doi:10.13440/j.slxy.1674-0033.2022.02.013

基于出土文献数据库的集外字数字化处理方法研究

唐杰,刘铭,陈镱文

(西北大学 科学史高等研究院,陕西西安 710127)

摘 要:相比较现代汉语和传世文献数据库而言,出土文献数据库中的集外字占比较大。因此在利用数字人文手段对出土文献文本处理时,集外字在数据库中的存在形式及参与方式将直接影响信息处理。参考藏文、甲骨文、西夏文的数字化文本的处理方式,提出一种适应于出土文献集外字在文本信息处理中的程序化方法。该方法不仅提高了出土文献数据库中的数据完整性,也可使得以往不能被用于文本信息处理的集外字参与至自然语言信息处理技术中。利用目前主流的分词工具jieba进行试验,结果显示该方法在出土文献的文本信息处理中是有效的。

关键词:出土文献;集外字;文本信息处理;分词

中图分类号:G230.7 **文献标识码**:A **文章编号**:1674-0033(2022)02-0073-06

引用格式:唐杰,刘铭,陈镱文.基于出土文献数据库的集外字数字化处理方法研究[J].商洛学院学报,2022,36(2):73-78.

Research on Information Processing of Character out the Set Based on Ancient Turpan Document Database

TANG Jie, LIU Ming, CHEN Yi-wen

(Institute for Advanced Studies in History of Science, Northwest University, Xi′an 710127, Shaanxi)

Abstract: Compared with Modern Chinese and Handed Down Literature Databases, the proportion of Character out the Set in the Ancient Turpan Document Database is relatively large. Therefore, when using Digital Humanities to process the texts of Ancient Turpan Documents, the existence form and participation mode of Character out the Set in the database will directly affect the information processing. By referring to the processing methods of digitized texts in Tibetan, Inscriptions-Bone Inscriptions, and Xixia Characters, an attempt is made to propose a programmatic method suitable for the processing of text information of Character out the Set in the collection of Ancient Turpan Documents. This method not only improves the data integrity in the Ancient Turpan Document Database, but also enables the Character out the Set that could not be used for text information processing in the past to participate in the natural language information processing technology. Experiments are carried out using the current mainstream word segmentation tool, named jieba, and the results show that this method is effectiv for the Ancient Turpan Document Database.

Key words: Ancient Turpan Document Database; Character out the Set; Text Information Processing; word segment

收稿日期:2022-01-08

基金项目:陕西省重点研发计划项目(2019ZDLGY17-03)

作者简介:唐杰,男,陕西渭南人,硕士研究生

自 20 世纪 80 年代始建“古汉语语料库”以来，经过 30 余年的积累，现已整理出数亿字的古籍数据库[1]。现有古籍语料库中含较多字符集以外的文字，即集外字。该类文字由于无法被 OCR 识别或通过键盘直接录入，在一定程度上影响了古籍的数字化及其转换与检索[2]，因而成为古籍数字化的重点工作之一；又由于目前的集外字技术不能适用于计算机的自然语言信息处理技术，也成为古籍数字化中汉字处理难点之一。出土文献数据库是关于中国出土文献简、牍、帛书的封闭式数据库，如《简帛金石资料库（全文）》《引得市数据库》《汉代简牍数位典藏》《汉达文库》《瀚唐典籍》等[3]，其中存在大量的避讳字、异体字、俗体字、生僻字，多以集外字为主。《瀚唐典籍》采用替换法将集外字认定为集内字，字库所用字符集编码为 Unicode5.0 字符集[4]，《引得市数据库》采用造字法建立集外字字库，字库所用字符集编码为 Unicode 字符集编码，参考中央研究院汉字部件检索系统，利用部首结构来查询集外字[5]。香港理工大学开发的“中文古籍网上出版平台”采用描述法对集外字进行描述，字库所用字符集编码为 Unicode 编码[3]。目前，这些数据库对集外字的信息处理均集中在解决集外字录入、显示及检索功能，尚未关注集外字在文本信息处理中的应用。本文基于课题组建立的《中国出土文献数据库》中的集外字字库，提出出土文献集外字在文本信息处理中的程序化方法。该方法基于《汉典》网站及相关工具书对先秦至晋期间出土文献的集外字进行整理；使用造字法将所造集外字存储至 Unicode 编码私用区码位，建立集外字字库；结合四角号码检字方案生成输入法码表，借助多多输入法生成器生成集外字输入法。并以里耶秦简为例，探讨了分词工具对文本的处理效果。

