e; 彝文字符 的编码为 4e01, 导出时的输入为 104e, 依次…)打开新建的“彝文导出.txt”文档并另存为 Unicode 编码标准. 用 UltraEdit 打开新建文档并使用 HEX 模式进行反向编码[8]. 特别说明, 在任何有关“滇南彝文输入法”操作的前提下必须选择对应的 OLD_YI 古彝文字体, 否则输入无法显示彝文.

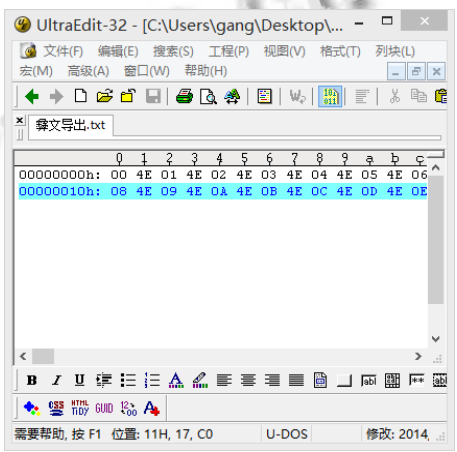

图 7 UltraEdit 反向编码

考虑人工将彝文字符(共 8223 个)导出是极其复杂而需耐心的工作, 编码稍有误差就会造成输入法字库漏字或重字. 课题组利用上述的反向编码规律设计简单的自动导字 C++程序, 该代码适用于任何少数民族

语言字符编码导字, 具体代码如下:

```
#include "stdafx.h"
int main(int argc, char* argv[])
{
 int i=0, cnt = 0;
FILE*  pfile=fopen("彝文导出.txt","wb+");//新建命名为“彝文导出”的 txt 文件
unsigned short buff =0xFEFF; //设置固定的文件格式
if (pfile == NULL)
{
   printf("File open error!");
 return 0;
}
printf("导字个数: ");
scanf("%d", &cnt);
fwrite(&buff, 2, 1, pfile);
buff = 0xe400; //这里 e400 代表你要导字区间的开始
while(i<cnt) //i 代表导字个数
{
    fwrite(&buff, 2, 1, pfile);
   buff++;
i++;
}
fclose(pfile);
return 0;
}
```

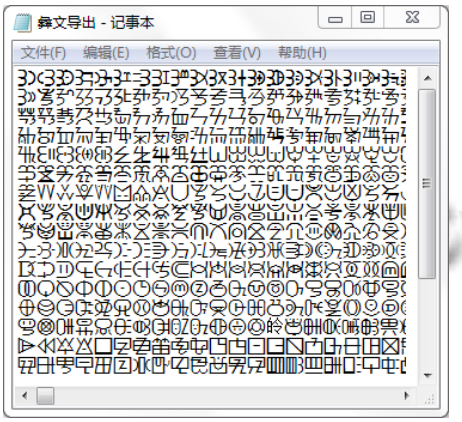

图 8　彝文字体导出

(4)字库编码规则: 在编码规则上采用课题组提出的“基于自由拆分模式的形态编码[9]”的编码设计方案. 考虑到在对同一个彝文字符进行拆分的过程中, 不同的人可能会出现不同的拆分结果, 我们在此方案基础上进行了优化改进, 提出“基于自由拆分模式一对多形态编码“规则. 从多次统计数据来看, 该编码方案快速、简单、易学, 无需死背字根, 输入方式具有极大的灵活性和极低的重码率, 且输入效率提高.

在彝文字符的拆分过程中, 根据彝文文字自身结构的特殊性, 对某些彝文字符的部首可能与多个英文字母存在相似性来进行形态编码. 而且对于要同一个彝文字符所可能出现的拆分结果都进行了编码, 以提高输入字的候选频率, 具体拆分原则如下:

1) 只要是一笔写成, 无论什么方向, 无论如何弯折, 都定义为最小拆分部首. 如果某个最小拆分部首存在棱角, 则按两笔算; 反之, 一律按一笔算.

