分类号：G250.7　　学校代码：10697

密　级：公开　　学　　号：201920531

西北大学
Northwest University

硕 士 学 位 论 文

MASTER' S DISSERTATION

基于出土文献数据库集外字字库的数字化处理方法研究

学科名称：科学技术史

作　　者：唐　杰　　指导老师：陈镱文

西北大学学位评定委员会

二〇二二年六月

Research on Digitalization of Character out the Set based on Ancient Document Database

A thesis submitted to

Northwest University

in partial fulfillment of the requirements

for the degree of Master

in Digital Humanities

By

Tang Jie

Supervisor: Chen Yiwen Professor

June 2022

西北大学学位论文知识产权声明书

本人完全了解西北大学关于收集、保存、使用学位论文的规定。学校有权保留并向国家有关部门或机构送交论文的复印件和电子版。本人允许论文被查阅和借阅。本人授权西北大学可以将本学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存和汇编本学位论文。同时授权中国科学技术信息研究所等机构将本学位论文收录到《中国学位论文全文数据库》或其它相关数据库。

保密论文待解密后适用本声明。

学位论文作者签名：唐杰 指导教师签名：陈鑫文

2022 年 6 月 11 日　　2022 年 6 月 11 日

西北大学学位论文独创性声明

本人声明：所呈交的学位论文是本人在导师指导下进行的研究工作及取得的研究成果。据我所知，除了文中特别加以标注和致谢的地方外，本论文不包含其他人已经发表或撰写过的研究成果，也不包含为获得西北大学或其它教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示谢意。

学位论文作者签名：唐杰

2022 年 6 月 11 日

摘要

集外字在古籍文献的重要性，不仅体现在文字自身的学术研究价值，而且在古籍文本的信息处理中有至关重要的作用。这些集外字不论是对于汉字规范进程或是特定时期的历史研究都具有非常重要的研究价值。

相比较现代汉语和传世文献语料库，出土文献语料库中的集外字所占比重较大。利用数字人文手段对出土文献文本进行处理时，集外字在语料库中的存在形式及参与方式将直接影响其文本信息处理。

本文进行的统计包括现已出土并释读出版的全部出土文献，涉及 52 种，其中先秦时期的有 9 种，集外字占比约为 3.5%-8.9%；秦时期的有 7 种，其集外字占比约为 0.3%-0.5%；两汉时期的有 34 种，其集外字占比约为 0.1%-0.8%；三国至魏晋时期的有 2 种，其集外字占比约为 0.07%-0.2%。

以现有集外字为研究对象，首先参考甲骨文、引得市等数字化文本处理方式，对以里耶秦简集外字为例进行造字，并建立出土文献集外字字库；其次，在此基础上结合前人研究生成出土文献集外字输入法，并依此提出一种适应于出土文献集外字在文本信息处理中的普适化程序化方法。该方法不仅提高了出土文献语料库中的数据完整性，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过程中。其次，本文利用中文分词手段对普适化程序化方法下被处理的集外字进行文本信息处理的验证。

最后，将里耶秦简的集外字完善至其语料库中，再利用其语料库进行三种不同方法下的分词实验，分词实验采用了基于规则的分词方法、基于统计的分词方法以及主流分词工具 jieba、Hanlp 的方法进行了实验。实验结果显示，对普适化程序化方法下被处理的集外字，可被直接应用于自然语言的处理过程中。因此该方法对于以里耶秦简为例的语料库构建具有有效性及可行性，并尝试将此方法推广至全部出土文献集外字的文本信息处理工作中。

关键词：出土文献，集外字，文本信息处理，分词

ABSTRACT

The importance of Character out the Set in ancient books is not only reflected in the academic research value of the characters themselves, but also plays a vital role in the information processing of ancient books. These sets of Character out the Set have very important research value for the process of Chinese character standardization or historical research in a specific period.

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.

The statistics in this paper include all the Ancient Turpan Document that have been unearthed and interpreted and published, involving 52 kinds, of which there are 9 kinds in the pre-Qin period, and the proportion of Character out the Set is about 3.5%-8.9%; The proportion is about 0.3%-0.5%; there are 34 kinds in the Han Dynasty, and the proportion of Character out the Set is about 0.1%-0.8%; there are 2 kinds from the Three Kingdoms to the Wei and Jin Dynasties, and the proportion of foreign words is about 0.07%- 0.2%.

Taking the existing set of Character out the Set as the research object, firstly referring to Inscriptions -Bone inscriptions, Index and other digital text processing methods, the Character out the Set of Qin Bamboo Slips Collection are used as an example to create characters, and the unearthed literature collection of Character out the Set is established; secondly, on this basis, combined with The predecessors have studied the input method of generating Character out the Set in Ancient Turpan Document, and based on this, they have proposed a universal programming method suitable for the processing of text information. This method not only improves the data integrity in the corpus of Ancient Turpan Document, but also enables out-of-set words that could not be used for text information processing in the past to participate in the natural language information processing process. Secondly, this paper uses the Chinese word segmentation method to verify the text information processing of the out-of-set words processed under the universal programming method.

Finally, the Character out the Set of Qin Bamboo Slips are perfected into its corpus, and then the corpus is used to conduct word segmentation experiments under three different methods. The word segmentation experiments use the rule-based word segmentation method, the statistics-based word segmentation method and the mainstream word segmentation tool jieba , Hanlp's method was experimented. The experimental results show that the out-of-set

words processed under the universal programming method can be directly applied to the natural language processing process, so this method is effective and feasible for the construction of corpus with Qin Bamboo Slips as an example, and the Try to generalize this method to the text information processing of all characters outside the collection of Ancient Turpan Document.

Keywords: Ancient Turpan Document Database, Character out the Set, Text Information Processing, Word Segment

目录

第一章 绪论

1.1 研究背景及意义

集外字是汉字系统中特殊的一类汉字，它在汉字发展的各个历史时期均有存在。广义地说，集外字是指所用字符集以外的字，若不采用特定的技术方法，将无法对集外字进行录入、处理及显示[1]，即我们使用电脑输入法时不能直接输入、输出及显示的文字。目前，由于中国古籍数字化的要求，古籍数字化项目中最常用的字符集是Unicode 字符集，而其中部分文字是不能利用现有的计算机输入法写入，这类未收录进字符集的文字为集外字。

从语言类型上来看，集外字的范围主要包括两类，一类是我国古代的甲骨文、金文和书写在简牍、帛书上的部分文字及在传世古籍中较少出现的部分文字；另一类是民族语言文字，包括了西夏文和藏文等，还有一类是外籍语言文字，例如楔形文字。甲骨文作为我国最古老的成熟文字系统，其文字构成方式多样，笔画根据刻痕深浅不一，与现代中文字符构建相差较大[2]。西夏文是西夏党项族独有的记录其语言的文字，文字字形由点、撇、捺、横、竖、折、提等笔画构成，类似于汉字的方块字。但与现代汉语也有显著不同：其一，字结构规律性差，不能利用汉字的结构偏旁发原理描述字；其二，左右互换的相似字较多，这些相似字是利用部位左右互换而形成的近义字[3]。藏文是一种由拼音组成的文字，是与不同于汉语、西夏文、甲骨文等由笔形构成的文字。其文字整体由三大部分组成，分别是辅音字母、元音符号和标点符号[4]；其次，在书写藏文时有横向拼写及纵向拼写两种方式。每种拼写方式具有其严格且完整的并由以上三大部分组合而成的特殊排列规则；第三，其文字的字母在音节中由其独特的结构组合而成，其结构字母分为基字(1)、上加字(2)、下加字(3)、前加字(4)、后加字(5)和再后加字(6)和元音符号(7)七个部分[5]。

在 Unicode 字符集建立之初，上述甲骨文、西夏文及藏文均是集外字，但由于数字化项目的建设及研究学者对这三类文字的整理及规范，目前已将甲骨文、西夏文及藏文收录进 Unicode 字符集中，从而实现了“集外字”转化为“集内字”的过程。但金文、书写在简牍、帛书上的部分文字及在传世古籍中的部分文字在现阶段研究中仍是以集外字的形式被呈现。

这些集外字对于汉字规范进程及特定历史时期的探索都具有非常重要的研究价值。对于古籍中的集外字，其产生和存在必有其特殊的原由和意义。集外字文字种类

较多，其中包含部分避讳字、部分异体字、部分俗体字、部分生僻字、部分专造字[6]等。其中有些文字已被学界释读，并具有其含义及意义，但大多数集外字还未被释读，或在学界还未有统一的认识。在汉字的演变过程中，从甲骨文到隶书再到楷书，同一个字的字体往往会不断变化[7]，在文本传抄过程中有可能会出现讹抄、讹写或是特殊历史时期原因干扰，最终产生集外字，其没有非常统一的文字字形格式。

简牍、帛书此类出土文献，年代涵盖春秋、战国、秦、汉、三国及魏晋，蕴含丰富的历史文化信息，包含了政治、经济、文化、军事、法律、哲学等诸多领域[8]。目前关于简帛相关的研究多为微观研究，而数字人文范式的出现，为学者们研究出土文献提供了新思路和路径。由于出土文献的数字人文研究中，需要将其经数字化后保存至计算机内，再形成语料库。现代汉语语料库集外字含量极少，故具有极大地开放性及广泛传播特性；传世文献语料库文字体量巨大，因此集外字相较其数量所占比重较小。而出土文献语料库同传世文献语料库类似，均是封闭式语料，虽然已有部分语料库面世，但内容全面、系统且具备文本信息处理技术功能的语料库还有待建设。

在自然语言处理技术中，利用计算机式的信息处理技术对小文本、大文本、超大文本具备高效率、多维度等的优势[9]，因此，对数字化古籍文本研究已经成为一种学术趋势。目前，已有学者将自然语言处理技术应用至传世文献的文本中，并已取得了较为可观的成果[10]，因此，我们尝试将这种方法拓展至出土文献文本中。此方法不仅可为传统的出土文献史学研究提供技术补充，还能够应用计算机在自动挖掘古籍文本的信息中提供新问题。

目前已有部分学者对此问题进行讨论，但其成果目前仅能支持集外字的转化与检索功能，且其所含集外字体量较小。因此，本文在整理全部出土文献集外字工作的基础上，尝试提出新的普适化程序化方法，在支持更大量的集外字可更易于被转化及检索的基础上，可被直接应用于自然语言处理技术中，完成了出土文献全文本的信息处理过程。

1.2 国内外集外字研究现状

集外字信息处理的研究概况主要分为两方面，第一方面是古籍语料库及特殊文字字库建立的相关研究，第二方面是特殊文字的输入法的相关研究。

（一）常用的中国古籍语料库有《中国基本古籍库》《文渊阁四库全书（电子版）》《瀚唐典籍》《引得市》，其语料库建设现状包括：

《中国基本古籍库》涵盖的典籍文献的时间上至先秦时期，下至民国时期，收录

的文献种类包含了经史子集等[11]。其文献种类丰富，并提供了通行版及重要版的全文图像是中国有史以来最大的历代典籍总汇的语料库。该库于 1997 年开始规划，至今已纳入文字 18 亿字。该文字格式是 ABT 数据格式，并根据古籍制作了其专有字库，其所收录的字数已超过 13 万。其字库特色有：第一，收录的汉字多，涵盖了 Unicode12.0 所定义的已有的 CJK 汉字以及 Unicode 未定义古籍集外字；第二，字体范围广，包含宋体风格的楷书汉字、小篆、日文等外文字符；第三，字库主要以 Unicode 字符集编码为主，同时又有 CJK 字符集编码。

《文渊阁四库全书（电子版）》包含了《四库全书》的所有内容，其字库是委托方正公司建立，该字库使用字符集编码为 CJK 编码，字库中自定义编码 4296 个，共收录字数 31780 个。对于集外字，其处理方法分为 4 种：第一种是采用替换法替换文中的异体字；第二种是对于模糊难识别的字用符号“□”代替并添加说明；第三种是对于无法认同的集外字，保留其图像并添加字形描述，根据字形描述统计出高频集外字，对高频集外字进行造字，对低频集外字用构字符或“□”代替。第四种是用图像替换集外字[12]。

《瀚唐典籍》是采用超大字符集且图文并存的古籍语料库，语料库收录大量甲骨文、金文、简帛文、印章、石刻等出土文献资料，简帛类字库只有《郭店楚简》《上博楚简》《九店楚简》《武威汉简》《张家山汉简》《银雀山汉墓竹简》，其对外宣称最大特点就是文本中不包含集外字，使用替换法将所有集外字认定为集内字，字库所用字符集编码为 Unicode5.0 字符集[13]，使用时需安装微软的宋体-扩展 B-simsunb.ttf 字符集以及特定微软的字符软件才可以正确显示并使用特定字符。

《引得市》是古文字相关的字典索引资料库，收录了 238 种各类出土文献及传世文献，累计并制作出的索引语料库，并根据其收录文献建立缺字字库。该字库不仅包含了商周、战国、秦汉等文献文字，也涵盖了部分古文字、传抄古文、今文字、训诂、俗语词等文献中的缺字。对于集外字，其采用造字法建立集外字字库，V1.x 版本的 seal 字库中集外字的 Unicode 编码是按照 glyphwiki 上的顺序自动顺延产生，在最新一版的 seal 字库中，语料库建立者陈信良将每一个集外字编码表示为 sXXX-YYY，其中 XXX 取值范围为 001～215，代表康熙字典的 214 个部首，YYY 取值范围为 001～999，对应此部首下具体文字编号，并为每一个集外字编码映射一个 Unicode 编码。Seal 字库借鉴了中华字库的做法，编码码位选择未定义的第 10～13 平面(U+A0000～U+DFFFF)，参考中央研究院的汉字部件检索方式，可利用部首结构来查