1 特殊文本的集外字信息处理方法

少数民族文字、西夏文字、甲骨文字相关的数字化研究工作都是从 20 世纪下半叶开始的，在这之前，我国使用的是 GB2312-80 字符集，少数民族文字、西夏文字、甲骨文字都属于集外字。通过分析此类集外字的信息处理技术并借鉴其处理方法，可对出土文献的集外字处理提供帮助。

1.1 藏文字的信息处理

藏文由辅音字母、元音符号和标点符号构成[6]，其同时包含横向拼写及纵向拼写，藏文是从 20 世纪 90 年代开始编码研究工作，1997 年发布了《信息交换用藏文编码字符集》，该标准包含 192 个编码点、169 个编码字符，是按照拼音文字编码规律对藏文进行编码[7]。藏文是通过使用造字法，在长宽不等的点阵中造字，字体有点阵字、矢量字、曲线轮廓字，将其保存为 Unicode 编码，Unicode 编码范围为 0F00~0FFF，其键盘输入法有字丁、音节和词组的形式，通过藏文输入法实现藏文自动排版，实现藏文的输入输出。目前藏文和其他少数民族文字已收录在 CJK 字符集中[8]，已被规范为集内字，信息处理更加方便。

1.2 西夏文字的信息处理

西夏文是记录西夏党项族语言的文字，其文字特征类似于汉字，同为方块字。自 20 世纪后半叶，西夏文数字化研究逐渐展开，宁夏大学开发并研制出了《汉夏字处理及电子字典》，创立了西夏文字字符集[9]。景永时等[10]通过使西夏字与 GB2312-80 标准汉字共用同一编码，制作了西夏文字库，此方法在处理西夏文与汉字同框的文本时需不断切换字库。马希荣等[11]采用位面技术在 GBK 的用户自定义区分配西夏字编码，避免了与汉字或其他字符的码位冲突问题，但其本质上并没有解决西夏文字符编码问题，字库不便应用于文本信息处理。2016 年 6 月发布的 Unicode9.0 版本收录了西夏文字符，使西夏文字规范为集内字，解决了占用汉字码位、夏汉同屏共存等问题，对于西夏文字库在文本信息处理方面的应用具有非常重要的作用[12]。著名西夏学研究专家李范文[13]根据西夏字结构编排了西夏字的四角号码，使得西夏字检索更加方便，为西夏文字数字信息化开辟了捷径。

1.3 甲骨文字的信息处理

甲骨文字形的特点是笔画繁多、无法区分、构造复杂[14]。自 1990 年开始就有专家关注甲骨文的信息化技术，徐松[15]开发了“甲骨文象形码编码系统”，可实现甲骨文笔画检索。江铭虎等[16]同时制作两套甲骨文字库，通过区位码和拼音形式输入甲骨文。顾绍通等[17]根据甲骨文字形的拓扑结构建立了甲骨文输入法，实现甲骨文的字形和拼音输入。刘志祥等[18]建立了字形编码型甲骨文 6 位数字码输入法，类似于汉字检索的四角号码，实现精确的甲骨文字的输入输出。刘永革等[19]通过对甲骨文的笔画特征进行分析，将构成甲骨字的笔画归纳为九种笔画，在此基础上采用香港中

文大学的甲骨文字库设计了甲骨文笔画输入法。

2 字符集编码及集外字处理

2.1 字符集编码

字符集是遵循国家或国际标准，对每一个字符进行定义的唯一代码[20]，常见的汉字字符集有GB2312，BIG5，GBK，GB18030，Unicode [21]，其中Unicode 编码字符集是国际标准字符集，可实现跨语言文本信息转换[22]。

集外字是指字符集以外的字，不采用特定的技术方法无法对集外字进行录入、处理及显示[23]，字符集的选择与集外字的数量呈负相关，故在建立字库时优先选择收录字符最多的 Unicode 字符集，其满足共享、国际通用的条件，同时也是目前古籍数字化项目最常用的字符集。