2) 在采用形似拆分方式的基础上, 不指定最小拆分部首对应某个字根, 即对于最小拆分部首, 如果在平面旋转任意角度后形似于某个英文字母, 则其可用相应的英文字母表示;反之, 一律归为笔画.

3) 英文字母没有大小写之分, 对于形似大写字母的彝文文字部首, 一律用小写英文字母输入.

结合以上方案, 给出彝文文字形码编码的三个特性.

表 1　彝文文字形码编码的三个特性

	彝文文字	英文字母
相似性	S	s
	U	u
大小写同一性	E	e
	Ɛ	
平面旋转性	C	c
	Ɔ	
	ꓤ	dv
	ꓥ	ovc

编码流程举例示意图如下:

至此字库的编码工作全部完成, 码表 txt 文档生成完毕, 见下图 10 所示, 下一步转向输入法的实现. (例如彝文字  一对多形态编码基本可能为 5 种, 编码分别为 zic、zc1、ziv、zvi、ec、ev)

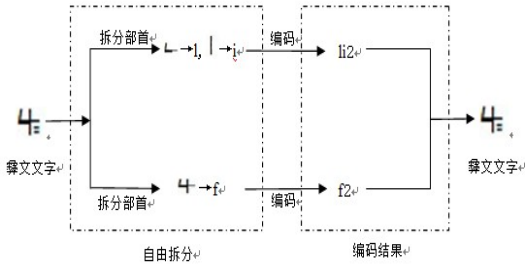

图 9　编码流程示意图

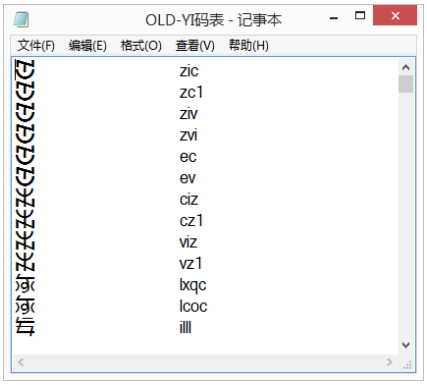

图 10 彝文码表示意图

2.2 读字规则

在上述字库编码及用户输入时字符拆分都要遵循一个最基本的准则“英文字母+剩余笔画数”，具体步骤如下:

1)输入彝文文字中各个最小拆分部首所对应的英文字母，其中字母输入顺序按照彝文文字结构进行输入，彝文文字结构分为: 上下结构(先上后下)、左右结构(先左后右)、上中下结构(先上再中后下)、左中右结构(先左再中后右)、包围结构(先外后内).

2)输入不能用英文字母表示的剩余最小拆分部首的总笔画数.

课题组制定本项最基本的准则，旨在保证开发者在字符编码和用户使用输入时能够按照规定读字方向，这样就避免了对于同一个彝文字的编码会因英文字母的顺序不同导致多码，而仅仅只会因形态差异产生多码，而这正是我们设计“一对多形态编码”的原由所在.

从理论上分析，“一对多形态编码”相对于优化后的“一对一形态编码”方案，会增加输入法重码字键选率，但却解决了目前初学者对于彝文的识别根本问题，而且从后期实际应用测试来看，改进后对重码字键选率影响并不是不大，在牺牲少许重码字键选率来提高输入法操作实用性和易学性是有必要的.

在后期输入法公开推广时[10]，课题组拟在软件下载页挂载使用说明和问题反馈，是为了更好的帮助广大用户和优化软件使用效果. 那么，初学者在使用过程中，只要首次配合说明书熟悉输入规则，具体流程，用户必须完成一系列转换: 汉字→笔顺或字的几何图形→根据拆分方法拆分为基本笔划、部件→数字编码→输入编码，在后期读字便可盲打，这就是该软件的使用易学性所在. 所谓易学是指用户经过较短时间的学习，对照输入法说明书会进行完成由汉字到输入编码的一系列转换，但是要达到一定输入速度，必须反复练习，达到汉字→数字编码的直接转换，这是需要一定的练习时间的.

