



## 面向语音拟构的西夏古文献数据库 结构设计及其实现

叶建雄 单迪

**摘要:** 本文主张应用计算语言学的研究方法,为西夏音韵学专题库设计一个优化的系统结构,把内部数据、共时数据和历时数据有机地整合起来,以充分发挥相关古文献在西夏语音拟构中的作用。

**关键词:** 西夏文 数据库 语音拟构

### 一 缘起

西夏文字的语音构拟经过聂历山、<sup>①</sup>王静如、<sup>②</sup>西田龙雄、<sup>③</sup>龚煌城<sup>④</sup>等前辈学人的不懈努力,取得了瞩目的成果。随着计算机技术的发展,尤其是计算语言学的兴起,这些成果面临前所未有的挑战。前辈学人胼手胝足、披肝沥胆,在浩如烟海的古文献中,寻找西夏语音的蛛丝马迹,为西夏文字读音的构拟奠定了基础。然而,当代语言学正经历着从转换—生成向语料库蓦然回首的进程,这种发展动向为西夏学界提出了新的课题:如何应用计算语言学的研究方法,把西夏文字读音的构拟向西夏语语音模拟与重构推进一大步。

在西夏学研究领域较早引进计算机技术的是日本亚非语言文化研究所,1996年该所研发的西夏文字库和排版系统发布。而后,李范文教授与中岛干起合作,<sup>⑤</sup>史金波教授与日本东京外国语大学合作<sup>⑥</sup>,推出了一系列研究成果。从日本文字镜研究会的西夏字库<sup>⑦</sup>到宁夏大学的共存方案,<sup>⑧</sup>对于西夏文的计算机处理、电子出版、信息系统等等应用,作出了巨大的贡献。然而,由于这些方案自身的结构局限,很难适应于计算语言学研究方法所需的专题数据库。

为了在西夏音韵学领域运用计算语言学的研究方法,需要研制适用于音韵学专题库的西夏文信息处理方案,加州大学伯格里分院的多位专家在 Unicode Consortium 的麾下进行了数年艰苦卓绝的努力,终于在2008年发布了 UniTangut(L2/08—259/336),<sup>⑨</sup>为建立西夏音韵学专题库提供了权威的标准。本课题的研究旨在应用该项标准,为西夏音韵学专题库设计一个优化的系统结构,把内部数据、共时数

① 《西夏研究》第6辑,中国社会科学出版社2007年,第1—9页

② 王静如《西夏研究》第3辑,国立北平图书馆馆刊1933年。

③ 西田龙雄《西夏语韵图〈五音切韵〉的研究》,京都大学1981年。

④ 龚煌城《西夏语言文字研究论集》,民族出版社2005年,第243—268页。

⑤ 李范文、中岛干起《电脑处理西夏文〈三才杂字〉研究》,日本亚非语言文化研究所1997年。

⑥ 史金波《电脑处理西夏文〈文海宝韵〉研究》,日本外国语大学2000年。

⑦ 文字镜研究会: <http://www.mojikyo.org/>, 2002年。

⑧ 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》2001年,第45—47页

⑨ Richard Cook: <http://linguistics.berkeley.edu/~rscook/UTC/Tangut/>, 2008.

据和历时数据有机地整合起来。

其中,内部数据主要指西夏文文献,包括宗教文献和世俗文献两部分,其中音韵文献采用数据库的格式存储,其余的则采用语料库的格式存储;共时数据则包括夏汉、汉夏、夏藏、夏梵等双语对照文献;而历时数据指的是羌语支诸语言以及跨语系相关语言历时演变发展的文献。处理如此异构而庞杂的古文献数据,必须设计一个优化的数据库结构,才能发挥这些古文献在西夏语音拟构中的作用。西夏音韵学专题库以数据库为主,以语料库为辅,为生成基于计算语言学的西夏语音拟构算法提供一个有效的平台。本文着重介绍数据库的结构设计,期望起到一点抛砖引玉的作用,请西夏学界的专家学者斧正。

## 二 编码空间与古文献数据组成

笔者迄今为止所见的西夏文字库及其处理软件的机内码与人—机交换码的设计,都不能满足面向语音模拟与重构的西夏古文献数据库的基本要求。先来看看机内码的问题,最近的例子就是 2008 年出版的《西夏研究》第 7 辑,西田龙雄的前言<sup>①</sup>中所引用的西夏字,全都变成了繁体汉字,其中有几行抄录如下:

噴杉璽筱迅噴 (給我獎賞—給我)