2.2 集外字处理方法

在古籍数字化项目中常见的集外字的处理方法大致分为造字法、替换法、描述法[24-25]。

造字法是在字符集的自定义区为集外字定义编码，这些编码与集外字的字形一一对应。优点是只要有字符集自定义编码区的支持，造字的检索与显示和集内字完全相同，缺点是自定义编码区只有 13 万余个码位[26]，如果不加限制地造字，码位很快会消耗殆尽，且不同的古籍数字化项目对私用区码位的使用可能完全不同，若同时使用这些古籍数字化项目，可能发生私用区编码冲突。

替换法是将集外字变换为其他可以输入的形式，如符号、图形和集内字。此方式的优点在于录入方便简洁，但是缺点也很明显。第一，集外字统一替换为某个符号，这些符号基本没有检索意义。第二，替换符号未能保留集外字的任何信息，当字符集变更时，集外字无法管理。第三，在不清楚替换规则的情况下，用户无法理解替换的意义。

描述法是将集外字表示为一个字符串，这个字符串描述了集外字的字形。优点是可使用标准化的描述符序列对汉字的构造进行说明，解决了自然语言描述法的不规范问题。缺点是很多生僻字结构极为复杂，拆解困难，一种字存在多种描述方法，且描述后的字不是一个 Unicode 编码，而是一组编码，例如  字，需要十三个编码才可完整描述此字[23]，其缺点有：占字节较多、不利于文本信息处理、需额外软件支持、所造字形与原字符存在一定差距。

相对而言，造字法是建立出土文献数据库的较好选择。第一，出土文献语料库是封闭式语料，其字数有限且相对于传世文献与自然语言而言较少，据目前统计的已释读的出土文献数据，集外字有 2 万余个。第二，造字法可将所造集外字保存为 Unicode 字符集编码，将集外字转为集内字，目前 Unicode 供用户自定义编码的码位有 137 468 个码位，可自由编辑 137 468 个集外字。第三，计算机的文本信息处理是识别文字所对应的字符集编码，要求集外字“一字一形一码”。造字法满足以上三种限制条件，故本文选择造字法建立出土文献集外字字库。

3 出土文献集外字输入法方案及实现

3.1 出土文献的集外字整理

计算机的应用范围越来越广泛，对汉字进行数字化已经成为了中文信息处理的必要前提。出土文献中包含大量的俗体字、异体字等，其中不乏集外字，给研究者合理规范使用文字带来了诸多不便。因此，建立出土文献数据库的必要工作就是对出土文献的集外字进行整理，但古籍版本众多，且没有非常严格的统一标准，故应选择专业的文物校释小组和权威作者所著的释文进行数字化和整理工作。基于《说文解字》《康熙字典》此类工具书及《汉典》网站，筛选查找《清华大学藏战国竹简》《望山楚简》《天水放马滩秦简》《里耶秦简牍》（壹与贰）《张家山汉简》《悬泉汉简》等先秦至晋的出土文献中的集外字。

以里耶秦简集外字整理工作为例，出土文献释文书籍选择湖南省文物考古研究所编著的《里耶秦简牍》，其释文一般按照原文字形释写，不识字是按照原样摹写。工具书选择陈伟主编的《里耶秦简牍校释》。在里耶秦简两卷的集外字整理过程中，对例如 9-475 中的字“  ”“  ”，此字是按照原样摹写的字，将其认定为集外字进行收录；例如 9-478 中的字“  ”，此字是“  ”下侧无法识别，所以对其不进行收录；对文本中的符号例如“  ”“  ”，不进行收录；例如 8-181 中的“  ”字，在汉典网中未收录此字，此字也不存在于 Unicode 字符集中，所以将其收录为集外字。根据整理已出版的两卷《里耶秦简牍》中的集外字，其在全文不重复字数的比重约为 11.2%，其中集外字个数 202 个，两卷不重复字数约 1 800 个。

3.2 出土文献集外字字库的建立

根据现筛选查找到的出土文献集外字，在现有的藏文字库、西夏文字库及甲骨文字库建立方

法的基础上构造出土文献集外字字库。该工作主要分为两步：第一步，采用造字法描绘集外字，将集外字描绘成汉字字符形式。藏文、西夏文及甲骨文对其文字采用了“造字法”，即利用造字软件采用描绘的形式将文字描绘在画板上。鉴于此，本文也根据造字法利用计算机描绘出土文献集外字，使得出土文献集外字成为“汉字”形式。第二步，由于藏文、西夏文及甲骨文在建立字库时，是将描绘好的字符以 Unicode 字符集编码的形式储存至计算机中，故本文也通过参考该方法，将所描绘的集外字存储至 Unicode 编码私用区码位。由此，基于以上两项工作建立起出土文献集外字字库。具体的操作步骤如下。