询集外字。截至 2021 年 4 月，所发布的 seal_20210424 版本字库共 19175 字，数据整理来源于文字编、部分简帛文献、金文、古汉语字词典等[14]。

（二）特殊文字包括甲骨文字、西夏文字、少数民族文字以及楔形文字，此类文字现在已被规范为集内字，但在其未被收录进 Unicode 字符集前属于集外字行列，特殊文字的字库建立研究具体包括：

甲骨文字库建立的相关研究有：马小虎等[15]根据甲骨文的字形特点，提出了设计建设通用甲骨文字库的设想，即选用出现频次高、具有代表性的甲骨文字形构成通用字库，基于 Windows 平台，使用 Visual C++6.0 开发出甲骨文轮廓字形造字系统，建立了最初的通用甲骨文字库；江铭虎等[16]在同一项目中建立了两套字库，使用其中的现代汉字字符代替甲骨文字，建立了计算机处理的甲骨文字库；陈年福对甲骨文字形及字样进行整理并重新摹写，北京古联公司根据陈年福的摹写样本对其进行字符编码与字库制作[17]，字库中包含 1.3 万余甲骨文字原形，本方法优点在于字形统一且全面，应用范围较广。

西夏文字库建设的相关研究有：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[17]；景永时、贾常业等[18]利用替换法，将西夏字符用集内字代替，制作了“基于方正典码的西夏文字库”；马希荣等[19]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；21 世纪初 Unicode 组织将西夏文字收录到字符集内[20]，分配的码位为 U17000-U187EC，共计 6125 个，其中 U18800-U18AF2 码位存储的是西夏文部首。首先，该组织利用 Unicode 编码私用区部分对西夏文进行占位处理，在解决了占用汉字码位的基础上，同时也可将夏汉同屏共存的问题进行分解；其次，这种处理手段也可以使得西夏文成为规范的“集内字”。这种方法不仅推动了西夏文字库的建设工作，还对其文本信息处理阶段存在深远影响。

藏文字库的建立的相关研究有：普顿[21]等通过分析，对比了汉文和藏文在信息处理中的方法及其技术手段，并将其不同之处分为文字特点、编码方法、字库制作、输入法设计、文字排版方法及识别技术六大方面；同元藏文网站首先建立了藏文字库，并通过在用户电脑上预安装藏文字体安装文件，再将字体设置为 IE 浏览器内含字体，实现了网页中的藏文显示[22]。微软公司在 2007 年所发布的 Vista 操作系统中采用藏文编码国际标准 Unicode，实现了藏文与中文相同的系统内核处理。因其售价高且输入效率低，未得以推广[22]。目前藏文和其他少数民族文字已收录在 CJK 字符集中[23]，

已规范为集内字，信息处理更加方便。

楔形文字字库的建立的相关研究有：Unicode 组织在 Unicode5.0 版本中就已经将苏美尔楔形文字收录在 Unicode 字库中，楔形文字的码位分布在补充多语言平面（SMP）中，其码位为 U+12000-U+1254F[31]，属于集内字。Unicode 组织对于楔形文字的规范化及标准化，为楔形文字数据库建设以及数字化应用提供了便利。

（三）特殊文字的输入法建立研究

甲骨文输入法的建立的相关研究有：徐松[24]建立了“甲骨文象形码编码”系统，该系统可实现对甲骨文的输入及检索，检索方式多样，包含象形码检字、笔画检字和页码检字；江铭虎等[16]采用区位码和拼音的形式实现甲骨文的电脑输入及现代汉字与甲骨文的对应；顾绍通等[25]根据甲骨文字形的拓扑结构制作了甲骨文输入法，建立了甲骨文字形码表与拼音码表，通过笔形检字与拼音检字的方式实现甲骨文的输入输出；刘志祥等[26]在前人基础上，通过提取甲骨文的字形特征进行分析，并将 6 位数字码技术与上述特征融合，开发了甲骨输入法。此输入法的检字方式与四角号码检字方式类似，可快速准确的输入输出甲骨文字符；刘永革等[17]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。

西夏文输入法的建立的相关研究有：西夏学研究专家李范文首先对比了现代汉字与西夏文字的笔形分类特征[27]，发现两者存在共同性；其次利用现代汉字中可以数字 0-9 表示文字笔形的方法，对西夏文字的笔形进行归类处理；最后，根据西夏文字左上角、右上角、左下角、右下角的四角笔形构件的形态对每个字符进行编码。此方法使得西夏字检索更加方便，为西夏文字输入法发展开辟了捷径；柳长青等[28]在西夏文四角号码输入法方面进行了深入探讨。首先基于键盘钩子技术(Hook)将西夏文字与计算机链接，再采用字体链接技术将西夏文字库与系统 True Type 字库挂接，实现在 Word 中西夏文字与其他宋体、楷体等字库的切换编辑；其次，用微软公司的通用输入法生成器(IME Generator)软件生成了“夏汉通”西夏文字输入法。该工作在一定程度上实现了西夏文本的计算机输入，并提供了一种模糊输入法可以复原部分残缺文字。

藏文输入法的建立的相关研究有：高玉军等[29]通过“HTML+Javscript+码表”实现了藏梵文嵌入式字库的在线藏文输入法；仁青诺布等[30]在前人制作的输入法基础上，进行相应改进，实现了内嵌法和外挂法两种在线输入方式；目前使用率较高的两种藏文输入法是班智达、喜马拉雅。上述所有语料库及字库对集外字的信息处理均集

中在解决集外字录入、显示及检索方面，尚未关注集外字在文本信息处理中的应用；且其字库中的集外字不能直接应用于出土文献语料库中。随着传世文献文本在自然语言处理技术中的应用愈来愈广，出土文献文本的计算机应用尚显不足。本文通过对出土文献中的所有集外字进行整理，使用造字法建立以里耶秦简为例的出土文献集外字字库，并结合四角号码建立集外字输入法，以期在实现集外字录入、显示及检索功能的基础上，使集外字可应用于文本信息处理，尝试为出土文献语料库的集外字信息处理提供一种普适的程序化方法。

1.3 研究内容与方法

研究内容包括：通过阅读、整理、汇总目前已被释读并出版的全部出土文献资料，较为全面地对出土文献的出土地点、数量、主要内容进行整理，并对其所包含的所有集外字进行地毯式、系统化的梳理。首先，通过分析集外字数据，对每部简牍、帛书的集外字占比进行统计；其次，在此基础上，为支持出土文献文字的输入、输出、显示及检索功能，主要使用造字法建立以里耶秦简为例的出土文献语料库的子库——集外字字库；第三，为实现数据可直接参与利用计算机对文本的信息处理工作中，与前人研究相结合，利用多多输入法生成器生成“集外字字库输入法”，并将里耶秦简全文本投入分词实验中。具体研究路径如下：

(1)建立集外字字库

随着不断有新的出土文献被挖掘解读，文献中所含集外字数量也在不断增加。将所有出土文献集外字进行整理建成字库，这样可为后来学者研究集外字或使用出土文献语料库提供便利。基于汉典网站和《汉典》《说文解字》《康熙字典》此类工具书，筛选查找《清华大学藏战国楚竹书》《里耶秦简》《张家山汉简》《长沙走马楼三国吴简》等先秦至晋期间的出土简帛文献中的集外字。

建立出土文献集外字字体库。使用 win10 的专用字符编辑程序和 High-Logic 公司的 FontCreator 曲线轮廓造字软件，使用造字法绘制集外字，并选择国际化标准的 Unicode 字符集编码进行保存，建立字符与字符集编码形成一一映射的关系，实现输出功能。随着字符集的不断发展，越来越多的集外字被纳入相应的字符集内，集外字数量也在不断减少，建立出土文献集外字字库可实现对集外字的动态管理，当有新增集外字时将其扩展至集外字字库，当现有字符集更新时，对已被认定为集内字的集外字从字库中移除，保证字库的更新及可扩展性。

(2)建立出土文献集外字输入法

首先，需要建立出土文献集外字检字码表。绝大多数集外字为方块文字，其基本笔画可为单笔和复笔，笔形可分为头、横、垂、点、叉、插、方、角、八、小十种，参考《四角号码新词典第十版》对集外字进行四角号码编码。

其次，建立出土文献集外字输入法。在以上内容的基础上，建立集外字与四角号码相结合的输入法码表,其中包含集外字,对应的四角号码以及部首结构。再基于多多输入法生成出土文献的集外字输入法。该输入法支持字库更新及共享，便于出土文献语料库对集外字的输入输出。

(3)实现集外字应用于文本信息处理

通过研究字符在计算机中的处理原理,将所有集外字以可被计算机处理的字符形式存储至字库中，并且将其从“字”上升到“词”再到“句”的维度参与文本信息处理，将集外字应用于文本分词。将集外字应用于规则的分词方法和统计的分词方法，并使其参与主流分词工具的分词过程中，测试其分词效果，实现集外字参与文本信息处理的应用。通过实验证明本方法处理的集外字可应用与文本信息处理的可行性，为建立功能性出土文献语料库奠定基础。

(4)提出集外字参与文本信息处理的普适化程序化方法

在众多古籍数字化项目中，对集外字的关注点都集中于集外字显示及检索上[13]，并未关注于其在文本信息处理中的应用，本研究针对建立语料库过程中，提出一种集外字可参与文本信息处理的普适化程序化方法,为建立语料库中所遇到的集外字信息处理问题提供借鉴。

研究方法包括文献研究法、定量分析法、实验分析研究法及跨学科研究法等。

第二章　出土文献集外字整理与分类

从中国出土文献语料库文字规范的角度来看，集外字的出现影响语料库文本内容的统一性、完整性，语料库中数字化文本要求文本形式统一为字符形式，便于文本信息处理应用；从文字研究价值角度来看，集外字包含有各种繁难字、异体字、专造字、隶定字，在对先秦至魏晋期间所有出土文献的集外字进行整理后，其整理结果对于研究文字随着朝代的变化过程、文字自身所蕴含的意义、文字在文本中应用的变化等方面都具有可参考性，所以整理所有出土文献的集外字非常重要且必要。

目前，世界上最通用也是含字符量最大的字符集为 Unicode 字符集[31]，又称为“万国码”或“统一码”，且该字符集已对世界各国文字进行了编码及整理工作。Unicode 字符集虽然已经对绝大多数常用的现代汉字及部分古文字进行了整理、编码的工作，但对于出土文献中所含有的先秦及秦汉时期的隶定字并未涉及。

由于集外字数量多少与字符集所含字符量呈负相关关系，字符集所含字符越多，则对应的集外字就越少。而且，大多字库都是依古文字研究项目而整理并建设的，例如专门研究甲骨文的项目对甲骨文字进行专项整理；研究金文铭文的项目对所有铭文字进行专项整理，研究楚简的项目对部分楚简帛文字进行专项整理。但是，对所有出土文献集外字进行专项整理的工作尚未有所涉及，而建立出土文献语料库需要出土文献集外字字库的支持，而字库的建立离不开对集外字的整理。

2.1 出土文献简帛集外字收录范围

本文所收录的出土文献源于上世纪至今出土的全部已被释读的简牍及帛书，其年代范围上至先秦时期，下至魏晋时期。本文从简牍及帛书的名称、数量、释文、及其内容概况四大方面进行了整理，根据整理结果进行分析，所得：

第一，出土的汉简帛种类最多，涉及 34 部文献。按照其地域划分，大致可被分为西北屯戍简及其他汉简帛。西北屯戍简包含了《玉门花海简》《甘谷汉简》《甘肃敦煌酥油土汉简》《敦煌汉简》《居延汉简》《居延新简》《悬泉汉简》《额济纳汉简》《肩水金关汉简》共 9 部。它们的发掘位置集中在今甘肃、内蒙地区，在两汉时期为边塞地区，由于饱受匈奴的入侵，故边塞地区常设“烽燧”进行屯兵等军事作用[8]。

第二，出土的三国时期的简的种类最少，但其出土的地理位置最为接近，在今湖南省长沙市。同样在今长沙市出土的还有东汉的长沙五一广场简牍。有意思的是，五

一广场汉简、三国吴简和嘉禾吏民田家荊简的具体出土位置很近，同样的还有东牌楼东汉简牍和尚德街东汉简牍[8]，这两部简牍小组正在计划建立数据库，因此本文也计划继续整理其集外字。

第三，里耶秦简在秦简牍中的体量最大，也是目前出土的体量最大的秦简牍文献。在中国简牍的发展史上，若以时代因素为标准，秦时期的简牍面世最晚，直至 1975 年底，云梦睡虎地两个墓葬中的简牍才被发掘。睡虎地秦简和岳麓秦简出土于低级官吏的墓葬之中，其校释小组推断，该批简牍均应为墓主人收藏或生前常用物品[8]。但这两部简牍的内容性质较为单一，均为秦国至秦代的法律文书。从内容上来看，里耶秦简涵盖了秦代洞庭郡迁陵县基层社会的各个方面，其体现的是学界一直无法窥探的动态的秦代基层社会，对秦的历史研究提供更好的史料。本小组基于此，首先以里耶秦简为例，建立其熟语料库及其集外字子库。