噴棘輓盍杉璽樺噴 (我給庶民獎賞—我)

这就是该书排版的西夏文字库的机内码与汉字机内码重叠所造成的结果,尽管我们看到西夏文与中文排在一个版面,其实两者却用了两个不同的字体库,一旦忘了转换,就会出现错位。上述例子说明目前流行的那种机内码重叠的字库,不仅不能适用于计算语言学研究的专题数据库,用于排版也有时会出错。

### (一) 机内码的合理区间

面向语音拟构的西夏文字机内码设计,要求汉字、梵字、藏字、西夏字在同一个平面内连续编码,互不重叠,这样才能符合语音拟构的专题库的基本需求。加州大学伯格里分院在 2008 年公布的多字节 Unicode 方案,为达到上述目标提供了权威的参照标准。

现行的西夏文字库之所以占用汉字的机内码空间,<sup>②</sup>或者西夏字本身的机内码相互重叠(宁夏大学的国家自然科学基金资助项目),其主要原因是排版应用软件大多支持双字节内码,对于多字节内码的支持,是在 WindowsVISTA 之后才推广开来的。美国加州大学伯格里的方案为解决了这个问题提供了权威的标准,该方案的西夏字机内码的区间如下:

U—17000—1871FH<sup>③</sup>

而中日韩超大字符集汉字统一码 CJK 机内码的编码空间如下:

U—03400—09FAF

由此可见,这两者完全独立,互不重叠。加大伯格里方案还有一个重要的贡献,那就是采用了西夏学界最新的研究成果——西夏文正字法,并根据笔画数按照 400 多个西夏文部首排序,这样,整个字库的布局显得非常整齐合理,也为西夏字形的标准化提供了权威的样板。

### (二) 人—机交换码的设计原则

在互联网时代,小键盘字符成为人—机交换码的主流。看看网上虚拟社区的交流平台,德语区、法语区、土耳其语区,还有斯堪的纳维亚语区等等那些文字中包含非小键盘字符的虚拟社区,现在都清一色用小键盘字符进行网上交流,这个趋势正符合信息交流中简洁快速的原则。

在古文献研究领域,诸如梵文、突厥文、犍陀罗文等,尤其是在网上传播文本,最不容易产生歧

① 西田龙雄《西夏研究》第 7 辑,中国社会科学出版社 2008 年。

② 中易: <http://www.china-e.com.cn/main/product/xixia.htm>, 2005.

③ Richard Cook: <http://www.unicode.org/unibook/UniTangut20080821.dbf>.

义的就是标准 ASCII 文本，那些采用自定义字库字符编码的文本往往难以统一，从而导致错误百出。顺应这一浩浩荡荡的世界潮流，采用“仅取小键盘字符”作为设计人一机交换码的原则是行之有效的方案，其次再考虑编码的冗余度以及“见字识码”和“见码识字”的原则：

#### 1. 仅取标准小键盘字符

标准小键盘字符是目前互联网信息交换领域里的国际统一标准，把人—机交换码的边界，设定在这个区间内的好处是显而易见的；

#### 2. 全息编码与适度冗余

鉴于西夏文是表意文字，其一字多音与一音多字的特性，否定了用字音编码的可行性，唯有采用字形编码，而且用“全息”的原则给编码空间以适当的冗余度，做到一字一码；

#### 3. 见字识码与见码识字

指编码与西夏字之间的双向映射的唯一性，而且要求一目了然，即编码与字形严格对应并建立充分的联想机制，按照加大伯格里方案确定的部首分类逐一编码。

### （三）古文献数据的基本组成

值得注意的是，加大伯格里发布的数据库是目前西夏文研究数据库中较完备的一个系统。其中凝聚了近一个世纪以来，各国专家的研究成果。该系统直接在页面中呈现 256 个西夏字的方阵，检索效率较高。西夏音韵学专题库采用加大伯格里的标准码本，与小键盘字符串组成的人—机交换码全息映射，构成了专题库系统的人—机模块。

西夏音韵学专题库的古文献数据主要由如下三大块组成：

#### 1. 内部数据

指众所周知的西夏古文献，包括音韵文献《文海宝韵》、《同音》、《同音文海宝韵合编》<sup>①</sup>和《五音切韵》<sup>②</sup>等，采用数据库格式存储，而更大量的西夏世俗文献和宗教文献则采用语料库的格式存储；