2.3 输入法实现

本输入法是基于 Windows 平台而实现，利用多多输入法 API 进行程序开发而生成的彝文输入法，因为 DuoDuo-IME 结构在 Windows XP/7/8/10 所提供的输入环境下不仅系统稳定性高，且制作的彝文输入法相比于其他少数民族输入法，操作更加人性化. 为美化该软件的使用UI界面，我们制作了对应的皮肤包Skin_icons. 下图 11-12 所示为输入法实现的相关示意图

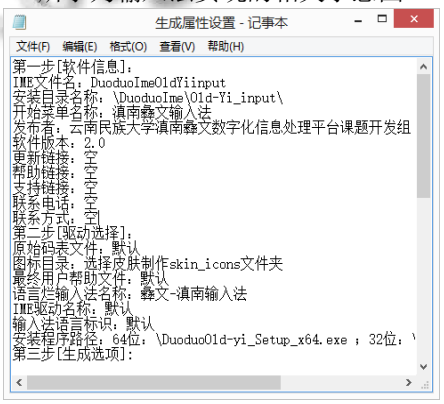

图 11 生成属性设置

只要选择 Windows 操作系统的应用环境，分别安装上对应版本输入法(32/64 位)程序包，以及我们开发的滇南彝文 TrueType 字库软件，就可实现对应的“滇南彝文输入法”. 滇南彝文输入法安装成功后，可以在 Word、Txt 等文字编辑软件中使用，还可切换输入法以实现中、英、彝文兼容显示. 下图 13-14 为输入状态效果示意图

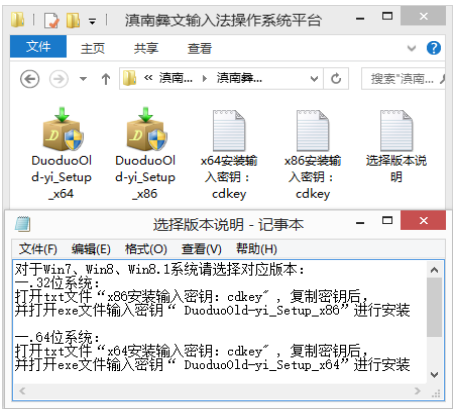

图 12 输入法生成安装程序

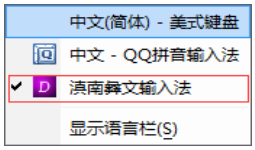

图 13　输入法安装完成

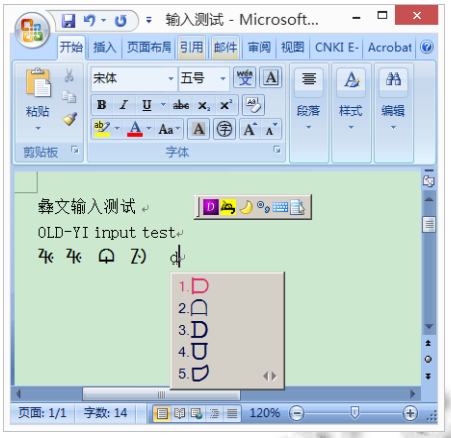

图 14　Word 中输入实现效果

3　输入性能分析

国家标准将编码层次和软件层次视为统一的键盘输入系统进行性能考核. 《GB18031 数字键盘汉字输入通用要求》提到的系统性能指标有三个方面, 在编码层次上要求达到定性指标, 如易学性; 在软件层次上要求达到量化指标, 如平均码长、重码字词键选率, 国家标准给出的指标是当前应达到的最低要求[11]. 因此本文对上述三个重要性能指标进行统计研究, 通过统计数据分析来测试输入法的可行性和高效性.

(1)易学性“学会使用汉字编码输入系统的时间应尽量短, 并应符合使用者的思维习惯”. 汉字编码标准(国标 GB18031)对数字编码更进一步提出要求: “做到上手能用”.