第四，有四部简未找到具体的发掘地址，是因为有部分先秦简、秦汉简是以捐赠形式而面世的，分别为战国时期的清华简，上海博物馆藏战国简、香港大学藏秦汉简牍及北大汉简。

第五，从简牍帛书的主要内容进行分类，可被分为文书类和典籍类[8]，根据整理，文书类文献共有 29 部简牍，典籍类及遣册类文献共有 20 部简牍，还有 3 部简牍目前无法获取相关内容的资料。综合来看，出土的简牍中，文书简占比居多。

第六，按今地理位置对出土文献进行大致比较，出土地点分布于甘肃、青海、内蒙古、新疆、四川、北京、河北、河南、山东、安徽、江苏、江西、湖北、湖南、广东、广西多省、市、自治区等[32]。出土地域较为集中的一为华中地区，二为西北地区。其主要原因在于这两地的土层和地表气候适合简牍的长期保存。

2.2 出土文献集外字分类

出土文献中集外字数量庞大，若要将出土文献中所有集外字展现至本章节有些许困难。故本文选取出土文献中部分较具代表性的集外字，进行整理并展示。出土文献集外字大致可分为三大类，分别为已释集外字、未释集外字、存疑集外字。其中已释的集外字根据文字所表示的意义，又可分为表示名物、行为动作两类。例如：

（一）已释集外字

1.表示名物的集外字

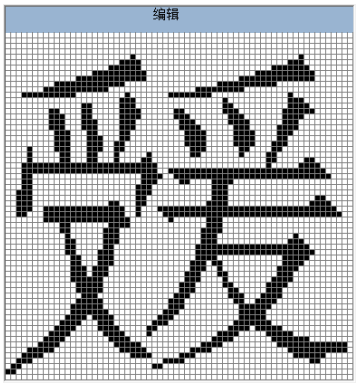

图 1 里耶秦简集外字“𤼵”

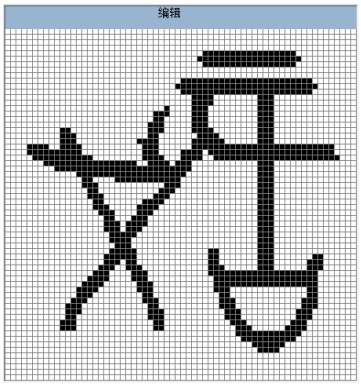

图 2 里耶秦简集外字“𤽙”

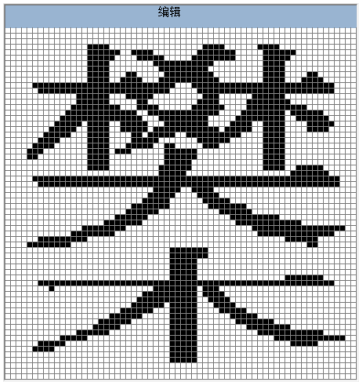

图 3 里耶秦简集外字“𤽾”

表示人名。“𤼵”、“𤽙”、“𤽾”在里耶秦简校释文中被释为“緩”、“棓”、“攀”，音从“緩”、“棓”、“攀”。因为里耶秦简内容为程式化的官府文书，故其行文格式有严格要求，简牍中提到的职官名如守、佐、司空、鄉之后一般跟随人名[35,36]。例如简 8-39 中“佐見已死。廿九年鄉歜、佐緩已死”，其中“佐”、“鄉”为职官名，职官名之后的“見”、“歜”、“緩”都为人名。

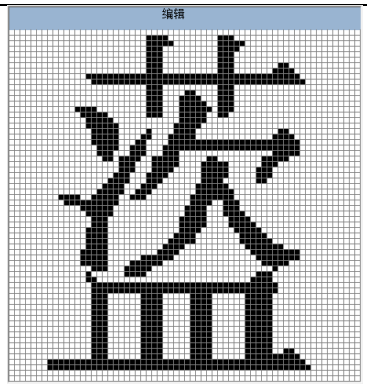

图 4　里耶秦简集外字“[illegible]”

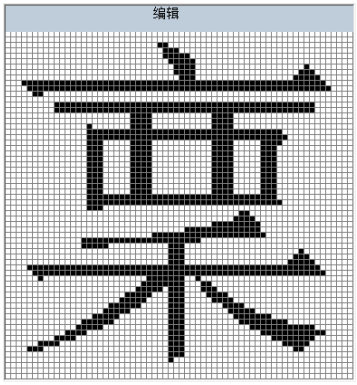

图 5　敦煌汉简集外字“[illegible]”

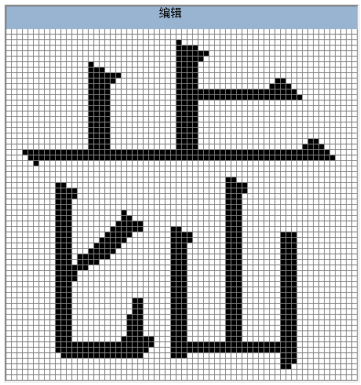

图 6　里耶秦简集外字“[illegible]”

表示植物、作物名。“[illegible]”在里耶秦简校释文中被释为“葵”，音从“葵”。《本草纲目·草五·葵》中有提到“葵”是一种菊科草本植物，有绵葵、秋葵、向日葵等[35,36]，《说文解字》中有“葵，菜也”。“[illegible]”音从“稟”，在校释文中被释为“稟”，在敦煌汉简 941 中以“出穀九十七石二斗，給稟莫府馬食”出现。“稟”在《说文解字》中有“賜穀也”[37]。“[illegible]”被释为“[illegible]”，此字在秦汉简帛中多次出现，也通“腦”，是穀物名。在《说文解字》中有“穀，楮也”[37]。

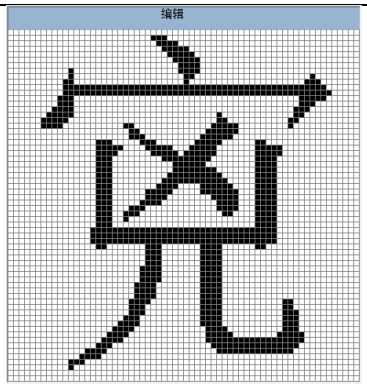

图 7　里耶秦简集外字“𡧓”

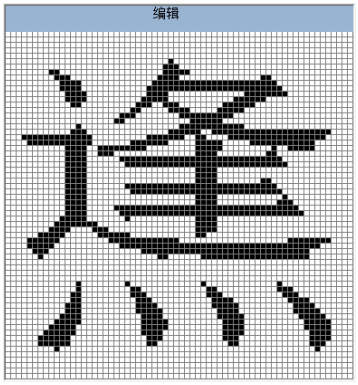

图 8　敦煌汉简集外字“”

表与军事相关名词。“𡧓”音从“兕”，在里耶秦简 8-458 中以“甲𡧓廿一”形式出现，该简为武器装备记录，文书格式为“武器/装备+数量”，“甲𡧓”为一种防具装备名称。“”音从“燙”，表示边防报警的烟火，在校释文中被释为“燙”。

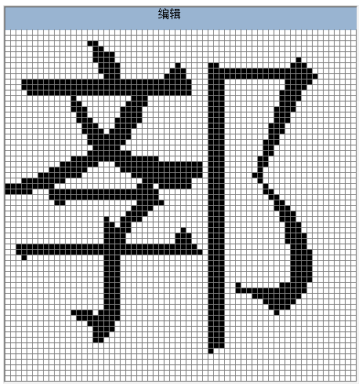

图 9　里耶秦简集外字“𨒗”

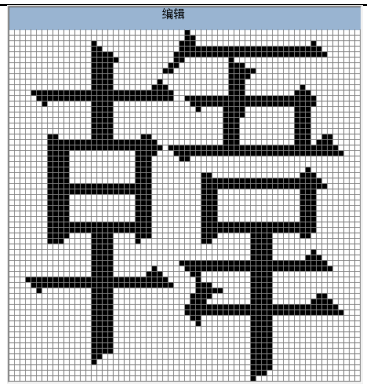

图 10　里耶秦简集外字“韓”

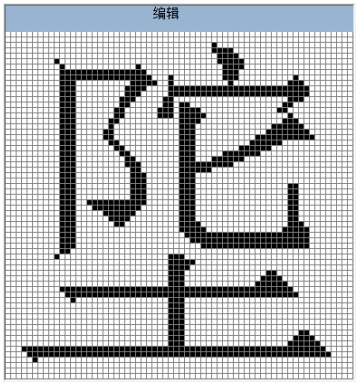

图 11　郭店楚简/清华大学藏秦简集外字“墜”

表示地名。“郚”、“韓”在里耶秦简校释文中被释为“□”、“韓”，“郚”音节不明，“韓”音从“韩”。简 8-761 中的“同郚”、简 8-894 中的“韓審里”均表示里名[35]。《清华大学藏秦简》的简“8-治邦之道-20”和《郭店楚简》的简“太一生水-1”中均出现集外字“墜”，音从“地”，其被释为“地”，表示大地。《说文解字》有“元气初分，輕清陽爲天，重濁陰爲地”[37]。

2.表示行为动作的集外字

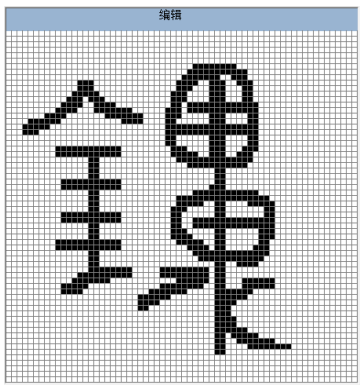

图 12　里耶秦简集外字“[illegible]”

表示铸造锤炼。“”音从“铸”，在里耶秦简校释文中被释为“鑄”[35]。在简8-454中以“鑄锻”出现，《拾遗记·后汉》中有“時人谓：郭氏之室，不语而雷。言其鑄锻之聲盛也”[35]。

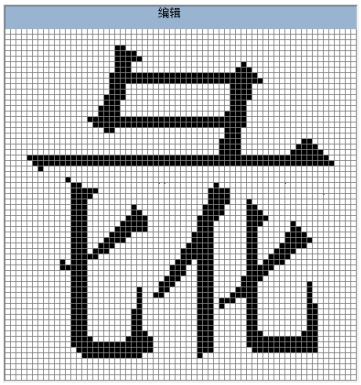

图 13　里耶秦简集外字“”

表示分割。“”音从“剥”，在里耶秦简校释文中被释为“剝”[35]。在简8-490中以“徒隸牧畜死負、剝賣課”出现，“畜死負、剝賣”是指牲畜死后，赔偿或分割售卖。

（二）未释集外字

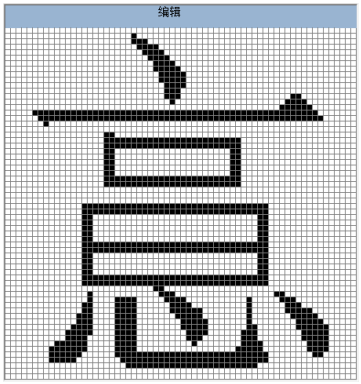

图 14　里耶秦简集外字“”

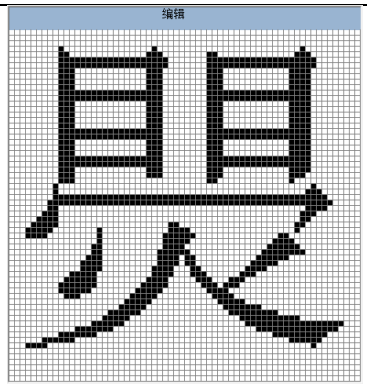

图 15　里耶秦简集外字“㬎”

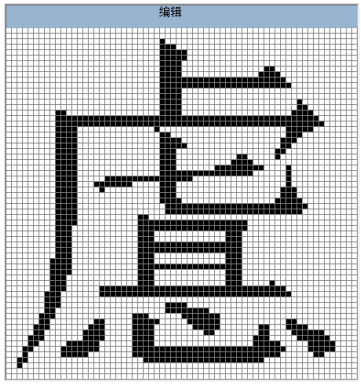

图 16　郭店楚简集外字“虙”

此类集外字在校释文中均未给出校释，无准确音节及音调。

（三）存疑集外字

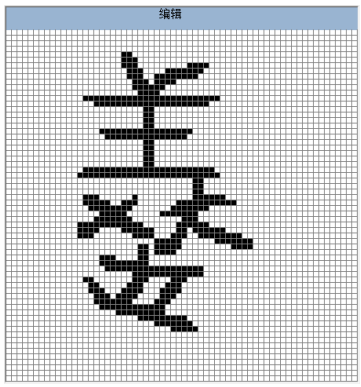

图 17　里耶秦简集外字“”

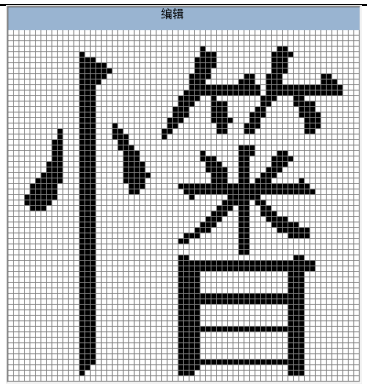

图 18　里耶秦简集外字“𤔔”