#### 2. 共时数据

指当时在西夏境内流行的外族语言，包括汉语、藏语、回鹘语、契丹语、女真语以及佛教梵语等，这些语言对西夏语有着广泛而深刻的影响。其中夏汉双语的《番汉合时掌中珠》尤其关键。另外，夏藏并行文献和夏梵对堪文献也是必不可少的古文献语料；

#### 3. 历时数据

指藏缅语族羌语支的各种活语言，包括羌语、普米语、道孚语等，特别是木雅语，被认为是西夏语的直接后裔<sup>③</sup>；还有阿尔泰语系的文献语言，包括突厥语、蒙古语、满语等，其中《蒙古字韵》<sup>④</sup>等与宋元汉语官话音系相关的音韵文献尤其重要。

## 三、音韵学专题库的结构设计

音韵学专题库的应用目标是西夏语的语音模拟与重构，采用计算机语音仿真技术（phonetic emulation），通过专题库系统的支撑达成语音重构（reconstruction）的目标，也就是说最终得到让计算机开口朗读西夏语文本的结果。无庸讳言，目前各家推出的西夏文语音构拟方案，还有很多悬而未决的问题，诸如松紧音的问题、卷舌音的问题、长短音的问题以及小舌音系的问题<sup>⑤</sup>等等。当我们拓展视野在语言学广阔背景下，考虑在西夏这个特定的时空区域内，汉藏语和阿尔泰语跨语系的交互影响及相互制约，这就确定了西夏语语音格局的边界，为达成目标找到了可能的途径。

① 韩小忙《〈同音文海宝韵〉合编整理与研究》，中国社会科学出版社 2008 年。

② 李范文《〈五音切韵〉与〈文海〉比较研究》，《西夏研究》第 2 辑，中国社科出版社 2006 年。

③ 黄布凡《藏语藏缅语研究论集》，中国藏学出版社 2007 年，第 215—312，562—574 页。

④ 罗常培、蔡美彪《八思巴字与元代汉语》，中国社会科学出版社 2004 年，第 116—164 页

⑤ 孙宏开、刘光坤《也谈西夏语里的小舌音问题》，《宁夏大学学报》2001 年，第 58—63 页

### （一）专题库的总体设计

西夏古文献数据库作为音韵学专题库系统的架构，以音韵学古文献为主轴，以其它文献以及周边的共时与历时文献语料为分支。具体地说，就是用音韵学古文献建立 DBF 文档，用其它古文献建立 CF 文档，广泛采集并筛选古文献语料，聚焦于西夏语音格局重建这个目标，设计词类标记、句法标注以及深加工程序，协同操作构建西夏音韵学专题库系统。西夏音韵学专题库的总体结构图示意如下：

