西夏文计算机数字化现状与展望*

柳长青

摘　要：近年来西夏学及计算机科学快速发展，推动了计算机数字化的进程。从 20 世纪末开始，国内外学者在西夏文数字化研究方面不断有成果面世，这些成果对西夏文数字化研究产生了深远的影响。通过整理西夏文数字化研究的历程，对于今后进一步开展西夏文数字化研究工作有着极其重要的作用，对把握今后的研究方向也有一定的借鉴意义。

关键词：西夏文　数字化　计算机处理　数据库

一、引 言

随着西夏学研究的不断深入,以及西夏研究成果的不断问世，西夏学已越来越多地受到了人们的广泛关注。大批学者投入到与西夏有关的研究当中并开始关注电脑处理西夏文问题。计算机处理西夏文研究最早可以追述到 1972 年，丹麦人格林斯蒂德曾设计了一套西夏字的计算机编码方案，但最终未能实现。[①]到了 90 年代初，西夏学者为了出版与西夏有关的著作，亟需有一套计算机西夏文字库及排版软件，但限于当时的经费和技术条件的限制，最终未能实现开发软件的愿望。[②]这期间西夏文各类出版物几乎是靠照相制版，剪贴和校对的方法完成西夏文字的出版，耗时费力且效果不佳，由于制版方法繁琐还导致了的出版物中遗留了不少本不该有的错误。90 年代末，西夏学者李范文教授为西夏文录入计算机设计了四角号码和类似汉字的五笔字型输入法。[③]国内的大部分西夏文计算机编辑软件都采用了李范文教授的四角号码输入法。由于计算机技术的不断发展与进步，使得以个人或课题组形式开发字库及字处理系统成为可能。国内外的西夏学者与计算机学者合作展开了对西夏文字库及配套字处理系统的开发研制工作。这其中，国际上在西夏文的计算机处理方面，主要有日本、俄罗斯和中国台湾等地区的学者进行过研究工作。日本国立亚非语言文化研究所 1996 年制作了西夏文字库和排版系统，1997 年中国学者李范文教授和日本学者中岛干起利用该排版系统合著出版了《电脑处理西夏文〈杂字〉研究》一书。

1999 年 11 月由马希荣主编，柳长青为主要完成人的国家自然科学基金项目“基于汉字字形的西夏文字研究”的成果“夏汉字处理及电子字典”软件由清华大学出版社正式出版。该成果是按照四角号码和顺序号检字法对西夏字进行排列、注音和释义的 Windows 单机版应用软件。[④]它建立有 6000 号西夏字的两种西夏字库、西夏字与汉字、英文混合排版编辑，实现了西夏字的任意缩放输出，具有

* 本文所获基金资助项目：国家自然科学基金项目资助（批准号：60803104）

① 杜建录《二十世纪西夏学》，宁夏人民出版社 2004 年，第 121 页。

② 李范文《〈夏汉字典〉的编撰、四角号码分类和输入电脑问题》，《宁夏社会科学》1997 年第 4 期。

③ 同上。

④ 马希荣《夏汉字处理及电子字典》，清华大学出版社 1999 年。

字处理软件的所有功能。成为当时在国内外第一个能够独立完整的在个人计算机上进行西夏文、中文和英文互译，并同屏混排、输入、输出的软件产品。中国台湾中央研究院语言学研究所与资讯科学研究所于 1999 年开始研制西夏文字库，并于 2000 年顺利完成。其原始字体是依据《同音研究》校勘字形并加以数字化[①]。该所研究人员还利用 Access 数据库建立了“西夏文字形属性资料库”，并将电子文档经由计算机程序进行西夏字字频统计，试图找出西夏字的常用字、次常用字等，这在当时具有一定的先进性。这对于当前的西夏文献数字化研究也有一定的借鉴作用。2005 年国内西夏学者针对已有西夏文处理软件存在的不足，利用工具软件制作了基于方正典码系统之上的西夏文字库，借助“万能五笔”输入法实现了西夏文的计算机外挂方式输入。“基于方正典码的西夏文录入系统”利用汉字软件工具制作出了外形类似汉字楷体风格的西夏文字库。

上述西夏文处理系统的研制与开发大大改善了西夏文计算机处理的状况，并逐渐在西夏学与计算机学科之间产生了一门新的交叉研究方向——西夏文信息处理。国内的计算机学者们也积极开展了一系列的科研工作，并获得了从国家科研基金到地方基金的支持。至此，西夏文计算机处理的领域已不仅仅局限于字库的建立和排版系统的开发上，学者们更多地将目光投向了西夏文献数字化、网络化、西夏文网络输入法以及西夏文的国际编码和字形的标准化等问题上。本文则主要论述当前西夏文计算机数字化的现状及今后的发展趋势。