本文设计的滇南彝文数字信息化处理平台, 包含三套输入法软件操作平台均基于文字象形方式编码设计而成, 不仅适用于懂彝文的专业人员, 还是方便所有彝文的初学者的学习工具. 用户只需根据彝文字的结构特点和掌握输入规则即可完成彝文键入, 易学性要求已经达到. 课题组成员参与过内部测试, 普遍反映使用方便, 约在 10 分钟能学会如何使用.

(2)输入平均码长

即在输入给定的测试样本时, 测得的输入每个汉字的平均击键次数. 计算公式如下:

$$\text{平均码长}=\frac{\text{输入样本的击键盘次数}}{\text{测试样本总字数}}\ (\text{键/字})$$

GB/T18031 给出了通用键盘和数字键盘汉字输入平均码长的指标[11], 如下表 2 所示.

表 2　GB/T 18031(数字键盘)给出的指标

编码类型	平均码长(键/字)
逐字字段输入	<6
字、词混合输入	<4

本文的滇南彝文数字信息化处理平台包含“云南规范彝文数字健笔画拆分模式的形态编码输入法”和“云南规范彝文自由拆分模式的一对一形态编码输入法”及“滇南彝文自由拆分模式一对多形态编码输入法”三款输入法. (以下简称数字健拆分输入法、一对一形编输入法、一对多形编输入法)其中“数字键拆分输入法”采用的是定长编码, 编码字符(UsedCodes)=0123456789, 所有码字对应的编码长度相同, 码长均为 5; “一对一形编输入法”采用不定长编码, 编码字符(UsedCodes)=0123456789abcdefghijklmnopqrstuvwxyz 最小码长为 2, 最大码长 6, 字符-码字的映射采用一对一方式; “一对多形编输入法”采用不定长编码, 编码字符(UsedCodes)=0123456789abcdefghijklmnopqrstuvwxyz, 最小码长为 2, 最大码长 6, 字符-码字的映射采用一对多方式. 前两种编码的过程不在此加以熬述.

以下图示 15-17 分别是三种编码形式的码表文件.

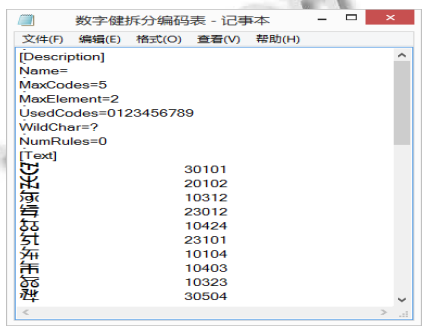

图 15　数字键拆分码表

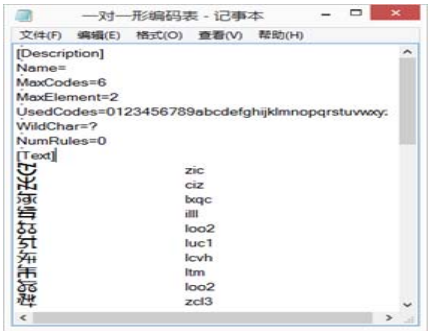

图 16　一对一行编码表

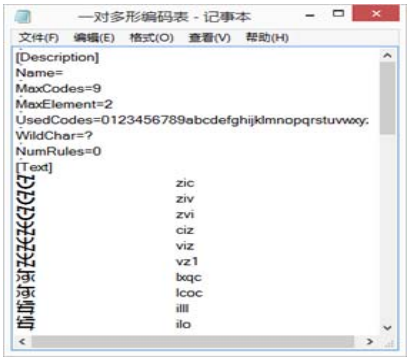

图 17 一对多形编码表

(3)重码字键选率

重码率是用来衡量输入法输入效率的定量指标, 重码率太高会导致输入时增加翻屏选字的次数, 因而我们在保证输入质量的前提下, 尽量降低重码率. 定义为在输入给定测试样本过程中, 通过重码选择键确认的文字字数与测试样本总字数的百分比. 计算公式如下:

$$\text{重码字键选率}=\frac{\text{重码选择确认的字数}}{\text{测试样本总字数}}\times 100\%$$

(4)性能测试及分析

在测试时字码本的字符集要保持统一, 以彝文标准字库中随机抽取 2000 个云南规范彝文字作为输入测试样本, 在以上前提下我们对三种彝文输入法进行了综合评估. 三款输入法测试结果见表 3.