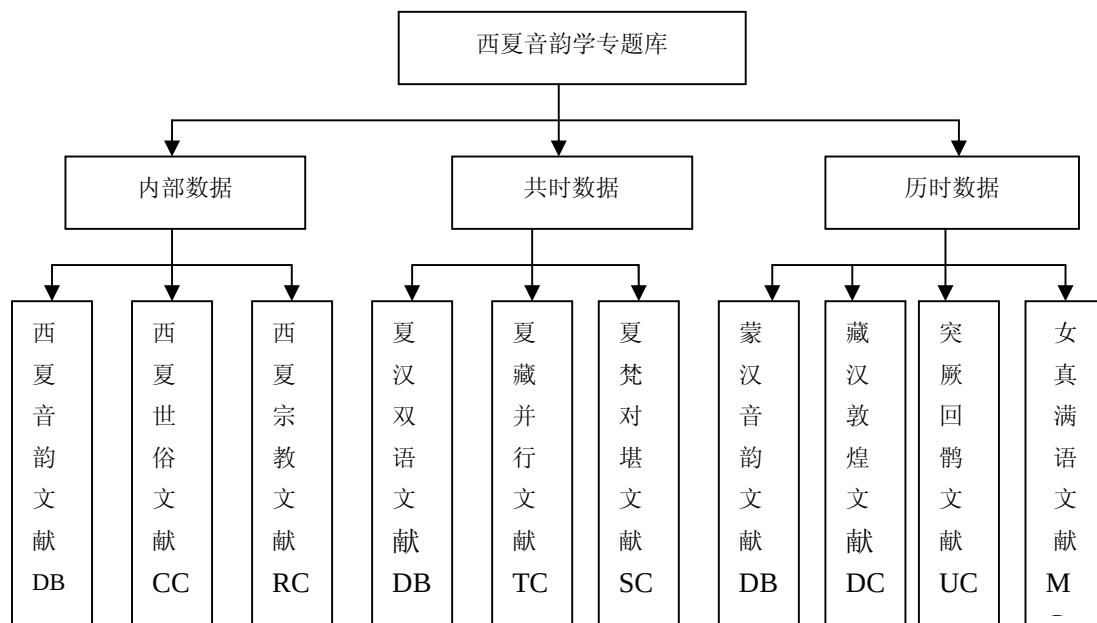


图 1. 西夏音韵学专题库总体结构图

其中，DB—DataBase；CC—Civil Corpus；RC—Religion；TC—Tibetan Corpus；SC—Sanskrit Corpus；

DC—Dunhuang Tibet—Chinese Corpus；UC—Uighur Corpus；MC—Manchu Corpus。

藏汉敦煌文献（DC）中，最重要的是用藏文字母拼写汉语的那些古写卷，黄布凡教授有很多的研究成果，还有托马斯等前辈学人的成果可以参考。

### （二）数据库的逻辑设计

数据库的概念设计和逻辑设计就是给数据库建立用户视图和管理员视图。这些视图一般用 E—R 图（即实体—关系图）、数据流程图等来表示。西夏音韵学专题库的 E—R 图和数据流程图，总共有 36 页，这里选择几组典型的 E—R 图示意如下：

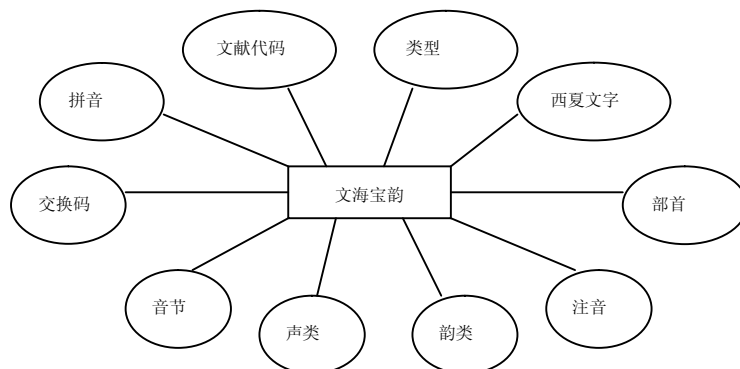


图 2. 文海宝韵数据库表 E—R 图

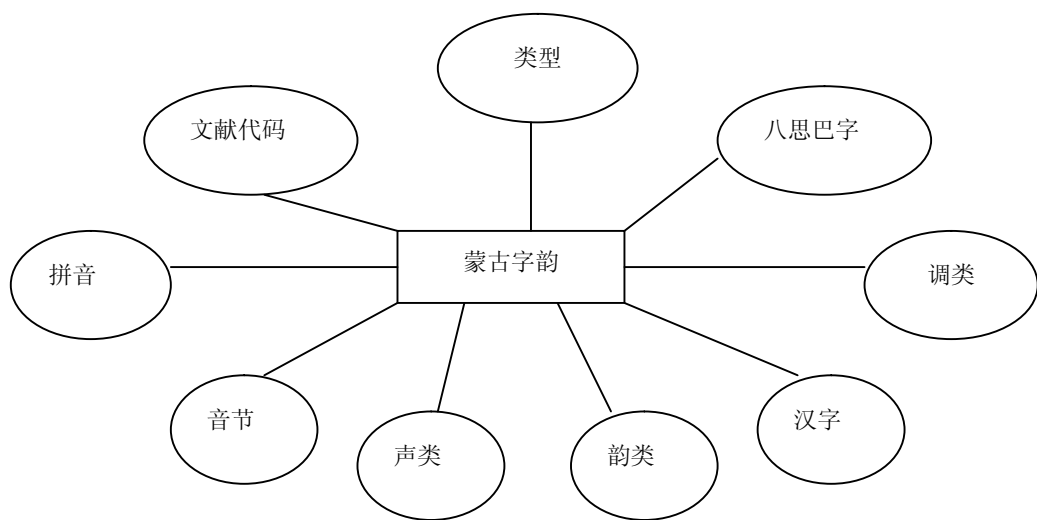


图 3. 蒙古字韵数据库 E—R 图

专家学者们感兴趣的不仅仅是实体与属性的关系，更重要的是实体与实体之间的互动关系，以及系统数据流程的概貌，这些才是西夏音韵学专题库的核心所在。限于论文篇幅，只能选一组单元 E—R 图示意如下：