“”在校释文中疑为“羡”字的变体，音疑从“羡”，在简 8-1087 中以“羡小畜”出现，应是针对小畜的某种动作[35]。“𤔔”在校释文中疑为“篚”字，音疑从“造”，待考[35]。

2.3 出土文献集外字字频分析

字频作为文字系统的第四大特征，对其进行分析可将文字系统的性质、特点呈现出来。因此，字频的分析研究也成为了文字学中的重要组成部分。但对于出土文献的字频分析研究随着出土文献的考古发掘工作有了进一步的探索。本文在小组工作的基础上对以上简帛中出现的集外字进行整理研究。整理工作首先将整理出的集外字按照其时期分为先秦、秦、汉籍三国至晋，按照时间顺序，将每部简牍或帛书的全文文字个数、集外字个数以及集外字在全文字数中的占比进行统计，对整理结果进行分析可得：

第一，从总字数来看，出土文献总字数呈现出字数随着时间的后移不断增大。三国时期文献的体量最大，字数越多，说明简牍的保存程度越好，且也与古人如何处理这些简牍的方式和简牍被埋藏的地理位置都有极大关系。三国吴简是在 50 余口古井窖群中存放的，因此其出土后，成为了体量最大的简帛文献[8]。

第二，部分先秦时期的简帛文献的总体量也高于部分秦简牍，主要原因是先秦时期的文献，例如上海博物馆藏战国楚简及清华大学藏竹简均是捐赠的方式，并不是出土的方式面世。且其主要内容涉及的是典籍文献。这类简帛文献与以里耶秦简及睡虎地秦简为代表的秦简牍和以居延汉简及悬泉汉简为代表的汉简牍的实用官府行政文书不同。典籍类简牍文献是古人专门组织一批写手将成文的典籍誊写或抄写至简牍上的，使得这些简牍成卷成册，有其唯一且特殊的编连顺序。但行政文书不同，它的作用主要是记录秦汉基层社会中的政府行为的，如兵器出入簿、粟米出入簿、律令、奏

谳书等。这些文书的主要用途是传递信息，且由于书写材料是简牍，在当时较为珍贵，因此或许比“成卷成册”的先秦时期简帛文献的字数较少一些。

第三，集外字占比最大的时代是先秦时期的文献。究其原因，首先是因为先秦古文字由于国别地域差异的问题，导致了文字的使用与文字的编写形式未得到统一。因此，此时的文字书写方式百家百态。其次，先秦古文字与现代汉字的书写风格截然不同[33]，其文字通用性程度较现代汉字而言更低，所以集外字占比较大。第三，细化集外字字频，先秦时期集外字的不重复率也是三个时期中的最低值。每部文献的每条简文中，几乎均存在集外字。

第四，集外字占比较小的时代为三国时期的文献。主要原因是：其一，三国时期的简牍出土量大，且出土时简牍的保存程度较其他时期的偏好，因此，其所含文字的体量最大。其二，虽然三国时期的出土文献集外字占比较小，但在实际的统计数量过程中，其集外字的出现频次远高于先秦时期及秦汉时期的简牍、帛书。但由于其字数更多，体量更大，所以集外字的占比偏小。以《长沙走马楼三国吴简》为例，其中集外字字频为 3814 频次，而这 3814 频次的集外字中，“垈”这一集外字就出现了 2100 余次。其三，三国时期的文献所含的集外字的整体个数虽然远高于其他两个时期，但细致分析，其集外字的重复率最高。例如，若将“垈”这一集外字频次计为 1，那么《长沙走马楼三国吴简》中集外字字数在全文占比约在 0.175%左右。

第五，整体上来看，除了先秦简帛，其余时代简帛集外字的占比大约在 0.1%-0.8%范围内，虽然集外字字数在总文本中的占比较小，但是文本中集外字占全文不重复字数比重较高，以里耶秦简（壹）、（贰）为例，其中集外字个数与总文的文字个数的占比约 12.8%(集外字个数 231 个，全文不重复字数约 1800 个)，虽然集外字字频相对于出土文献文本而言较低，但是其占全文不重复字数比重较高，所以有必要对文本集外字进行整理并建立字库。

第六，从发展趋势来看，集外字在各出土文献中的占比情况体现出汉字的流变的大致情况为：从先秦至晋期间集外字占比逐渐下降从侧面可以反映出在此期间文字发展系统逐渐走向成熟，文字传播程度增加，这一点不仅可以从集外字占比逐渐减小这一现象得出，从刘志基[34]在对出土文献字频分析中也可看出。

2.3.1 曾侯乙墓简所见集外字

曾侯乙墓简的内容涉及车马及攻击武器、防御武器的情况。下表从曾侯乙墓简中

的集外字、对应简号以及字频三方面，对字频率较高的前五个集外字进行整理，如下表所示：

表 1 曾侯乙墓简集外字字频 Top5

集外字	简号	字频
𣪊	入車-2,入車-4,入車-6,入車-8,入車-10,入車-14,入車-17,入車-18,入車-19,入車-20,入車-23,入車-25,入車-26,入車-29,入車-32,入車-36,入車-38,二甲胄-122	18
𩰫	入車-2,入車-6,入車-8,入車-14,入車-17,入車-19,入車-20,入車-23,入車-25,入車-26,入車-29,入車-32,入車-36,入車-38,入車-39,入車-46,入車-49,二甲胄-122、	18
𩈦	入車-2,入車-15,入車-17,入車-19,入車-23,入車-25,入車-26,入車-29,入車-32,入車-36,入車-38,入車-39,入車-46,入車-49,入車-6,入車-8,	16
𩠐	入車-5,入車-8,入車-9,入車-10,入車-13,入車-14,入車-19,入車-21,入車-29,入車-32,入車-36,入車-39,入車-40,入車-45,入車-78	15
𧟌	入車-1,入車-8,入車-13,入車-16,入車-18,入車-23,入車-24,入車-25,入車-26,入車-29,入車-32,入車-36,入車-39,入車-78,	14

由上表可得，“𣪊”和“𩰫”在文本中的字频最高，出现了 18 次，其中有 15 次都是以“屯𣪊𩰫”的形式出现。有的简用貧或紛代替，有学者认为该字是“放”的异体字；罗小華认为曾侯乙墓中有“衡”和“𣪊衡”，疑读为“黂”，表示粉色花纹[61]。

“𩰫”在校释文中，整理者根据简文记录格式推断其为附属物或附饰[61]。

“𩈦”在释文中多以“二𩈦”形式出现，校释疑释读为“鞔”，表示马缰绳。《说文解字》中有“车具也”[37]。

“𩠐”在释文中有“𧟌𩠐之聶”、“貍𩠐之聶”和“貂𩠐之聶”。校释文未给出释意。

“𧟌”在释文中以“𧟌韔”、“𧟌𩠐”形式出现。“𧟌韔”表示用虎皮做的韔[61]。

2.3.2 里耶秦简所见集外字

里耶秦简的内容涉及官府文书。下表从里耶秦简中的集外字、对应简号以及字频三方面，对字频率较高的前五个集外字进行整理，如下表所示：

表 2 里耶秦简集外字字频 Top5

集外字	简号	字频
𦥑	8-348,8-406,8-690,8-839,8-865,8-902,8-972,8-1517,8-1771,8-1809,8-1973,8-2269,9-202,9-546,9-1157,9-1203,9-1281,,9-1547,9-1641,9-1913,9-2385,9-3023	22
犴	8-474,8-565,8-575,8-763,8-765,8-800,,8-890,8-1239,8-1286,8-1360,8-1580,8-2249,9-813,9-948,9-99,9-1454,9-2334,	17
雔	8-487,8-1351,9-1099,9-1139,9-1271,9-1451,9-1766,9-2067,9-2203,9-3295	10

顥	8-36,8-477,8-1108,8-1476,8-2087,9-160,9-563,9-2561,	8
窊	8-458,8-760,9-177,9-1164,9-1547,9-2041	6

由上表可看出“臂”的字频最高，在简 8-348、8-406 与 8-865 中以“臂手”出现；在简 8-690 中以“佐臂以来”出现，其均表示为人名，其职官为佐。

“犴”在简 8-474 中以“令史犴視平”出现；在简 8-565 中以“佐犴四甲”出现，均表示为人名，对应的职官有令史和佐。

“雔”在 8-487、9-1451 中以“雔手”出现，在简 9-1099 中以“佐雔以来”，在 9-2067 中以“守雔敢言之”出现，其均表示人名，对应的职官有佐和守。

“顥”被释为“願”，在 8-477 中以“式願寫之”出现。但是校释小组在文中并未给出释意。在《说文解字》中有“願，大頭也”，指主要的部分[37]。

“窊”在简 8-458、9-1164、9-2041 中均以“甲窊”出现，表示防具装备。

从整体上来看，字频最高的前三个字均为人名，第四个高频字在校释文中未给出释意，第五个高频字表示防具装备，其他集外字字频均小于 5 次，且大多字频为 1。绝大多数集外字字频为 1，其原因可能是因为在文字发展过程中，不常用或通用性不高的汉字都会被其常用或通用性较高的字代替。类似情况在其他简牍中也有体现。例如在《上海博物馆藏战国楚竹书》中的集外字“  ”，“”，“”，“  ”，这四个集外字均被释为“亂”。而“亂”的异体字有“乱”、“亂”、“釠”、“乿”、“乱”、“乳”、“率”、“率”、“率”、“孌”、“孿”、“敳”、“敳”、“敳”、“爰”、“孿”、“爨”、“爨”、“爨”、“爨”、“爨”，这些通用性较高的字被收录为集内字，而“”，“”，“”，“”这些字因通用性较低，在文字发展过程中被归为集外字行列。

人名类集外字字频最高，原因可能是因为里耶秦简是官府文书，文书记录要求记录者或被记录者要有署名。但实际上这些字单独成词，且为人名，在文字发展过程中因通用性不高而被归为集外字。

2.3.3 银雀山汉简所见集外字

银雀山汉简内容包括军事类的《孙膑兵法》、《晏子》，律令类的《守法守令》等。下表从银雀山汉简中的集外字、对应简号以及字频三方面，对字频率较高的前五个集外字进行整理，如下表所示：

表 3 银雀山汉简集外字字频 Top5

集外字	简号	字频
顯	1290,1294,,1299,1303,1319,1346,1350,1351,1370,1400	10
牧	1305,1306,1307,1310,1312,1313	6
篡	1747,1781,1795,1871	4
仳	1113,1116,1117	3
襲	1534,1546,1547	3

“顯”被释为“願”，在简 1290 中以“願聞有國之大失”出现，但是校释小组在文中并未给出释意。此字的被釋字“願”在里耶秦简中也有出现。但两个集外字字形不同。

“牧”，校释小组校释：该字不见于字书，字当从牛勺声[62]。牟成牧，即见于古书的務成昭。牟与務，勺与昭，古音相近可通。

“篡”被释为“葬”，在简 1747 及 1748 中以“可葬貍（埋），分異”出现，校释小组并未给出具体释义。

“仳”被释为“背”[62]，在简 1113 中以“……背之勝”出现，

“襲”在简 1534 中以“玄襲之陳者”出现,校释文未给出释意。

2.4 小结

本章首先从简牍及帛书的名称、数量、释文、及其内容概况四大方面对集外字的收录材料范围进行整理，并从时间、空间、数量、内容的角度对其进行了简要分析。

其次，对材料中每部简牍中的集外字字频及文本内容占比进行了整理，并根据整理结果详细论述了文字字频及占比在不同朝代和不同简牍中所呈现的整体特点、发展趋势及产生原因。

最后，以曾侯乙墓简、里耶秦简、银雀山汉简中的集外字为例，对其进行了具体地整理并详细地分类。并分为已释字、未释字、存疑字三大类，其中已释字部分根据文字表意，又细分为表名物和表行为动作两类。同时，对集外字字频最高的前五个集外字进行了整理分析，又以《上海博物馆藏战国楚竹书》中的集外字为例来展现集外字的重要研究价值。

第三章　出土文献集外字字库的建立

随着古籍数字化和古籍语料库的不断发展，语料库信息处理的应用范围越来越广泛。但在建立“出土文献语料库”过程中，无法使用 OCR 识别技术对古籍进行数字化后校对录入。主要原因是集外字本就不存在于现有的字符集中，而 OCR 识别是建立在现有字符集基础上的一种将图像类字符转换成集内字技术。在 OCR 识别集外字时会根据算法将其识别为与集外字字形轮廓相似的集内字，这会导致集外字识别录入错误[38,39]。

为了保存集外字的原本字形信息及其语义信息，和快速并完整地建立语料库，前人会将集外字以图片形式替代并录入语料库。此种方法虽然在建库前期有快速准确的特点，但是集外字无法直接以字符形式录入语料库，而计算机的文本信息处理过程是对字符所对应的机内码进行识别并运算。因此，图片形式的集外字在语料库中只可以显示、检索，而不可以参与文本分词此类过程。本文为了实现“出土文献语料库”的完整性以及文本信息处理功能实现，故需要建立出土文献集外字字库。

建立出土文献集外字字库，首先是为了对出土文献中的所有集外字进行整合。因为集外字的字形及其语意和用法等就具有非常高的研究价值，故集外字字库对出土文献研究而言其可研究性价值潜力巨大；其次，是因为集外字在语料库中是一个动态增长的过程，随着不断有新发掘的出土文献面世，所对应的集外字数量会随着出土文献数量的增加而增多，所以对于“出土文献语料库”而言，有必要建立集外字字库对集外字进行动态扩充及管理；最后，是为了在使集外字在完成显示及检索的基础上，同时可参与到文本信息处理过程中，所以建立出土文献集外字字库非常有价值且必要。