利用 High-Logic 公司的 FontCreator 曲线轮廓造字软件，根据二次 β 样条曲线拟合算法，自动将扫描的集外字图形数字拟合成尽可能接近原样摹写的集外字，可通过调整文字的轮廓点、线、角度及位置，描绘出较为理想的集外字并保存在对应的 Unicode 编码私用区码位。

现有较常用的出土文献集外字字库有“引得市古文字缺字资料库”“古联瀚字输入法字库”，但是建立出土文献集外字字库无法直接采用其结果，原因有二：其一，不能包含先秦至晋期间所有出土文献中的集外字。其二，无法对其字库进行动态管理。本文所建立的集外字字库优势在于：第一，集外字数据更全面，可研究范围及角度更广，对每篇出土文献中的集外字进行整理并梳理成表。第二，对字库实现动态管理，根据出土文献的更新和 Unicode 字符集的更新，对字库中的字进行增加或删改。

3.3 生成集外字输入法

在西夏文字库的建立及夏汉通输入法生成过程中，西夏文字存在拼音难以读写且结构复杂的特点，由于其是仿造汉字而创制的，即以偏旁部首组成方块字，因此研究者利用此特点建立了西夏文四角号码检字法，对所有文字进行四角号码检字编码，生成了西夏文输入法码表。现有的检字方案可分音码、形码两种[27]，四角号码检字法属于根据文字字形查找文字的形码检字方案，是用最多不超过 5 位的阿拉伯数字将所有汉字进行归类，此检字方案对于无法正确读写拼音及结构部首繁琐的文字效率极高，可在计算机中高效率录入此类文字。而出土文献集外字也存在与西夏文字类似的特点，故也可采用四角号码对其检字。出土文献集外字的检字码由 5 位阿拉伯数字组成，遵循《四角号码新词典》第十版的规则，对集外字进行编码拆分取号码顺序为左上角，右上角，左下角，右下角，附加码为左下角上方距离最近的笔形对应号码，例如：

“**宽**”的四角号码为：30212

“**旂**”的四角号码为：08240

本文首先利用四角号码检字法对字符编码，进而建立出土文献集外字的输入法码表，在输入法码表中需将“0123456789”用其拼音声母首字母“oyesxwlqbj”代替[28]。其次，按步骤将整理好的输入法码表文件导入多多输入法生成器中并设置相关参数，生成“出土文献集外字输入法”。图 1 为出土文献集外字输入法的构造流程图。

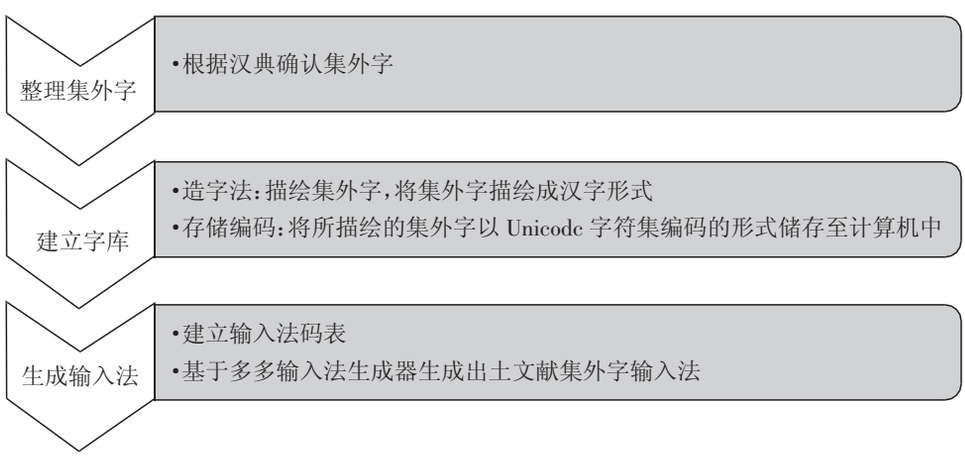

图 1 出土文献集外字输入法构造

4 集外字文本分词试验

里耶秦简于 2002 年在里耶古城一号井第一次出土，计三万八千余枚简牍；及于 2005 年在护城壕第十一号坑第二次出土，计五十一枚简牍。

简牍起止时间为秦始皇二十五年（前 222 年）至秦二世二年（前 208 年），内容为秦朝洞庭郡迁陵县遗留的公文档案，涵盖了当时社会的各个层面[29]。目前已出版《里耶秦简牍》第一卷和第二卷，两卷共包含四层，共含 6 050 条简，总字数约为 12.25 万字。本文在小组建立里耶秦简数据库的基础上，建立集外字字库，并应用于里耶秦简语料库的建设中。