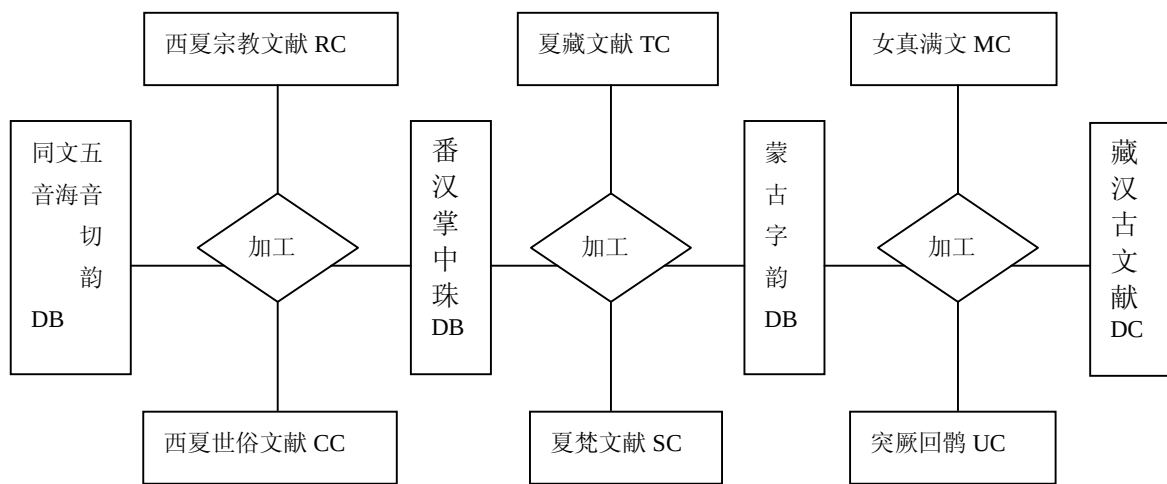


图 4. 西夏音韵学专题库 E—R 图

### （三）跨语系古文献语料的组成

西夏音韵学专题库中的西夏语料、夏藏语料、夏梵语料以及跨语系的阿尔泰诸语言的古文献语料，为生成基于计算语言学的语音模拟与重构算法提供平台支撑。其中跨语系的阿尔泰诸语言，例如突厥古文献、回鹘古文献、女真满文古文献的语料选取，基于这些语言在当时西夏境内作为较强势的语言，对西夏语的演变起到了至关重要的作用。

首先是回鹘文。由于西夏人笃信佛教，而回鹘僧人常常作为西夏的国师甚至帝师，在西夏各地漫游。许多流行的佛经都由回鹘僧人翻译成西夏文，回鹘僧人受到西夏各阶层的崇敬<sup>①</sup>。因此，回鹘语

① 耿世民《敦煌突厥回鹘文书导论》，新文丰出版公司 1994 年。

音对西夏语的影响不可低估。

还有稍早一点的突厥如尼碑铭及文献，则是解决回鹘文中诸多难题的钥匙。所以，建立回鹘文语料文档，必须同时建立突厥文语料文档：

☆ 回鹘佛教文献语料文档；☆ 回鹘世俗写本语料文档；☆ 突厥如尼文本语料文档。

其次，女真的强大，构成当时金、西夏、南宋三足鼎立的局面。女真语对西夏语的制约也是不容忽视的。然而，由于女真文的资料数量太少，要还原西夏时代的女真语，还必须借助女真语的直接后裔满语<sup>①</sup>，所以，加入了女真文语料库，必须同时加入满文语料库：

☆ 女真文书文本语料文档；☆ 满文历史典籍语料文档；☆ 满文佛教经律语料文档。

四、音韵学专题库的实现

西夏音韵学专题库，广泛采集了《文海宝韵》、《同音》、《同音文海宝韵合编》、《五音切韵》、《番汉合时掌中珠》以及《西夏文世俗文献》、《西夏文宗教文献》等经典古文献，还整合了阿尔泰语系诸语言的古文献语料，诸如《蒙古字韵》、《突厥回鹘文献》、《女真满文文献》等等，可以在此平台上生成基于计算语言学的语音模拟与重构算法。