3.1 现有集外字处理手段

目前古籍数字化已初见成效，对集外字也有不同的处理方法出现。现有的对集外字的处理方法主要有三种，分别是替换法、描述法及造字法[40,41]。以下将对三种方法分开说明，整理如下：

3.1.1 替换法

替换法是用计算机可直接输入的形式去替代古籍中的集外字，常见的替换有特定符号、图片、集内字这三种形式。第一种特定符号替代形式在古籍数字化项目较为常用，例如目前通用的有用“□”表示未识别字；或者某一集外字在文中出现较为频繁，

用特定符号去替代此集外字，这是最简单的集外字处理形式。不过此种方法存在如下几点不足：1.特定符号有限，所能表示的集外字数量少，且当集外字数量过多时符号数量就增加，增加用户学习成本;2.替换的符号类似于集外字的映射，但是不包含字形信息，对用户理解学习集外字没有帮助；3.符号在参与文本信息处理时会影响信息处理结果准确率。第二种是图片替代形式，即对集外字截图去替代文本中的集外字，此法可以完整保留集外字特征，利于其录入和显示，此方法存在不足有：1.图片形式的文字不便检索，在语料库中功能性会较弱。2.无法参与文本信息处理，在图片与文字同时存在的语料库中，进行分词或文本挖掘时会自动忽略图片字而只识别集内字，影响分词准确率。3.图片较多时占用数据内存较大。第三种是集内字替代形式，相对于前两种处理形式，此方法在现有的古籍语料库中使用更广泛。优势在于可直接录入语料库且方便检索，同时可参与文本信息处理，不足之处在于：1.集外字认定为集内字困难，需要有统一标准；2.无法保留集外字特征；3.文本中存在的集内字与所替换的集内字同时存在时，影响文本信息处理结果准确率。在大型古籍数字化工程中，集内字的替换形式对工作人员提出更高的标准要求[38]，根据肖禹老师[40]在国家图书馆数字方志工程中的统计，笔画略有差异的异体字中有 58%是由书写方式不同导致的；部首结构或隶定导致的异体字占比约为 2%，将中古汉语中的集外字录入进计算机时，其录字产生失误的平均概率为 0.22%。因此在数字化古籍工作中，集外字的写入会极大地影响数字化后的文本的准确性、阅读性等问题，且数字化文本的项目难度越大，其影响越深。

3.1.2 描述法

描述法分为两种，第一种是部首拼接法，第二种是动态组字法。第一种是拆解集外字的结构，通过在 word 文档中调整部首结构的上下左右位置，对集外字进行展示，此方法可粗略的展示结构简单的集外字，是一种非正式的描述方案，只是对于集外字的显示有帮助，对于结构复杂的集外字位置调整难度大且费时。第二种是动态组字法，也称动态描述法，该方法是通过表意文字描述符(IDC)及表意文字描述序列(IDS)的组合[42]，将汉字部首组合为集外字，相对于第一种描述法其描述更准确且扩展性更强，它的优点是可以用一个标准化描述符序列来解释汉字的结构，相较于自然语言描述更为规范。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，用 IDS 描述法就是

IDS:U＋2FF1，U＋5B80，U＋2FF1，U＋2FF0，U＋738B，U＋5C14，U＋2FFA，U＋8FB6，U＋2FF0，U＋96B9，U＋2FF0，U＋8C9D，U＋62D，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。

3.1.3 造字法

造字法是对集外字进行描绘，描绘成汉字字符形式并保存在字符集的私用区编码的码位中，集外字与私用区编码码位形成一一映射关系，此方法相当于将集外字认定为集内字。通过造字法可最大程度保留集外字的特征，并且满足与集内字显示输出别无二致。以 Unicode 字符集编码为例，在 Unicode 字符集编码的建设之初，首先在其基本面预留了私用区码位，共 6400 个码位；其次，还将本就用来增补私用的 A 区编码和 B 区编码进行扩充，共 131068 个码位。这样一来，已含有的 6400 个码位再加上后续增加的 131068 个码位[43,44]，这意味着用户可创造约 13.8 万个集外字。因为码位有限，大多古籍数字化项目都会缩减所创造的字数，以国家图书馆的数字方志项目为例，其将所描绘字数限制在 4866 字[40]。此方法有易录入、展示文字信息全面的优点，缺点有：1.古籍数字化项目众多，随意造字导致一字多码或一码多字；2.码位数量有限，最多只有 13.8 万个码位，若集外字数量超过 13.8 万会造成码位不足；3.混合使用多个数字化项目易造成字库混乱。

通过比较上述三种方法，造字法的原理其设置最符合出土文献集外字的特征，故本文首先采用“造字法”对里耶秦简文本中的集外字进行造字处理。

根据在里耶秦简文本中整理出的集外字，其字形大致分为两种，第一种是部首结构规整的方块字，例如“  ”、“”此类字形；第二种是部首结构不规整的临摹字，例如“  ”、“  ”等。在里耶秦简两卷的集外字整理过程中，们参考文渊阁四库全书电子版的做法，尽量减少造字。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。在创建集外字字体库的过程中，依照集外字字形不同主要使用了两种造字工具，选择其自适应的方法进行造字。

方法一为微软自带的专用字符编辑程序，对结构规整的集外字进行造字；举例如下：

例 1：集外字“宽”的造字步骤：首先，对所选集外字进行部首构建拆分例如集外字“宽”可拆分为“宀”和“兑”；其次，在“窗口”工具栏选择参照，在参照文字中对所需部首构建“宀”和“兑”进行截取平移，调整构件的位置大小，如图 19 所示：

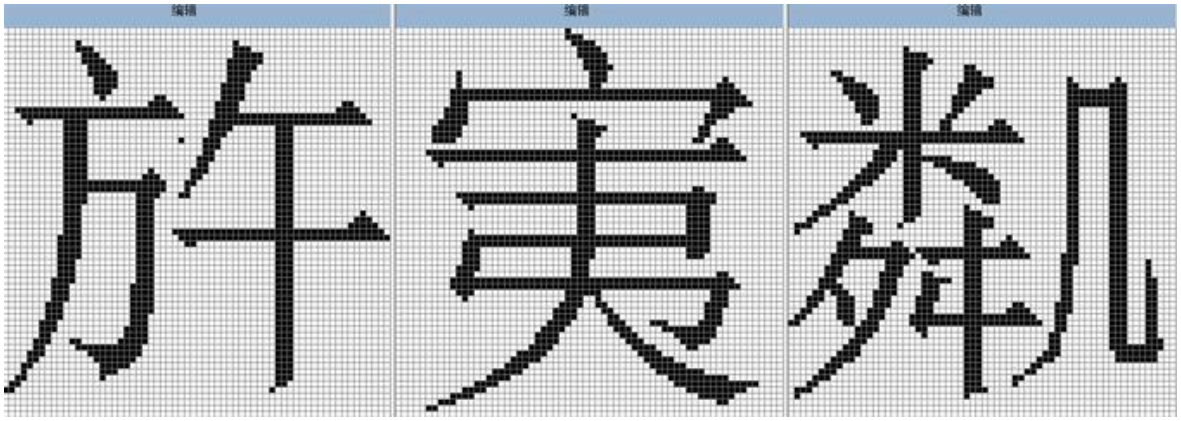

图 19　Win10 专用字符编辑程序所造集外字

方法二为 FontCreator 字形录入工具，对结构不规整的集外字进行造字，举例如下：

例 2：集外字“  ”的造字步骤：首先，将图像拖至 Font Creator 中进行平滑处理；其次调整理文字的轮廓点、线、角度和位置，如下图 20 所示；

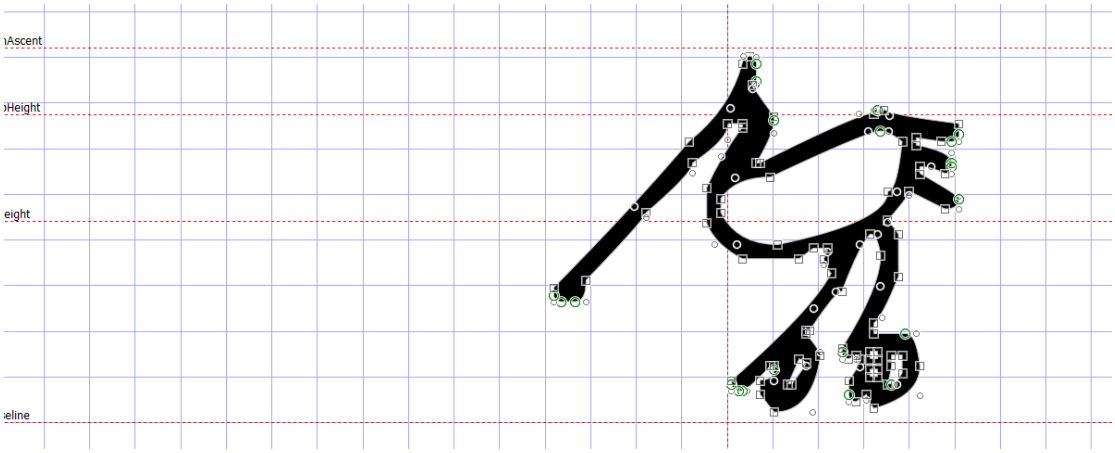

图 20　Font Creator 工具所造集外字

3.2 计算机处理字符原理与字符编码

3.2.1 计算机处理字符原理

为满足字库所具备的以上三点功能，在建立字库时，需明确以下两个步骤，其一是汉字字符在计算机中的存储方式；其二是计算机对汉字的处理方式。且我国为统一汉字在计算机中存储及处理制定了《信息交换用汉字编码字符集—基本集》[45]，在该

标准中，定义了存储及转译所需的区位码、国标码、机内码，及其相互之间的转换关系。

第一，区位码是将所有汉字储存在一个方阵区位中。每一行为所对应的数值为“区”，每一列所对应的数值为“位”，方阵区位内的每个汉字都以区位码来表示。

第二，国标码是所有汉字都必须遵循的编码标准，包括建立字库、输入及输出汉字的区位码都必须以此为标准。国标码规定了一个汉字可表示为两个字节，即在区位码的基础上将“区码”和“位码”各加上 32(20H)即可得到国标码。由于国标码会与通用 ASCII 产生二义性，因此不能直接应用于计算机内部字符转译处理。而能够解决此问题的方式是建立机内码。

第三，在计算机诞生之初，其处理的数据对象是数值数据以及字符格式数据，因此 ASCII 码对所有的数值数据及字符数据都进行了定义，且这两种数据都是以 ASCII 码的形式存在于计算机中。由于计算机仅对 ASCII 码定义范围之内的数据进行处理，而为了使定义范围外的数据可参与计算机处理过程中，机内码随之产生。对于汉字字符而言，机内码是为了解决国标码和 ASCII 码产生冲突，而存在的二义性问题。

以国标码“3123H”和 ASCII 码“3123H”为例说明二者在计算机处理中的二义性问题。国标码“3123H”表示的是汉字“保”，而 ASCII 码“3123H”表示的是“1#”，两者码位相同但其所示含义截然不同。而二者均为计算机处理字符时所必需的码位，因此将国标码的字节高位数值改为 1，与 ASCII 码进行区分。例如“保”字，其国标码的高字节是“00110001B”，低字节是“00100011B”，转换为机内码后高字节是“10110001B”，低字节是“10100011B”，对应的机内码为“B1A3H”。

区位码、国标码、机内码其互相转换关系入下：

区位码 + 32(20H) = 国标码
国标码 + 128(80H) = 机内码
区位码 + 160(A0H) = 机内码

计算机中对字符的处理方式可分为存储、通信及显示，处理步骤流程各有不同，计算机内部处理字符原理流程如下图所示：

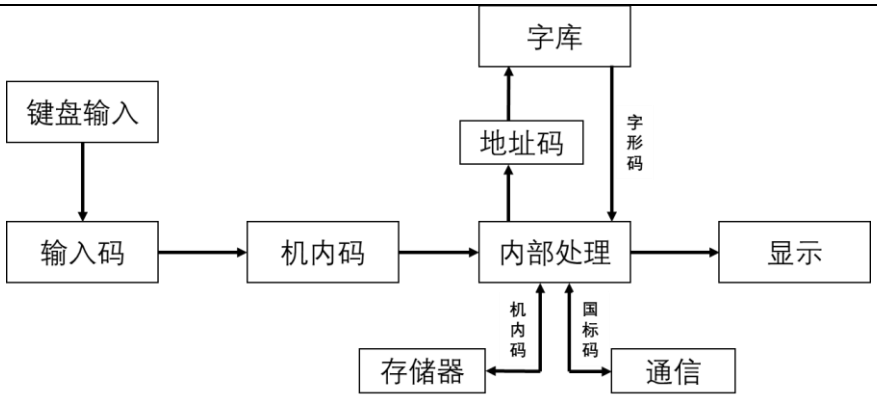

图 21　计算机内部处理字符原理流程

对字符的存储上共分为四步，第一步获取字符字形，根据字符集编码规则的不同对其进行编码，这个编码也就是查找该字符所需要的地址码；第二步是将字符集中的字符按照对应的字符编码规则汇集在一起形成字库；第三步是将字库中每个字符所对应的字符集编码转换为机内码储存至存储器。