中文分词是文本分类、信息检索、文本挖掘等中文信息处理工作中的难点和技术关键点，为验证所建字库及输入法可使集外字参与到文本信息处理过程中，本文将所造集外字录入语料库，以里耶秦简两卷文本数据库为例，测试集外字可参与到主流分词工具的分词过程。

试验选择目前自然语言分词的主流工具 jieba[30]。jieba 分词工具本身内含词典，该词典的主要内容是现代汉语及部分古代汉语，但也可根据用户需求选择是否添加自定义词典，以此保证分词的准确率。由于简牍语言以单字词居多及与现代汉语实体名词有较大区别的语言特殊性，故在分词时采取两种方案，以验证集外字字库及输入法建立的必要性。

试验数据选择里耶秦简第一卷的简 8–458，简文中包含集外字“**兖**”：

遷陵庫真□

甲三百卌九

甲**兖**廿一

鞮瞀卅九（第一欄）

胄廿□

弩二百五十一

臂九十七

弦千八百一（第二欄）

矢四萬九百□

戟二百五十（第三欄）

试验分为两个方案：方案一，基于 jieba 分词工具的默认精确切分模式下，不添加用户自定义词典进行分词。方案二，添加自定义的包含集外字的分词词典进行分词。试验结果如表 1。

表 1　集外字文本分词试验

方案	分词
方案一	遷陵庫/真甲/三百/卌/九甲/**兖**/廿一/鞮/瞀/卅九胄/廿弩/二百五十/一臂/九十七/弦/千八百/一矢四萬/九百/戟/二百五十
方案二	遷陵/庫/真/甲/三百卌九/甲**兖**/廿一/鞮瞀/卅九/胄/廿/弩/二百五十一/臂/九十七/弦/千八百一/矢/四萬九百/戟/二百五十
标准	遷陵/庫/真/甲/三百卌九/甲**兖**/廿一/鞮瞀/卅九/胄/廿/弩/二百五十一/臂/九十七/弦/千八百一/矢/四萬九百/戟/二百五十

从表 1 可见，第一，根据分词结果，用 jieba 分词工具对含有集外字的文本直接进行分词时，集外字“**兖**”参与了分词。这说明了“甲**兖**”虽然包含集外字，但由于集外字字库及输入法的建立，这类集外字可参与至计算机对语言的信息处理工作中，最终使得集外字可被读取、被处理、被写入。

第二，“甲**兖**”在里耶秦简校释小组的校释中，被释为秦朝士兵作战时的一类防御类装备，但此名称在现代汉语中并未出现。故在方案一未添加自定义词典的分词结果中，该词被切分开，因为 jieba 自带词典中未检测到有“甲”一词。

第三，在添加自定义词典后，该词得以正确切分，这是因为“**兖**”保存在出土文献集外字字库中，已将其转换为集内字，有唯一的 Unicode 编码。计算机在对文本识别时所识别的为字符所对应的字符集编码，此处已成功识别“**兖**”字的 Unicode 编码，故可成功对集外字进行分词。这说明集外字字库的建立有利于建立出土文献词库，从而可将生语料转化为熟语料，为之后工作奠定可行性基础。

第四，方案一与方案二的分词结果略有出入，主要问题还是在 jieba 分词工具的内含词典主要是现代汉语，对于简牍这类特殊语言，其词库及数据库的建设还未完备，故分词的准确率较低。

5 结论

本文根据《中国出土文献数据库》的集外字字库提出了文本信息处理程序化方法，在实现集外字的数据库中显示和检索功能的基础上，为出土文献中集外字参与文本信息处理提供了可行性方案。

本文以里耶秦简为例，采用造字法将集外字转换为计算机可识别的编码并参与文本信息处理，并建立了出土文献集外字字库，提高了数据库的完整性。其次，结合四角号码，基于多多输入法生成器生成集外字输入法，并以自然语言处理技术中的基础任务——分词为例，利用已处理的文本数据进行切分，达到了较好的处理效果。

在数字人文背景下，本文利用该方法将适用于现代汉语与传世文献文本的信息处理技术泛化至出土文献文本中，不仅可直接应用于建立出土文献集外字字库及文本数据库，且为集外字参与古籍文本的计算机处理技术提供新方法。

参考文献：

[1] 马海丽,王曦.古籍数字化中计算器自然语言处理应用现状分析[J].古籍研究,2020,72(2):322-328.