（一）西夏音韵文献的数据字典

日本东京外国语大学发布的《西夏文〈文海宝韵〉电脑处理》<sup>②</sup>，为西夏文音韵文献的数据库结构设计提供了成功的样板。本课题处理的西夏文音韵文献主要是《同音》、《文海宝韵》、《同音文海宝韵合编》、《五音切韵》等，其典型的数据库字段结构表述如下：

| 序号 | 字段名称 | 数据类型     | 长度解析                        | 注 释          |
|----|------|----------|-----------------------------|--------------|
| 1  | 文献代码 | CHAR(10) | 卷号页码 4 页面行号 2 大字序号 2 小字序号 2 |              |
| 2  | 类型   | CHAR(4)  | 大字类型 2 小字类型 2               | 代/序/普/加/仅*   |
| 3  | 西夏文字 | CHAR(8)  |                             |              |
| 4  | 部首   | CHAR(6)  |                             | 452 类        |
| 5  | 汉字注音 | CHAR(16) |                             |              |
| 6  | 韵类   | CHAR(8)  | 摄 2—等 2—调 2                 | 105 韵**      |
| 7  | 声类   | CHAR(4)  | 纽 2 类 2                     | 唇/舌/牙/齿/喉*** |
| 8  | 音节   | CHAR(16) |                             |              |
| 9  | 交换码  | CHAR(12) |                             |              |
| 10 | 汉译   | CHAR(16) |                             |              |
| 11 | 拼音   | CHAR(32) |                             |              |
| 12 | 英译   | CHAR(64) |                             |              |

表 1. 西夏音韵文献的字段结构表

\* 大字类型：01 代表字/ 02 序文字/ 03 普通字/ 04 追加字/ 05 专有字/ 08 推定字；  
小字类型：11 构造字/ 12 意义字/ 13 解释字/ 14 左联字/ 15 右联字/ 18 推定字；  
\*\* 01 摄—01—04 等—01 平 02 上 03 去 04 入/  
02 摄/ 03 摄/ 04 摄/ 05 摄/ 06 摄/ 07 摄/ 08 摄/ 09 摄/ 10 摄/ 11 摄/ 02 摄；  
\*\*\* 01 重唇音/ 02 轻唇音/ 03 舌头音/ 04 舌上音/ 05 牙音/ 06 齿头音/ 07 正齿音/

① 佟永功《满语与满文档案研究》，辽宁民族出版社 2009 年。  
② 西夏学会：<http://www3.ua.tufs.ac.jp/~mnaka/tangutindex.htm>，2006 年。

08 喉音/ 09 来日音/ 10 舌齿音。

(二) 夏汉双语文献的库表结构

夏汉双语文本，主要是指《番汉合时掌中珠》这样的古文献。《番汉合时掌中珠》是 1190 年刊刻的双语双解词汇表，其词条结构比较特殊，以适应其音义双向译释的需要，其数据库表的结构与西夏音韵文献有很大的区别。《同音》、《文海》和《五音切韵》都是以字为单位，无须考虑词汇的切分，而《番汉合时掌中珠》则是以词为单位的，所以，从序文开始，就要考虑词的切分。这类序文，用语料库的格式存储就显得比较合理，而正文则用数据库格式更加有效。

《番汉合时掌中珠》的数据库表结构，考虑以词为单位的字段，还要涉及语料库的词类标记和句法标注等等互操作的因素。由此可见，汉夏对音文献的库表结构，是西夏音韵学专题库的枢纽，其典型的库表结构列表如下：

| 序号 | 字段名称 | 数据类型     | 长度解析                        | 注 释          |
|----|------|----------|-----------------------------|--------------|
| 1  | 文献代码 | CHAR(10) | 卷号页码 4 页面行号 2 大字序号 2 小字序号 2 |              |
| 2  | 类型   | CHAR(6)  | 大字类型 2 小字类型 2 章节类型 2        | 天/地/人/上/./下* |
| 3  | 西夏文字 | CHAR(10) |                             |              |
| 4  | 汉字注音 | CHAR(10) |                             |              |
| 5  | 汉字译意 | CHAR(12) |                             |              |
| 6  | 西夏注音 | CHAR(12) |                             |              |
| 7  | 汉注拼音 | CHAR(16) |                             |              |
| 8  | 汉译拼音 | CHAR(16) |                             |              |
| 9  | 词类标记 | CHAR(8)  | 分隔符 2 句法标记 2 专用标记 4         | 句法词类**       |
| 10 | 辞源标注 | CHAR(16) |                             |              |
| 11 | 英译   | CHAR(32) |                             |              |