二、几种常用西夏文处理软件示例

当前，投入使用的西夏文字库软件有：1.日本的今昔文字镜西夏字库。2.“夏汉字处理及电子字典”软件。3.“西夏文字处理系统”。4.北京中易公司开发的西夏文字库及基于郑码的输入系统。5.中国台湾地区中央研究院民族语言研究所开发的西夏字库软件。6.宁夏大学西夏学研究院正在开发的西夏数字化平台及古籍西夏字库系统。上述这些字库均为 Windows 系统下的 True Type 字体库。

2.1 今昔文字镜

该系统是日本今昔文字镜研究会制作，包括 24 个 TTF 格式的矢量字库，共包括九万个汉字。其中收录日本《ISO10646 字符集》汉字两万个，《大汉和字典》汉字五万个，其他四万个汉字包括：甲骨文、梵文、大陆、台、港汉字、水文、西夏文、越南字喃、汉字偏旁和造字部件、日文假名、俄文、拉丁文等各种常用字母和符号。另外，还提供了一个简易的检索工具，如图 1 所示。

图 1 今昔文字镜字符表

该检索工具不能使用四角号码检索，只能手工查找所需的西夏字后利用软件提供的“拷贝、粘贴”功能将所需西夏文粘贴到 Word 等字处理软件中，对于少量的西夏文字的录入这种方法可行，但对于大量的录入工作则显得不大方便。字形结构方面，其字形笔画锋利，整体结构平直。该套西夏字库在国际上采用较多。

① 高雅琪《西夏文字输入法》，《第三届西夏学国际学术研讨会论文集》，2008 年，第 153 页。

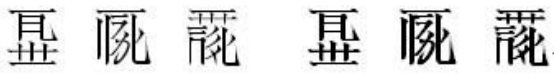

图 2 今昔文字镜字例

2.2“夏汉字处理及电子字典”软件

该软件包括两套西夏文字库，一部电子字典和一个西夏文字处理软件。软件开发语言采用可视化开发工具，字典数据库是开发者自行开发的数据结构文件。该软件提供汉夏互译及英夏互译功能，内建四角号码输入法，提供了一个西夏文字处理软件，并建立了包含 6000 西夏字的 2 套字库。其字库主要占用 Windows 的用户自定义私有码区域：AAA1H－AFFEH，F8A1H－FEFEH 及 A140H－A7A0H 三个区域，这些区域不与汉字码位冲突[①]。这三个区域一共可以放置 1894 个西夏字，为了将全部 6000 余西夏字放入，软件设计者巧妙采用了位面映射技术，如图 3。通过位面映射技术可以随意增加西夏字码位数，这解决了码位不够的瓶颈问题。

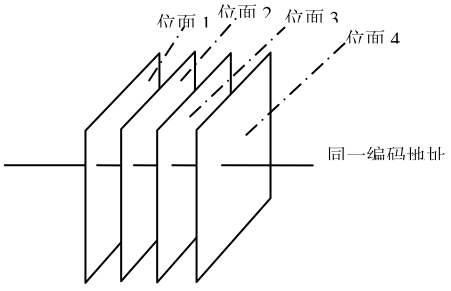

图 3 字体位面技术示意图

夏汉字处理输入法能够直接通过数字键盘输入 4 位四角号码加 2 位附号来检索西夏字。选字区还提供了字形放大镜以方便用户查看，如图 4。

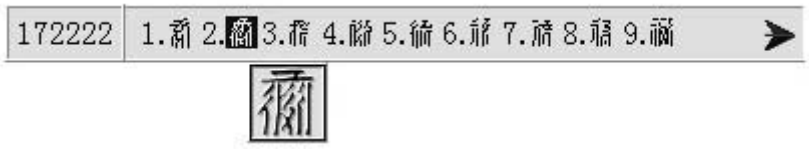

图 4 “夏汉字处理及电子字典”输入法

该软件还包括两套西夏文字库， 其原始字形分别来源于人工书写和夏汉字典 1997 年版字体。通过光学扫描技术将字形图片输入计算机后，再进行进一步的数字化处理。其中人工书写字体为毛笔手书，笔画较为粗重。

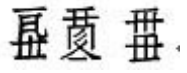

图 5 “夏汉字处理及电子字典”细体字库示例

“夏汉字处理及电子字典”软件从研制出版到现在已 10 余年，一度成为西夏学者计算机输入西夏文字的主要软件工具。尤其在 2000 年前后，其成为国内主流的西夏文计算机处理软件，被西夏学界广为使用。但也存在一些错码、漏码等问题，及软件作者未能及时修订和更新，现逐步被其他西夏