对字符的显示上共分为四步，第一步是根据输入管理模块获取输入码；第二步是将输入码转换为机内码并查找机内码所对应的地址码；第三步是调用字符地址码所对应字库中的字形信息，一般为矢量字形；第四步是将字形信息储存至显示缓存中，经过显示控制器处理进行显示。

对字符的通信应用上共分为三步，第一步是根据输入管理模块获取输入码；第二步是将输入码转换为机内码并查找机内码所对应的地址码；第三步是根据国家标准码的规则将字符集编码转换为国标码进行显示及通信。

综上，对集外字进行造字形成字形码，基于合适的字符集编码规则将所有的字形码整合形成字库，通过计算机内部处理将字库中每个字符编码转换为机内码可进行存储，则该字库中的所有集外字字符可进行显示和通信，在字符处理层面说明该方法可使集外字能够被计算机识别和应用于信息处理，所以选择合适的字符集对建立字库至关重要。

3.2.2 字符集编码

每个汉字字符集都由区位码、国标码、机内码组成，区位码表示汉字在字符集中的位置，国标码表示汉字字符的编码，机内码表示汉字字符在计算机中的存储位置[43]。常见的汉字字符集有 GB2312、GBK、BIG5、GB18030 及 Unicode[41]，具体如下：

第一个是 GB2312 编码字符集。1981 年 5 月 1 日国家技术监督局发布的简体中文

汉字编码国家标准，GB2312 编码字符集采用双字节编码，并以区位码为基础。GB2312 编码字符集收录的汉字分成两个层级：第一层存放常用汉字约 3755 个，该区汉字依照拼音声母顺序进行排序，据此可得到某个拼音在该层汉字区位中的范围，大多数输入法的检字程序就是根据此规律编写而来；第二层存放不常用汉字计约 3008 个，字符集中不仅收录了常用的简体汉字字符同时还收录有日文、韩文等此类外文字符。

第二个是 GBK 编码字符集。由于 GB2312 编码字符集中不包含 Unicode 基本平面中的大部分现代汉字字符，因此，GBK 编码字符集在收录了上述两者的基础上，与此用时，还收录了日文、韩文等外文字符及汉字部首、标点等字符。GBK 编码字符集的编码范围分为三个区，其一是汉字区，其二是图形符号区，其三是用户自定义区。其中用户自定义区可以提供 1894 个码位可供用户自由编辑，但是其用户自定义区只开放部分码位，不完全对用户开放。

第三个是 Big5 编码字符集。它采用的是“高字节+低字节”的编码形式，其字符集仅收录繁体汉字字符，且所收录字符数量小于且被包含于 GBK 字符集。两种字符集编码在表示相同的汉字时，所对应的编码不同。Big5 字符集编码中还设定了 0x8140-0xA0FE 的码位，共 8126 个，是用户自定义区，用户可在本区域的码位中保存自定义字符。

第四个是 GB18030 编码字符集，是我国定制的多字节标准字符集，不同于 GB2312 编码字符集的双字节。它类似于 Big5 字符集与 GBK 字符集的扩充集，GB18030-2005 字符集中共收录汉字 70244 个，同时也收录了世界大多数国家的文字字符。

第五个是 Unicode 编码字符集。它是国际标准字符集，囊括了世界各国的文字字符，可实现多国家多种语言的文本信息转译[31]，该字符集为世界各国字符分配了唯一的编码码位用以字符存储。Unicode 字符集持续在更新，最近一次的更新版本中的字符集支持超过 600 种语言的字符,克服了简繁体、西夏文与现代汉字、少数民族文字与现代汉字不能同屏的困难，同时也解决了多国语言无法同屏交流的问题，实现了世界语言之间的信息共享。

字符集的选择与集外字的数量是反向增长关系，即字符集所收录字符越多，则该字符集所对应的集外字就少[40]；若字符集所收录字符越少，则该字符集所对应的集外字就越多。故本文建立字库选择收录字符最多的 Unicode 字符集，其优点有：第一，所收录文字量大，此字符集所对应的集外字少；第二，Unicode 可满足国际互通，是因为该编码内所含的字符均具备全球统一性及唯一性[43]；第三，可为用户提供约 13.8

万个自定义编码区域，可充分满足出土文献集外字的字库建立需求。

其具体的操作步骤如下：选择所整理集外字的方块字图像，为其分配好对应的 Unicode 编码。首先如下图 22 所示选择 Unicode 编码为 UE3D0；其次如下图 23 所示将所造集外字与此 Unicode 对应：

图 22　Win10 系统字符映射表

图 23　集外字字符映射表

3.2.3 建立出土文献集外字字库及 Open Type 字体库的建立

计算机系统通过计算机字体以显示字符的外表形态，字体库将我们所熟悉的字符和图像汇集起来供计算机存储和文本处理，在字体库中每个字符都有所对应的字符集编码。计算机以字符集编码的形式对字符进行存储，用户通过文本编辑环境选择所需的字体库对文本进行录入、编辑及文本处理，根据字模描述方法所用技术不同，其大体可分为 2 类，分别是点阵字体、矢量字体，其格式特点如下：

(1)点阵字体

顾名思义点阵字体就是在二维像素方阵中用点描绘的字符，此类字符可清晰的显示文字轮廓，具有分辨率小、处理速度快的特点，但是受限于字体分辨率的影响，其放大之后的字形存在锯齿，故而只适合作 Windows 的系统字体显示，不适合作为集外字字库的字体格式。

(2)矢量字体

不同于点阵字体的描绘方式，矢量字体是通过保留字形的曲线信息进行存储的。

计算机通过读取该字形的数学曲线，实现对字形的描绘和渲染，此方法的优点是可进行任意缩放保证字体边缘依然光滑，字符不变形。因此，矢量字体优于点阵字体。

主流的矢量字体格式有 3 种：Type1，TrueType 和 Open Type。其中 OpenType 字体格式兼有 Type1 和 TrueType 的特点，字体美观且精确，并且支持 Unicode 字体。它在屏幕显示及打印中都呈现出其高度的优越性，所以 Open Type 适合作为集外字字库的字体格式，因此本出土文献集外字字体库就采用了矢量字体中的 Open Type 字体。

其具体操作步骤如下：将建立好的文字进行存储并生成扩展名为“.ttf”的 Open Type 文件，如下图 24 所示，展开文件后，界面如图 25 所示：

图 24 .ttf 文件

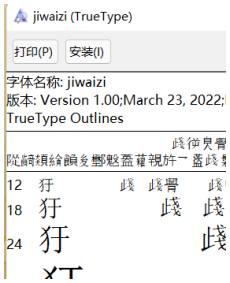

图 25 .ttf 文件

由 Font Creator 软件生成的字库如下图 26 所示：

图 26 字库展示

3.2.4 小结

本章首先利用两种不同方法实现造字法后生成的集外字，现整理如下：

方法一：利用 windows 自带专用字符编辑程序，它可编辑 eudc 字体文件，参照集内字的部首结构，创造出一部分集外字，并保存在相应的 Unicode 私用区码位，形成集外字字符映射表，将集外字与对应的私用区编码的对应关系制成图表。该法对于部首结构较为工整的集外字造字速度较快，所造字因为是对集内字字形的拆分，所以字形工整。

方法二：使用 High-Logic 公司的 Font Creator，该款软件是曲线轮廓造字软件，根据直线和二次 β 样条曲线，将扫描到的集外字字符，按照字符形状自动拟合至接近原始字符的笔形位置，通过调整文字的轮廓点、线、角度以及位置，造出较为理想的集外字并保存在对应的 Unicode 私用区编码，此法作为方法一的补充。对例如里耶秦简 9-475 简牍中的“  ”和“”字，用此方法操作简便，所造字还原度高。

其次，选择所整理集外字的方块字或临摹字字符，为其分配好对应的 Unicode 编码，将所造集外字与此 Unicode 对应；

第三，将此文字所对应的编码导入与之对应的字模文件，存储成扩展名为“.ttf”的 Open Type 文件，从而生成“集外字字库”。

图 26 为所创建集外字字体库中的部分字体样例，样例中的字形基本保持了集外字原本的字体风格，验证了本章节所提出的建立字体库方法的有效性。本文所讨论的出土文献集外字字体库的建立，其所造字形尽可能最大程度还原其在原文中所存在的形态，通过提取部首构件元素重新组合构成新的字模，最终生成字库，本字库的目标不局限于集外字的录入及显示，更重要的是其可参与计算机的文本信息处理，所以对于临摹字，我们只需尽可能的还原其在原文中的字形，而不必追求与原文字形的分毫不差，因为计算机所识别的不是此字在字库中的字形，而是其所对应的机内码，也就是 Unicode 编码，故在造字时不必过分追求字形的分毫不差，否则会导致集外字数量巨大、私用区码位溢出不够用。

在建成本集外字字库后，为方便本库中集外字在出土文献语料库中的输入及输出，需要建立本字库所对应的集外字输入法，而输入法码表是建立输入法的主要依据文件，码表中规定了调用字符的键盘方式以及该字符所具有的其他属性，是非常重要、基础且必不可少的文件。

第四章　集外字自定义输入法的建立

集外字字库建成后，每个集外字都有其唯一对应的机内码，但是仅有字库对于建设“出土文献语料库”远远不够，还有两个问题亟待解决：1.提升在建库时对集外字的录入效率。通常在建立集外字字库后录入集外字时，需要在字库的字符映射表中找到所需集外字，对集外字进行复制并粘贴录入至语料库中，此方法在字库中字的数量较少时效率较高，当字数量超过 100 时效率就会大大降低；2.实现集外字在其他电脑平台的输出共享。通常想要实现字库共享，需要将本字库.TTE 和.EUF 文件拷贝至其他电脑的 Fonts 系统文件的子目录内，而 Windows 系统规定不允许用户直接对系统文件进行更改，故需要用户使用 DOS 命令修改添加字符集字库，并且需要用户自行添加字符集注册表，若未修改系统字体文件注册表，会造成与系统其他字库文件冲突，造成集外字显示混乱，此方法操作步骤复杂繁琐，不宜推广。

方正公司所制作的古联瀚字输入法，通过将方正字库与检字方案结合起来，实现古籍生僻字的输入输出。输入法为连接字库和字符映射表以及键盘提供了连接工具，同时输入法也为字库的传播共享提供了平台，有了输入法即可实现对字库中所有字符的调用，也可实现不同计算机之间的集外字交流共享。本研究建立了基于 Windows 平台下 IMM-IME 框架的集外字输入法，输入方案的构成有集外字字库、输入法码表和字库注册表，其中集外字字库是输入法的基础文件，用以对集外字的存储；输入法码表是输入法通过键盘调用集外字的检字方案，码表中规定了对每一个集外字的检字编码；字库注册表是实现输入法在不同 Windows 电脑兼容的最终自运行注册文件，可省略繁琐的 dos 命令操作。

4.1 IMM-IME 框架输入法体系

计算机对于英文的输入转换流程相对简单，将键盘所键入的组合转换为与之对应的 ASCII 编码，将 ASCII 码进行编译显示即可[46]，但是对于中文此类方块字的输入，则转换流程相比英文就会复杂很多。

首先，需要对方块字进行编码，将捕捉的键盘组合转译为对应的机内码；其次，根据机内码与字符集编码之间的计算规则进行相应转换；最后，对转换后的字符集编码进行编译显示。IMM-IME 框架的输入法是 Windows 为了简化此流程提供的[46]。因此，本文所实现的集外字输入法就是基于此框架下的输入法，其工作原理如下图 27

所示：

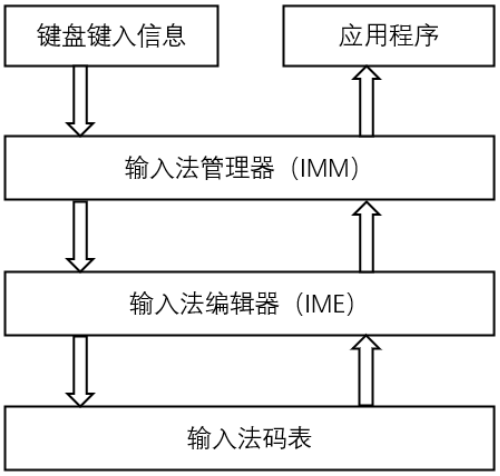

图 27　IMM-IME 输入法工作原理

4.1.1 输入法管理器(IMM)

输入法管理器(IMM)统一管理系统中的输入法编辑器(IME)，用户不可以直接选择 IME，而是要通过 IMM 选择所需的 IME，例如搜狗、微软拼音、出土文献集外字输入法。IMM 根据 IME 所提供的状态信息决定输入法执行的下一步操作，其提供了多种供 IME 调用的函数，负责 IME 与应用之间的互动操作。

4.1.2 输入法编辑器(IME)