表 2. 夏汉双语文献的字段结构表

\* 章节类型：31 天体上/ 32 天相中/ 33 天变下/

41 地形上/ 42 地相中/ 43 地用下/ 51 人体上/ 52 人相中/ 53 人事下；

\*\* 句法词类：n 名词/ t 时间词/ s 处所词/ f 方位词/ m 数词/ q 量词/ b 区别词/ r 代词/  
v 动词/ a 形容词/ z 状态词/ d 副词/ p 介词/ c 连词/ u 助词/ y 语气词/  
e 叹词/ o 拟声词/ i 成语/ l 习用语/ j 简称/ h 前接成分/ k 后接成分/ g 语素/  
x 非语素字/ w 标点符号；

专用标记，是指音韵学相关古文献专用的标志，在西夏古文献语料、夏藏古文献语料、夏梵古文献语料以及阿尔泰诸语古文献语料的标注与加工过程中，起到切入点的作用。

(三) 《蒙古字韵》在专题库中的位置

在《番汉合时掌中珠》中注音的汉语是 12 世纪西北方言的音系，该音系的语音面貌众说纷纭，莫衷一是。而《蒙古字韵》则是元代制定的汉语官话音系，<sup>①</sup>其发布时间是 1260 年，比《番汉合时掌中珠》的刊刻（1190）晚 70 年。《蒙古字韵》精确地描述了元代官话音系，为西夏语音格局的重建提供了可靠的音韵学参照框架。

《蒙古字韵》标注了元代官话音系，该系统是在皇家权威的笼罩下，网罗元代顶级的文字学音韵学人才，以“拼写天下一切文字”的目标构造出来的，与敦煌出土的古藏—汉文献相比<sup>②</sup>，理所当然

① 照那斯图、杨耐恩《蒙古字韵校本》，民族出版社 1987 年。  
② Thomas, F.W., *Tibetan Literary Texts and Documents Concerning Chinses Turkestan*, 2003.3, pp.31—114, 231—265.

应该置于专题库的关键位置。《蒙古字韵》是跨跃 12 世纪西夏语音格局重构之河的坚固桥梁，其数据库表结构如下：

| 序号 | 字段名称 | 数据类型     | 长度解析                 | 注 释          |
|----|------|----------|----------------------|--------------|
| 1  | 文献代码 | CHAR(8)  | 卷号页码 4 页面行号 2 汉字序号 2 |              |
| 2  | 类型   | CHAR(6)  | 大字类型 2 小字类型 2 章节类型 2 | 东/庚/./百家姓*   |
| 3  | 八思巴字 | CHAR(12) |                      |              |
| 4  | 调类   | CHAR(2)  |                      | 平/上/去/入      |
| 5  | 汉字   | CHAR(32) |                      |              |
| 6  | 声类   | CHAR(4)  | 声纽 2—守温三十六字母 2       | 见/溪/群/./日**  |
| 7  | 韵类   | CHAR(8)  | 韵摄 2—等 2—调 2         | 东/庚/阳/./麻*** |
| 8  | 音节   | CHAR(16) |                      |              |
| 9  | 拼音   | CHAR(16) |                      |              |
| 10 | 英译   | CHAR(64) |                      |              |

表 3. 蒙古字韵文献的字段结构表

\* 章节类型：01 东/ 02 庚/ 03 阳/ 04 支/ 05 鱼/ 06 佳/ 07 真/ 08 寒/ 09 先/ 10 萧/  
11 尤/ 12 覃/ 13 侵/ 14 歌/ 15 麻/ 18 百家姓；

\*\* 声母类型：21 见/ 22 溪/ 23 群/ 24 疑/ 25 端/ 26 透/ 27 定/ 28 泥/  
31 知/ 32 彻/ 33 澄/ 34 娘/ 35 帮/ 36 滂/ 37 并/ 38 明/  
41 非/ 42 敷/ 43 奉/ 44 微/ 45 精/ 46 清/ 47 从/ 48 心/  
51 照/ 52 穿/ 53 床/ 54 审/ 55 晓/ 56 匣/ 57 影/ 58 喻/  
61 来/ 62 日/ 63 邪/ 64 禅。

\*\*\* 韵母类型：01 东—01—04 等—01 平 02 上 03 去 04 入/  
02 庚/ 03 阳/ 04 支/ 05 鱼/ 06 佳/ 07 真/ 08 寒/ 09 先/ 10 萧/  
11 尤/ 12 覃/ 13 侵/ 14 歌/ 15 麻；