① 柳长青、马希荣《西夏字与汉字共存方案的实现》,《宁夏大学学报》(自然科学版)，2001 年第 1 期，第 45~47 页。

文处理软件取代。

2.3“西夏文字输入法”软件

“西夏文字输入法”软件是台湾中央研究院语言学研究所与资讯科学研究所合作研发的西夏文输入法软件。该软件包括一套西夏文字库、一个西夏文字形属性数据库及西夏字字频统计资料库。其西夏文字库原始字形主要来源于“同音研究”校勘本。[①]其字形是目前已正式出版的西夏文字库中最为接近西夏文原始字形的 1 套字库。“西夏文字形属性资料库”包括西夏文序号、字形、同音编号、笔画、部、品、型、修正音韵、页字、俄文编号、拟音、声类、调类、词义、例句、说明、夏汉字典编号、四角号码、附号、龚煌城先生的文海编号、《文海研究》编号、《义同一类》编号、西田龙雄先生的编号、Nevesky 的编号。[②]目前，该资料库仅对外开放《同音研究》索引的相关资料包括部、笔画、品、音韵、页字及俄文编号、夏汉字典编号等常用属性，其余资料还在不断修改中暂不对外开放。

其“西夏字字频统计”资料库对 9 种西夏文献中的西夏文进行了字频统计。这 9 种文献分别是：《黄石公三略》(11,484 字)、《根本说一切有部目得迦》(4,906 字)、《根本说一切有部毘奈耶杂事》(5,791 字)、《禅源诸诠集都序之解》(18,980 字)、《禅源诸诠集都序幹文》(3,651 字)、《维摩詰所说经》(26,042 字)、《大方广佛华严经》(15,157 字)、《月月乐诗》(1,398 字)、《将苑》(1,237 字)。[③]研究人员将这些文献作为样本，对 5777 个西夏字进行了字频统计。得到了出现频率最高的西夏字为“  ”，总共出现 1,376 字（占样本总数的 1.55%）。[④]该字频统计资料库对于建立智能西夏文输入法有一定的借鉴作用。

2.4“西夏文字处理系统”

该系统的前身是 2005 年出版的“基于方正典码之上的西夏文录入系统”，景永时、贾常业编著。该典码录入系统是借助方正典码输入法的开放接口建立的，能够运行于 Windows95/98/ME 操作系统之上。其主要包含一套西夏字库和一个典码输入法。值得一提的是，该系统所建立的字体库一经推出得到了西夏学者的推广和使用，尤其在书版系统中有较好的应用。由于其西夏字形是通过使用汉字笔画组合造字而来，因此与汉字混排后风格统一、格式整齐、笔画粗细均匀，是一套较好的印刷体西夏字体库，见图 6 所示。基于典码的输入法则是借用了方正典码汉字输入法软件外壳加入了软件编者的西夏字码表文件而建立的西夏文专用输入法。由于典码本身的输入码限制只能使用字母输入，因此该系统编者采用了转义字符方式将四角号码的 0-9 的数字码转换为其汉语生母即 1-y,2-e,3-s,4-x,5-w,6-l,7-q,8-b,9-j,0-o（字母 o），其中由于 3 和 4 的生母相同，故将 4 用 x 代替。例如，要录入四角号码为 174422 的西夏字则需输入 yqxxee。[⑤]这种转化实属无奈之举，在没有软件源代码支持情况下通过这种变通也算实现了西夏文的输入，只是对于初学者需要一个熟悉的过程，一旦使用熟练后也可快速输入西夏文。2007 年基于典码的西夏文处理系统的编者推出该系统的更新升级版本即“西夏文字处理系统”。该系统中的西夏文字库是对于典码系统中字库的修订及升级，而输入法则改换为万能五笔输入法为平台而制作的西夏文外挂式输入法。该输入法仍然沿用了典码系统中的数字-字母的转换方式。2007 版的软件配有光盘 1 张及使用手册一本。《西夏文字处理系统》的出版进一步推动了国内西夏文数字化的研究进程。

图 6 西夏文字处理系统字体库示例

① 高雅琪《西夏文字输入法》，《第三届西夏学国际学术研讨会论文集》，2008 年，第 153 页。
② 同上，第 155 页。
③ 同上，第 156 页。
④ 同上，第 157 页。
⑤ 景永时、贾常业《基于方正典码之上的西夏文录入系统》，香港社会科学出版 2005 年，第 8 页。