表 3 三款输入法测试结果

测试用字符集	云南规范彝文		
测试样本总字数	2000		
测试输入法类型	数字健拆分输入法	“一对一”形编输入法	“一对多”形编输入法
编码方式	定长编码 5 最大码长 6	不定长编码 1-6 最大码长 6	不定长编码 1-6 最大码长 6
击键次数分布	击键数 l_i 码字个数 m_i 2 39 3 1347 4 358 5 256	击键数 l_i 码字个数 m_i 1 53 2 508 3 753 4 418 5 145 6 123	击键数 l_i 码字个数 m_i 1 65 2 617 3 841 4 375 5 82 6 20
平均码长(键/字) $\bar{L}=\frac{\sum_i l_i \bullet m_i}{\sum_i m_i}$	3.4155	3.2315	2.926
重码选择分布	重码个数码字个数 0 1678 1 158 2 98 3 56 4 10	重码个数码字个数 0 883 1 246 2 213 3 174 4 169 5 93 6 76 7 56 8 53 9 18 10～20 17 >20 2	重码个数码字个数 0 699 1 231 2 267 3 213 4 178 5 128 6 86 7 69 8 63 9 46 10～20 19 >20 1
重码分布特点	无重码 0.839% 重码∈[1.4] 0.161%	无重码 44.15% 重码∈[1.5] 44.75% 重码∈[6.9] 10.15 重码∈[10,+∞]0.95%	无重码 34.95% 重码∈[1.5] 50.85% 重码∈[6.9] 13.2% 重码∈[10,+∞] 1%
重码字键选率	16.1%	55.85%	65.05%

测试结果分析:

从平均码长来看, 第一款输入法的平均码长在 3-4(键/字)之间, 后两款输入法的平均码长集中在 2 和 3 上, 且输入法的平均码长 3(键/字)左右, 都比较符合

《GB18031 数字键盘汉字输入通用要求》中的平均码长标准, 这在很大程度上说明三款输入法具有较高的输入效率. 且第三款输入法采用的编码方式是“一对多”映射, 相比于第二款输入法的优势在于, 对于大多结构简单的彝文字用户不用查询码表来实现输入. 其码表中的每个字都基本涵盖了所有的编码可能, 如彝文字 的输入, 所有编码可能 zic、zc1、ziv、zvi、ec、ev, 用户只要根据输入规则联想其一便可完成指定彝文字的键入, 以此来降低带有歧义字形结构识别的错误率, 提高文字的可辨识性. 若从信息论的角度来分析定长编码和变长编码的选择问题, 前者输入法采用定长编码, 这在冗余度压缩编码中更具有优势, 另一方面, 使用定长编码容易在手动输入时查错, 避免造成错误输入. 第二、三款输入法运用变长编码方式, 虽然说提高了编码效率, 但却需要兼顾时间花费和存储空间的关系, 存储变长码必定需要更多的缓冲设备, 一旦造成误码, 容易造成连锁反应.

从重码字键选率来看, 三款输入法的码字个数均随着重码个数的增加而减少, 首款输入法的重码率最低, 从整体来看基本可以说是无重码. 理论上分析首款输入法采用数字键的拆分方式, 定长为 5, 产生的编码共有 10^5 种, 加上编码方式对编码重复的可能, 预计可以编码 10^5 左右个字. 而测试的总字符数只选取了 2000 字, 占总字数的 2%, 说明产生重码键入的可能性是比较低的, 跟最终实验数据得到的重码字键选率 16.1%基本是吻合的. 而在后两款输入法中, 是对彝文字的结构特点进行形似编码, 固重码率必然比第一款要高出许多. 其中第三款输入法是在第二款输入法的基础上进行改进, 在保证输入效率上采用一字多编码的映射方案会加大重码率, 固最终三者重码字键选率比较: “数字健拆分输入法”<“一对一”形编输入法<“一对多”形编输入法, 上述实验数据都比较符合理论分析结果. 三款输入法中重码字键选率在 16.1%-65.05%之间,从整体的重码字键选率来看虽然还是比较高的, 但这不能说明具体问题, 在输入候选框是允许出现 1-5 字的差别, 也即我们只用考虑重码属于 1~5 的重码字键选率, 前者在 16.1%, 而后两款都集中在 50%左右, 这都是比较符合实际应用范围的.