IME 是微软公司为了统一计算机的编码输入规范而定制的一套输入法编程规范，所有 Windows 系统的计算机都需要遵从这套规范，规范里定义了输入法需要的基础操作，例如捕捉键盘输入信息、光标跟随、编译机内码至相应字符集中的字符、字符在应用程序中的显示，若要建立自定义输入法就要遵循 IME 的规范，IME 提供了对应操作的工具函数，开发者只需按照既定规则调用编写工具函数即可，若开发者未遵循此套规范而建立自定义输入法那就需要再重新编译实现 Windows 系统下的键盘捕捉、光标跟随这些功能的代码，这无疑会增大工作量。基于 IME 实现自定义输入法只需将其视为转换接口，编写、调用最重要的几个接口函数即可，这将大大提高工作效率，实现 IME 的接口函数编写可分为 UI 函数和 CI 函数两部分，其中 UI 函数主要用到的有 8 个功能函数，CI 函数主要用到的有 5 个功能函数：

UI 函数：

(1)WM_CREATE　　##传达创建输入法状态窗口、编码窗口、候选窗口消息

(2)WM_IME_SETCONTEXT ##传达窗口显示、拖动、隐藏消息

(3)WM_IME_STARTCOMPOSTION ##传达用户开始按键消息

(4)WM_IME_COMPOSITON ##传达用户编辑字符串消息并改变当前输入状态消息

(5)WM_IME_ENDCOMPOSITON ##传达用户结束编辑字符串消息

(6)WM_IME_SELECT ##传达用户选定新输入法编辑器消息

(7)WM_IME_NOTIFY ##传达当前 UI 组件状态消息

(8)WM_IME_DESTROY ##传达关闭状态窗口、候选窗口和编码窗口消息

CI 函数：

(1)ImeInquire ##获取输入法属性、UI 界面窗口信息名称设置

(2)ImeProcessKey ##判断键盘键入字符是否需要编译转换，若需要转换则返回 True，若不需要转换则发送至应用程序，当字符选择窗口出现多个集外字时，数字键 0-9 被认定为选择键，此时键入与集外字候选列表中对应的序号即可输入目标集外字；当无字符选择窗口时键入数字键会将数字直接输入应用程序

(3)ImeSlect ##启动输入法需调用的函数，控制输入法的启动、关闭以及数据初始化

(4)ImeToAsciiEx ##自定义输入法最重要的函数，当 ImeProcessKey 返回值为 True 时，IMM 激活此函数，首先捕捉键盘组合，然后按照键盘组合在字符映射表中寻找与之对应的字符，最后形成字符上屏序列表，同时当捕捉到键盘输入为 enter、backspace 等时执行与之对应的操作

(5)ImeConfigure ##当用户需要设置输入法相关属性时，被 IMM 调用，为用提供可选择的属性配置。

4.2 输入法码表

输入法码表是自定义输入法的主要核心文件，文件扩展名为.txt，一般的码表文件支持 Unicode 编码格式、ASCII 编码格式、UTF-8 编码格式，最简单的码表文件中包含了字符、词组以及所对应的检字编码，文件内容可分为两列，一列为字符或词组，另一列为字符所对应的检字编码，两列之间用水平制表符“Tab 字符”分割开。但是为了丰富 ImeProcessKey 访问码表文件的功能，一个码表中可包含更多的字符信息，

可在码表定义的末尾添加几个标记用以表示该行词条具有的特殊属性，本文所使用的输入法码表格式为多多输入法码表，其码表可添加的词条属性有：

- #固 固定词条，该词条总是显示在上屏候选中的最前列，且不受词频调整影响，对于词频较高的集外字可添加此属性。
- #用 指定当前词为用户自定义词
- #辅 辅助码词条，作为主输入方案的补充方案，例如在四角号码和结构部首的自定义输入方案中，四角号码为主输入方案，结构部首为辅助方案
- #次 次显码表，可将部分不常用字设置为次级显示，在普通显示候选列表时只显示主码表，当切换次显码表时会显示次级词序列
- #类 词条所属分类标签，定义方式为“#类+1-16”
- #序 候选列表排序，定义方式为“#序+0-65536”，后面的 0-65536 代表的是该字符或词组的出现频率，频率越高，排序越靠前

为了便于统一对候选列表中的字符和词组进行管理，对码表表头定义了两种排序数值，第一种为在表头添加“---config@全局排序数值=1000”，初始设定此码表中的所有词条的排序数值为 1000，第二种为在表头添加“---config@全局排序数值=2000”初始设定所有字符与词组的排序数值为 2000。在输入法配置程序中，可供选择的词频方式三种，分别为“一次到顶、逐渐调整、累加上屏”，“一次到顶”的词频调整策略是对于用户字词上屏次序为首位，无需调整，若上屏次序不为首位，则将用户上屏字词调整至首位，其他候选词序不变；“逐渐调整”的词频调整策略是对于用户字词上屏次序为首位，无需调整，若上屏词序不是首位，则将用户字词调整至当前候选排序的 1/2 位置，其他字词候选序列不变；“累加上屏数”的调整策略是当用户每选择一次上屏字词，则该字词在上屏候选列表的序号加 1。三种方法对于集外字字频调整策略各有优势，例如对于长沙走马楼三国吴简中的“  ”字，在释文中出现频率非常高，相当于常用字，就可将其在输入法码表中设置为“一次到顶”或“#固”，使得该字始终显示在上屏候选列表中的最首位，提高输入效率。

4.3 集外字输入法码表检字方案

现有的输入法码表中的检字方案大致可分为两类，分别是音码类检字法和形码类检字法，音码类检字法就是依据字符的拼音，根据拼音组合对字符进行编码并导入码表，主要的音码编码有全拼、双拼；形码类检字法顾名思义是按照字符外形结构对其

进行编码并导入码表，主要的形码检字法有五笔、郑码、四角号码编码。

音码是根据字符的声母、韵母的组合将所有字符进行归类，此编码方案的优点是简单易上手，缺点是重码率高且要求使用者必须熟悉所需字符的拼音，对于出土文献集外字而言，使用者若非专业古文字研究人员，无法正确输入其拼音，所以全拼或双拼形式音码都不适合作为本输入法的码表的检字方案。

形码是根据字符的笔画、部首、结构将所有字符进行归类[46]，此编码方案的优点是使用者可以不用了解字符拼音，只需根据笔画及部首结构即可输入字符对应编码，缺点是形码编码体系复杂，学习成本高。形码中的五笔编码法是根据汉字笔形将其分为撇、捺、横、竖、折 5 类笔形，每类笔形对应键盘中 5 个键位，5 个键位中又有此笔形对应的部首和结构，使用者必须熟练背诵五笔字根表才可掌握此检字法；郑码编码法类似于五笔输入法，将汉字的起笔分为横、竖、撇、点、折五大笔形，每一起笔的笔形对应数个键位，每个键位有对应的字根，使用者需熟练掌握郑码字根表和编码规则才可掌握此检字法；四角号码编码法是根据汉字结构将其分为十种笔形，笔形对应的编号为 0-9 之间的一个数字，任何一个汉字都可按照其左上角、右上角、左下角、右下角的笔形顺序进行编码，即对于任一汉字都可由 4 位阿拉伯数字的组合进行描述。以上编码方案中五笔编码与郑码编码对于使用者而言，需要背诵字根表，且字根表对应部首之繁琐，若非有专业教师教学，短时间内掌握较为困难，所需学习成本较高。但是四角号码编码法只需记忆十种笔形，有对应的记忆口诀，经过专业训练，半天时间即可掌握，易上手且学习效律高。

出土文献集外字字形分为两种，一种是方块字，可拆分为规整的部首结构且此类字在字库中占比超过 99%；另一种为临摹字，无法拆分为规整的部首结构且占比不足 1%，所以形码编码方案可描述出绝大多数的集外字。同时对于古文字不是非常精通的人，他们无法正确拼写字库中集外字的拼音，相较于音码码表，形码码表不用了解集外字拼音，只需正确切分笔形即可输入所需集外字，故形码编码方案更适合作为集外字的输入法码表编码方案。

根据上一小节对于现有检字方案的对比，就各类编码方案的重码数而言，四角号码检字方案在处理集外字方面优势比音码更大。以现代汉字为例，音码的重码率最多有 100 余个[47]，故四角号码检字法更适合作为集外字输入法检字方案。

四角号码是根据汉字笔画构造的特点，将汉字的四个角转换为一组数位码，并具有其唯一性[48-53]。其构造过程如下：首先，按笔形查字。汉字的笔形分为十类：头、

横、垂、点、义、插、方、角、八、小，并分别用数字 0-9 表示；其次，将汉字分成左上、右上、左下、右下四个角，且每个角会根据其笔形确定一个数字编码，此时一个汉字会被表示为一串四位数组；最后，给上述四位数组的末位再加上一个附加码。附加码是为了降低重码率而产生的，其方法是取靠近右下角上方一个笔形作“附号”。

因此将所有字可拆解为从 00000-99999 的编码。汉字对应的笔形、编码及说明规则如下表 4 所示：

表 4　集外字笔形代码规则及说明表

笔名		编码	笔形	规则
复笔	头	0	亠	取角方法: 1.一笔可以分角取号； 2.一笔的上下两段和别笔构成两种笔形的，分两角取号； 3.下角笔形偏在一角时，按实际位置取号，缺角作 0，但弓亏等字作偏旁时，取 2 作整个字左下角编码； 4.凡外围是“口、门、門”的三类字，左右两下角改取里面的笔形，但上、下、左、右有附加码的字都不在此列； 5.一个笔形如果前角用过，那么后角作为 0。
单笔	横	1	一 亻 𠂉 𠃌 乀	
	垂	2	丨 𠂆 丿 亅	
	点	3	丶 𠂉 乀	
复笔	义	4	十 寸 七 义 犭 𠂇	
	插	5	丰 扌 戈 丈 𢦏 丰	
	方	6	口	
	角	7	冂 冖 乚 𠂉 厂 𠂆 亅	
	八	8	八 𠆢 人 入 𠆢 𠂇 𠂉	
	小	9	小 忄 木 灬 巛 灬	

四角号码检字法根据字形获取编码，根据编码查找字形，其优点在于：1.容易学习，认真学习只需半天甚至几小时即可掌握；2.对于读者不认识或部首结构难以区分的字，查字速度快。但四角号码检字法目前也存在缺点：除《四角号码新词典》使用四角号码排序外，其他例如《新华字典》、《现代汉语词典》等也用四角号码检字法，因为作者不同，导致每个人对四角号码检字法的理解不同，进而对汉字的取角原则也有所差别，为统一输入法码表方案[52]，本文在使用四角号码时遵循《四角号码新词典》第十版的规则，对集外字进行编码拆分。以集外字“𡍮”为示例，其对应四角号码为“30212”，“3021”是基本号码，“2”为附加码。因为输入法码表中只能输入英文字母，所以需将“0123456789”用其拼音声母首字母“oiesxwlqbj”代替[28]，按照此规则，“𡍮”字在输入法码表中的描述结果为“soeie”。

基于上述的集外字四角号码检字方案以及多多输入法码表规范，建立所需的 txt

输入法码表文件，使用码表文件建立集外字输入法。

4.4 生成输入法

配置生成输入法需要八个步骤，其具体内容如下：

第一步：设定软件信息

- IME 文件名：在用户计算机 system32 中创建的 IME 文件名，指定为“MyIME”；
- 安装目录名称：在用户计算机 Program files 中创建的目录名称，指定为“集外字输入法”；
- 开始菜单名称：在用户计算机开始菜单中创建的菜单名称，指定为“集外字输入法”；
- 更新链接：输入法更新所访问的网站链接，如图 28 所示；

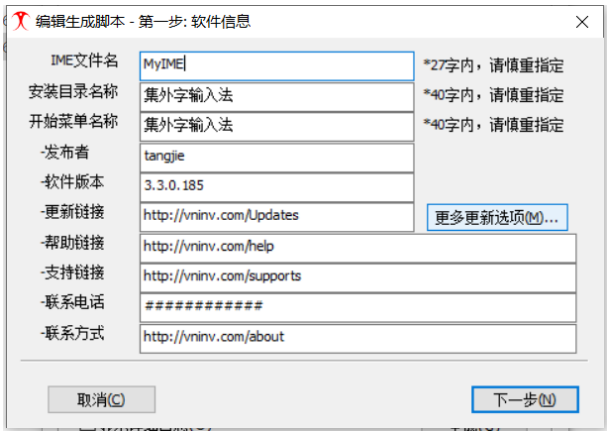

图 28 软件信息界面

第二步：驱动选择

- 原始码表文件：输入法码表文件，插入 5.2 小节所建立的 txt 集外字输入法码表文件，包含 2 个码表文件，主码表文件和辅码表文件，主码表包含集外字极其对应四角号码，辅码表包含集外字及其对应的部首构件拆分；
- 图标集目录：定义输入法栏外在图标；
- 最终用户帮助文件：使用 easy chm 编写 CHM 格式的 HTML 帮助文档，为用户提供软件使用帮助；
- 语言栏输入法名称：定义语言栏名称；
- IME 驱动设置：导入_ime_classic 驱动;
- 输入法语言标识：IME 驱动中的文字区域，选择与集外字字符集编码相匹配