[2] 陈力.中文古籍数字化的再思考[J].国家图书馆学刊,2006,15(2):42-49.

[3] 季培培.常见 10 种古籍全文数据库的比较研究[J].图书馆学研究,2020(20):71-80.

[4] 辛瑞龙,王雅坤.古籍数字化中汉字处理的现状、问题及策略[J].图书馆理论与实践,2017(9):103-107.

[5] 张玉金.出土文献语言研究:第三辑[M].广州:暨南大学出版社,2020:228-247.

[6] 尼玛扎西,李志蜀,拥措,等.实现计算机藏文快速输入的关键问题研究[J].电子科技大学学报,2009,38(1):102-107.

[7] 普顿,群诺,尼玛扎西.汉文和藏文在信息处理中的比较研究[J].西藏科技,2013(10):77-80.

[8] 尉迟治平.再论中文汉字字符集[J].语言研究,2020,1(40):78-89.

[9] 柳长青.西夏文古籍字库建立研究[C]//西夏学：第六辑——首届西夏学国际论坛专号(下).宁夏大学西夏学研究院，2010:204-210.

[10] 景永时.西夏文数字化的现状与未来[J].西夏学,2011(1):199-203.

[11] 马希荣,王行愚.基于汉字字形的西夏文字信息处理研究[J].计算机工程与应用,2001(21):26-27,95.

[12] 孟一飞.西夏文字数字信息化若干问题研究[D].北京：北京交通大学,2019:63.

[13] 李范文.夏汉字典[M].增订版,北京:中国社会科学出版社,2008:1-20.

[14] 马小虎,杨亦鸣.甲骨文轮廓字形生成技术研究与通用甲骨文字库的建设[J].语言文字应用,2004(3):105-111.

[15] 徐松,胡金柱.甲骨文象形码输入法的实现[J].华中师范大学学报(自然科学版),1995,29(3):299-302.

[16] 江铭虎,邓北星.甲骨文字库与智能知识库的建立[J].计算机工程与应用,2004(4):45-47,60.

[17] 顾绍通,马小虎,杨亦鸣.基于字形拓扑结构的甲骨文输入编码研究[J].中文信息学报,2008(4):123-128.

[18] 刘志祥,刘晓戎.甲骨文六位数字码检索字库[M].成都：四川辞书出版社,2019:1-24

[19] 聂艳召,刘永革.甲骨文自由笔画输入法[J].中文信息学报,2010,24(6):103-107.

[20] 尉迟治平,汤勤.论中文字符集、字库及输入法的研制[J].语言研究,2006(3):63-66.

[21] 石丽怡.MySQL 数据库字符集的问题研究[J].电子技术与软件工程,2020(12):149-150.

[22] 苗军.Unicode/XML 在电子出版物中的实现[D].天津：河北工业大学,2002:3.

[23] 郑军,林民.一种面向集外字输入的手写汉字轮廓提取算法[J].内蒙古师范大学学报(自然科学汉文版),2008,37(6):750-752.

[24] 肖禹.古籍数字化中的集外字处理问题研究[J].图书馆研究,2013,43(5):27-30.

[25] 张翼飞.古籍数字化中的字符集问题与解决方案[J].出版发行研究,2016(3):77-80.

[26] 高晶晶.中医古籍数字化生僻字的处理[J].中国中医药图书情报杂志,2014,38(3):28-30.

[27] 杨巧.输入法字库切换功能与分级编码研究[D].武汉:华中科技大学,2018:22-23.

[28] 柳长青,杜建录.西夏文四角号码输入法研究[J].宁夏大学学报(自然科学版),2010,31(4):324-328.

[29] 湖南省文物考古研究所.里耶发掘报告[M].长沙:岳麓书社,2007:179-180.

[30] 石凤贵基于 jieba 中文分词的中文文本语料预处理模块实现[J].电脑知识与技术,2020,16(14):248-251,257.

（责任编辑：张国春）