总体分析得出, 本文设计的第三款输入法软件操作平台相比于前者, 编码方案极大的解决了使用者易学性的问题, 在输入过程中方便快捷不用查询码表, 编码效率和输入效率都较高, 是比较理想的彝文文字处理软件.

4 总结与展望

本文论述了滇南彝文数字化信息处理平台的设计与研究方法. 主要论述了本平台彝文编码、彝文字库与彝文输入法三个模块的设计原理及其具体实现方法, 其中第三款滇南彝文输入法是基于 Unicode 的形似编码和多多 API 接口生成机制实现的,解决了长期以来中、彝、英不能混合输入、打印并显示的难题. 目前用于排版系统的彝文文字处理软件还尚未推出, 因此 Windows 操作平台下的滇南彝文输入法研究是一项应用创新性的工作.

本文对滇南输入法的编码原理与软件功能的研究方法和测试结果, 对形码编码发明者和输入系统设计人员有着指导作用; 通过对本平台下的各种彝文输入法性能的定量评测, 总结了各自的特点为用户选择合适的彝文输入法提供了依据; 对如何修改彝文输入法现行标准以及制定彝文输入法新标准有一定的参考价值; 对如何建立科学的彝文输入法评价体系有着积极的意义. 且本操作平台下的滇南彝文输入法可实际用于相关学术研究, 对彝文字的排版, 彝文古籍的印刷、彝文办公自动化和信息化的推广具有广泛的使用价值和示范价值.

参考文献

1 王正平.基于藏文国际编码字符集的输入法研究[学位论文].南京:南京师范大学,2008.

2 王明贵,吴勰,禄玉萍.论古彝文整理计算机输入软件的开发及其价值、意义与前景.中国科学技术协会、河北省人民政府.第十四届中国科协年会第 2 分会场:数字文化产业和技术创新国际研讨会论文集.中国科学技术协会、河北省人民政府,2012:4.

3 李永忠,刘勇,薛华,孔延香,郭秀峰.国际互联网上藏文信息交换平台的设计与实现.中国中文信息学会、中国科学院软件研究所、青海师范大学、五省区藏族教育协作领导小组办公室.第十届全国少数民族语言文字信息处理学术研讨会论文集.中国中文信息学会、中国科学院软件研究所、青海师范大学、五省区藏族教育协作领导小组办公室, 2005: 7.

4 宋建斌.基于 UNICODE 编码的蒙文编辑器[学位论文].呼

和浩特:内蒙古大学,2004.
5 贾海霞,沙马拉毅.《通用规范彝文方案》的研制与发展前景展望.西南民族大学学报(人文社会科学版),2014,9:19–24.
6 李昀姗,王嘉梅,郑晟.云南规范彝文字库设计及其字符集编码研究.电子科技,2011,5:97–101.
7 吴兵,张楠,刘玉萍,殷锋.X 窗口系统中彝文国标编码与显示.西南民族大学学报(自然科学版),2004,6:807–811.
8 李金发.试论计算机彝文字符编码的转换.云南民族大学学报(自然科学版),2008,1:80–84.
9 冯浩,王辉,王嘉梅.基于自由拆分模式的彝文输入法设计与实现.计算机应用,2010,S1:306–308.
10 冯浩,周莹,王嘉梅.彝文输入法自由编码方案的推广及应用.现代计算机,2011(9):3–5,11.
11 周克兰.汉字数码输入法评价体系研究[学位论文].苏州:苏州大学,2005.