（四）夏梵藏汉语料的加工功能

自从 N.A.Nevsky 发布藏文注音的西夏文资料<sup>①</sup>以来，前辈学人从藏文佛经中寻求对音，进而再从梵文的真言、咒语和陀罗尼对勘西夏文字的读音。这一学术路径的最早开拓者和实践者之一就是钢和泰（Alexander von Stael—Holstein,1877—1937）<sup>②</sup>。这种夏藏、夏梵对照的古文献，由于其形式多样不拘一格，采用语料文档格式存储是最有效的。不过，其标记过程一般采用手工方式，这样就可以针对灵活多样的文本做出灵活的处理。除了标准的词类标记之外，还要增加专题库的专用标记，以便为西夏语音格局的重建提供深加工的切入点。

美国华盛顿大学邵瑞祺教授所做的键陀罗文佛教文书（橡树皮介质）的研究，<sup>③</sup>为夏梵双语文献语料文档的构建提供了有益参考。在梵文文献的研究中，键陀罗文和巴利文是不可或缺的参照系。有了这些文本的精细研究成果，对梵文古文献语料的标记、标注、检索、统计、加工方法等等，提供了卓有成效的指南。

台湾中研院史语所的林英津教授采用逐词逐句遍历语料的方法，对西夏文本做出的研究，包括西

① N.A.Nevsky, *A Brief Manual of the Sihia Characters with Tibetan Transcriptions*, Research Review of the Osaka Asiatic Society, No.4, 1926.

② A.von Stael-Holstein 《音译梵书与中国古音》，《国学季刊》第一期，1923 年。

③ Richard Salomon: *A Gandhari Version of the Rinoceros Sutra*, the University of Washington Press, 2000. pp.105—189.



夏语译《孙子兵法》和《真实名经》<sup>①</sup>等，是我们构建西夏古文献语料库的范本。从林教授的煌煌巨著中，可以得出西夏古文献语料文档的标记、标注以及加工原则。由此原则出发，在系统设计上采用3层架构：应用层、接口层、实现层。接口层抽象出一套通用的专题库访问接口，能够有效隔离上层应用与底层语料存储。

西夏音韵学专题库的古文献语料加工功能可以概括如下：

- ① 采用基于XML的语料标注方式，利用XML结合音韵空间和概念空间标注信息，简化标注过程，提供句群切分和语境单元标注工具，将标注推向了语境单元层次；
- ② 采用全文索引技术实现了西夏语料的全文检索，并提供了音韵特征检索方式：基本检索、高级检索、XQuery检索、SQL标准检索等；
- ③ 采用常规统计和音韵特征统计功能，设计和实现了3种音韵特征统计模式：数量统计、比值统计、特定属性分布统计、自定义分布统计等。

在此基础上研制机器辅助标注，利用已有的标注语料，采用最大熵模型来解决语义块切分问题，采用基于实例的方法解决句类判断问题，进而建立句类重组语料文档。依托于基本语料文档，按照句类重组熟语料，并开发错误反馈和难点标注功能。依托语言学的广阔背景，以西夏音韵学专题为基础平台，重建西夏语音格局，进而拓展到西夏词句的计算机语音模拟，最终实现特定西夏文本的语音合成输出。

## 五、结 论

西夏音韵学专题库，以音韵学古文献为主轴，以相关古文献语料为分支，广泛采集并筛选12世纪前后西夏境内以及周边地区的古文献语料，包括内部语料、共时语料和历时语料。

首先，从西夏内部语料数据的采集着眼，《同音》、《文海宝韵》和《五音切韵》等音韵学文献，为构建西夏音韵学专题数据提供了可靠的素材，在优化了专题数据库结构之后，这些语料如虎添翼，在主支分明的结构树上，可以生成基于计算语言学的语音模拟与重构算法。

其次，从共时语料数据的层面来看，12世纪前后西夏境内诸异族语言的影响，尤其是像汉语、藏语等这些强势语言在西夏语语音演化的进程中，扮演了一个非常重要的角色。以《番汉合时掌中珠》为专题库的枢纽，协同阿尔泰诸语言文献语料，划出了西夏语音格局的边界。

最后，从历时语料数据的角度展望，现行羌语支诸语言发展不平衡的现状，为我们提供了一条羌语支语言演化的路径。采用计算语言学的研究方法，把《蒙古字韵》置于专题库的关键位置，关联全部的古文献语料，依托语言学的广阔背景，重建西夏语音格局，进而拓展到西夏语词句的计算机语音仿真，最终实现特定西夏文本的语音合成输出。

（作者通讯地址：同济大学西域古文献研究所 上海 200092；同济大学西域古文献研究所 上海 200092）

---

① 林英津《西夏语译〈真实名经〉释文研究》，《西夏研究》第8辑，中国社科出版社2008年。