2.5“西夏文古籍字库”软件

“西夏文古籍字库”软件是宁夏大学西夏学研究院研究人员开发的西夏文数字化处理系统的一部分。该系统包括 1 套西夏文字库，一个西夏文献数字化平台及西夏文智能输入法和在线夏汉电子字典软件。西夏文古籍字库字形主要来自于“同音”及“蕃汉合时掌中珠”西夏文献。其字形是将原始西夏文献中的西夏字切割，再扫描输入计算机，最后利用计算机图形学相关技术提取切割图像的轮廓信息，将提取后的字形存储并加以修饰。最后得到了 1 套基于原始西夏文字的西夏文古籍字库。该套字库建立的目的是希望能够尽可能的保存原始西夏文字的笔画及笔锋、力道等信息，尽可能体现原始西夏字的风貌。

图 7 西夏文古籍字库字形示例

西夏文献数字化平台则基于西夏古籍文字库基础上的文献数字化显示及检索平台。在该平台下，用户可以通过查看原始文献扫描图像得到第一手西夏文献资料，并能通过数字化得到该文献图像的纯文字版本的电子文档，且还能在该文档中进行检索及查询显示，如图 8-9 所示。通过网络在线平台，用户还可以将所检索的关键字内容进行全库查询检索，即对已入库的所有西夏文献进行关键字查询操作。最终可以得到该关键字有关的上下文内容条目。

图 8 西夏文献数字化平台

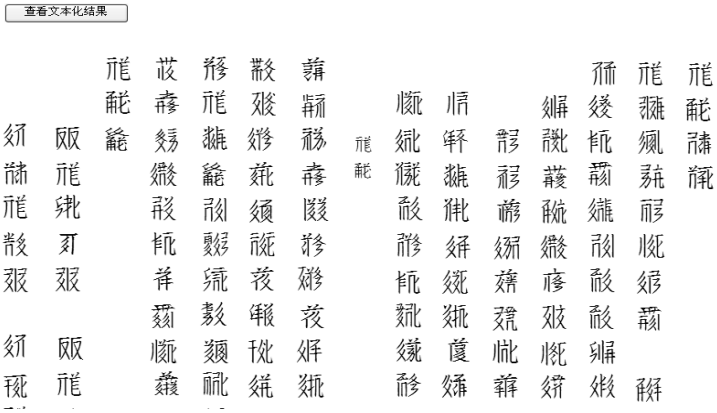

图 9 自动纯文本化处理后的电子文档页面

基于四角号码的西夏文智能输入法是 Windows 系统下的纯 IME 输入法，能同时支持中、英、日、俄四种语言的 Windows 操作系统，如图 10 所示。带自学习的智能拼音联想汉字输入功能。对于西夏文字的输入可以根据用户输入的频率自动将高频率的西夏字优先排列，从而达到提高西夏文录入速度的目的。该输入法还提供西夏字输入的汉文与英文的释义显示窗口，在选字窗口右侧同时显示西夏文所对应的汉文释义，如图 11 所示。

图 10 夏汉通西夏文智能输入法状态条

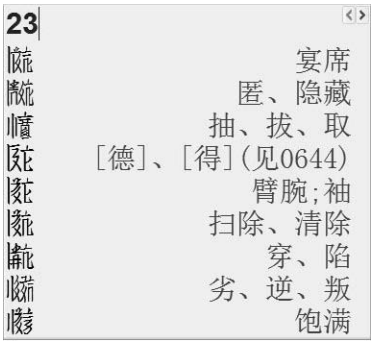

图 11 夏汉通输入法选字窗口

三、西夏文计算机数字化展望

进入 21 世纪，西夏文计算机数字化研究已经取得了长足的发展。从 1972 年最早期的格林斯蒂德设计的西夏字计算机编码方案到现在国内外计算机数字化软件的不断推陈出新，使得西夏文计算机数字化研究逐渐成为西夏学研究的一部分。回顾近十几年的西夏文数字化历程，其从无到有的发展，每一步都倾注了学者们的汗水与智慧，每一次的突破也推动着西夏文数字化向着实际应用的方向迈进。未来西夏文计算机数字化研究应从以下几个方面继续开展研究工作：1. 西夏文字库的标准化及其国际编码方案的建立。2. 西夏文及西夏文献数据库的建立。这部分工作内容最多，任务最重，所需时间也最长，可以说是西夏文计算机数字化的最终研究方向。所有的前期数字化工作都是为最终形成全面的西夏文数据库做准备。该数据库将涵盖西夏学所有的研究方向，无论从语言、文字、音韵、西夏文献及西夏艺术等方方面面都将以数字化的形式建立相关的数据资料库。3. 西夏数字化应用开发。如何将西夏学的成果应用于社会并产生效益也是今后的主要研究方向。总之，西夏文计算机数字化还是一个亟待挖掘和研究的新兴领域，在西夏学者与计算机学者的共同努力下一定会有新的发展潜力。

（作者通讯地址：宁夏大学数学计算机学院　银川　750021；宁夏大学西夏学研究院　银川　750021）