的 Unicode 字符集；

- 安装程序保存路径：输入法所安装的文件路径设置，如图 29 所示：

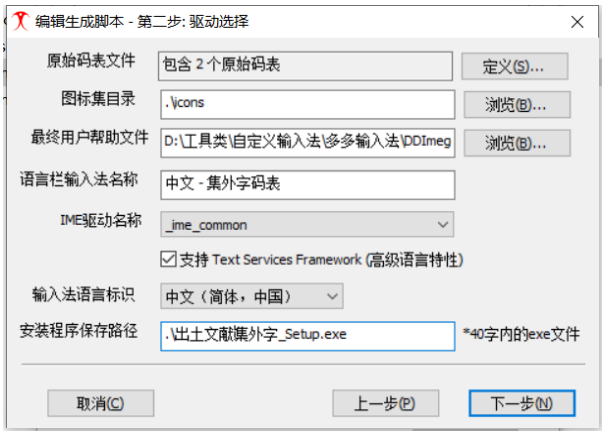

图 29　驱动选择界面

第三步：生成选项

- 包含联想库：上屏联想词组，可建立出土文献词库的 TXT 文档，将词库导入，若用户词包含词组中的任一字会自动上屏该词组；
- 支持用户造词、临时造词：连接用户最近上屏的词组并将其储存至临时词库，提高输入效率；
- 定义中文标点：定义键盘对应的标点符号；

定义按键功能：定义键盘按键功能，例如切换主次显码表、前翻页、后翻页等，如图 30 所示：

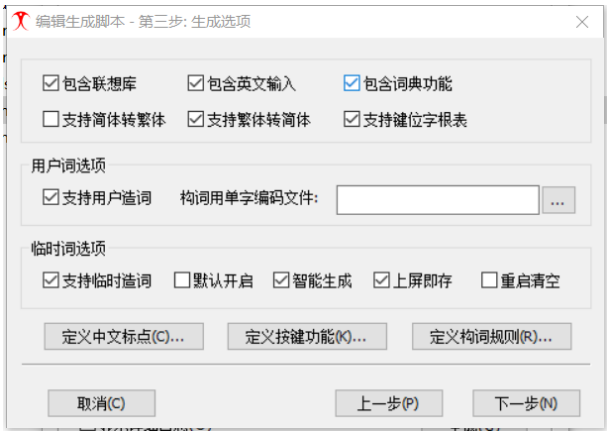

图 30　生成选项界面

第四步：特性设定

- 空码时空格行为：上屏英文为按下空格直接上屏所输入字符，上屏英文+空格为按下空格直接上屏所输入字符和空格，清屏为按下空格清除所输入上屏字

符；

- 输入系统行为：开启命令直通车时命令提示符定义的命令有效；
- TSF 嵌入窗口行为：仅重码时显示候选窗口是当用户词候选字符为 0 时直接上屏所输入字符，不显示候选列表；隐藏嵌入首候选项是候选列表中不含已上屏的第一字符，
- 初始化转化状态：可定义打开软件时的初始状态，如图 31 所示：

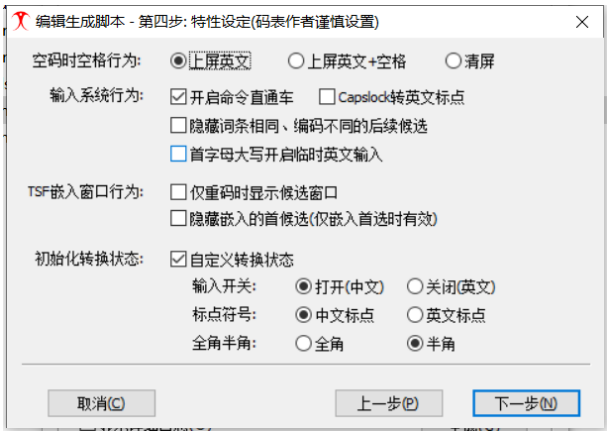

图 31 特性设定界面

第五步：IME 默认设置

- 候选列表排序依据：定义候选列表中字符的排序规律；
- 词频调整：根据字频定义候选列表中上屏字的排序；
- 输入方案：根据主辅码表的检字方案，包含四角号码检字及部首构件检字；
- 检索范围：检索键入编码的字符范围，范围为 Unicode 用户自定义区字符集，如图 32 所示：

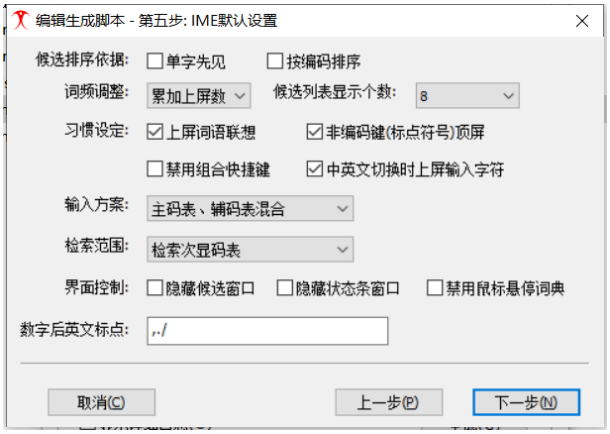

图 32 IME 驱动设置界面

第六步：DME 驱动设置

- 码长选项：定义键入编码长度与其上屏候选列表中的字词；
- 查询选项：定义键入字符后的查询输入；
- 引导符、截止符设定：设定引导或截止键，控制上屏或清屏，如图 33 所示：

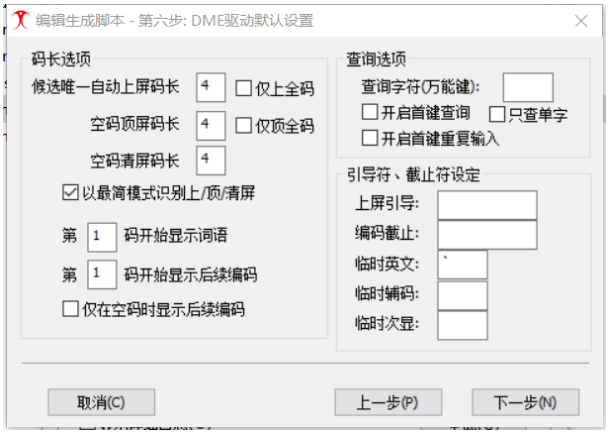

图 33　DME 驱动设置界面

第七步：SKIN 驱动设置

- 定义输入法的窗口布局方式、输入栏嵌入方式、光标符号以及输入栏皮肤如图 34 所示：

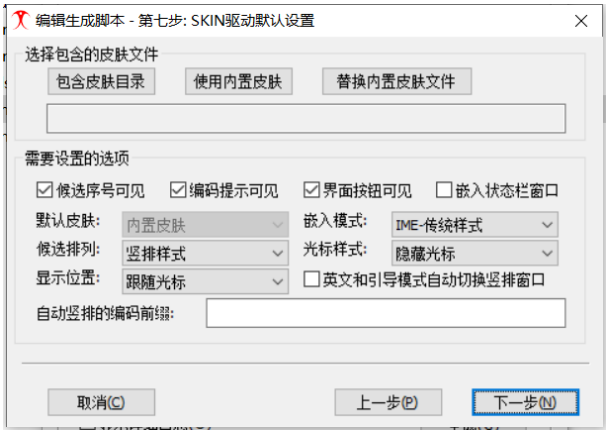

图 34　SKIN 驱动设置界面

第八步：软件安装设置

- GUID：输入法的唯一性标识；
- 码表密码及密匙：定义用户安装本输入法所需的密码及与密码相对应的密匙，如图 35 所示：

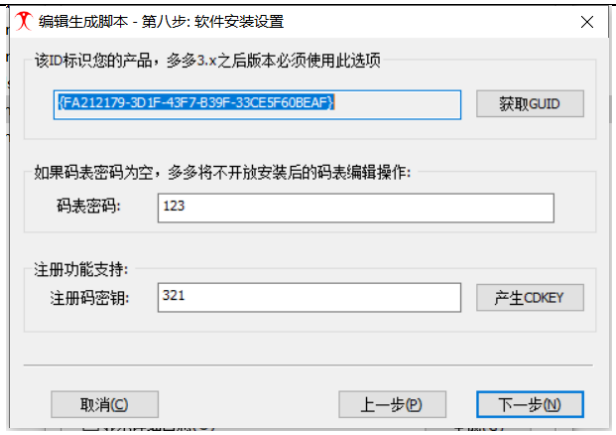

图 35 软件安装设置界面

第九步：软件内容设置

- 字体安装支持：在安装包中导入集外字字库文件；
- 开始菜单项：在开始菜单项中定义为用户可提供的功能辅助；
- 配置程序项：配置程序时可展示的项目，如图 36 所示：

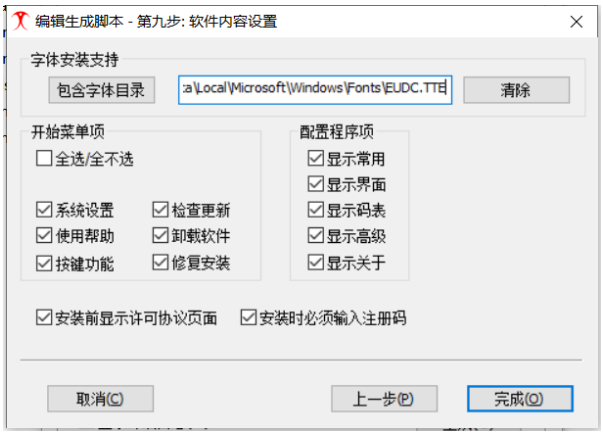

图 36 软件内容设置界面

完成上述操作步骤生成集外字输入法安装包程序“出土文献集外字_Setup_x64.exe”，用户只需运行安装程序即可完成集外字输入法安装工作。因为本输入法无法自动在用户计算机内完成字符映射工作，所以还需将集外字字库文件添加至系统注册表文件中，才可完成字符映射、字符显示以及参与文本信息处理。传统的添加注册表需在 dos 命令中完成，过程繁琐复杂，故本文提供一种可直接运行的“字体回溯注册表”文件，可使用户便捷完成字库注册工作。

4.5 集外字输入法码表检字方案

4.5.1 计算机字库注册表

由于集外字所对应的 Unicode 私用区字符集不是系统常用字体，属于扩展区字符集，一般系统是默认禁用这些补充字符集的，若只是将字库拷贝至用户电脑中不做其他操作，则字库中文字无法正常显示及使用。为集外字输入法可正常使用同时可以对集外字进行文本信息处理，还需要手动将字库文件添加到系统常用字符集中，在此本文提供一种便捷的方法完成字符集添加工作，根据系统注册表中的键(key)、分支结构(branch)、值项(Value)等数据结构在系统中添加集外字字体库链接及字体回溯键值项，完成对集外字库的跟踪配置。

注册表(Regedit)是计算机操作系统的管理核心，类似于一个功能丰富的语料库，不仅可以储存计算机的软硬件配置信息、用户信息、所有系统文件类型及关联程序，还可以记录输入法的配置及安装信息，并驱动输入法运行。系统注册表分支结构有 5 个，其分支如下表 5 所示：

表 5 系统注册表分支

表 5 系统注册表分支

名称	说明
HKEY_CLASSES_ROOT	存储系统文件及关联程序信息
HKEY_CURRENT_USER	存储实时用户的配置信息
HKEY_LOCAL_MACHINE	存储计算机软硬件信息
HKEY_USERS	存储所有用户的配置信息
HKEY_CURRENT_CONFIG	存储计算机当前配置信息

注册表常用 3 种键值数据类型如下表 6 所示：

表 6 系统注册表键值类型

名称	说明
二进制值(reg_binary)	二进制数
字符串值(reg_sz)	由字母和数字组成，一般用来描述文件
双字节值(reg_dword)	双字节类型数字

注册表修改值有两项，第一项是修改 HKEY_LOCAL_MACHINE 分支中的 LanguagePack 子键值，将禁用补充字符集项修改为允许；第二项是将补充字符集添加

至 HKEY_LOCAL_MACHINE 分支中的字体回溯(SurrogateFallback)子键中，具体注册表代码如下：

```
REGEDIT4

[HKEY_LOCAL_MACHINE\SOFTWARE\Microsoft\Windows NT\CurrentVersion\LanguagePack]
"SURROGATE"=dword:00000002

[HKEY_LOCAL_MACHINE\SOFTWARE\Microsoft\Windows NT\CurrentVersion\LanguagePack\SurrogateFallback]
"Plane10"="jiwaizi"
"Plane11"="jiwaizi"
"Plane12"="jiwaizi"
"Plane13"="jiwaizi"
"Plane14"="jiwaizi"
```

4.5.2 输入法选择及界面介绍

在用户点击 Windows 语言栏选项后，可选择集外字输入法，选择“中文-集外字输入法”后会出现输入法的界面，用户即可进行集外字输入，输入法状态栏及输入界面如下图所示：

图 37 集外字输入法状态栏

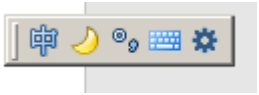

图 38 集外字输入法界面

软件界面信息如下：

- 图标“  ”表示输入法处于集外字输入状态；

- 图标“  ”表示输入法软键盘；
- 图标“  ”表示输入法的设置选项；
- 设置选项中包含输入法的属性设置、软件使用帮助、上屏选项等功能，如图 39 所示；

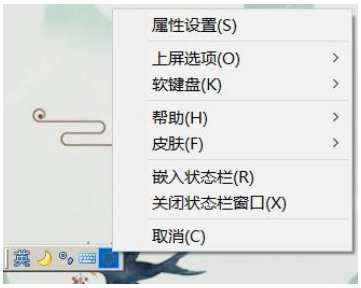

图 39　集外字输入法设置选项栏

- 设置选项的帮助栏下有软件更新检查，支持输入法及时更新，如图 40 所示：

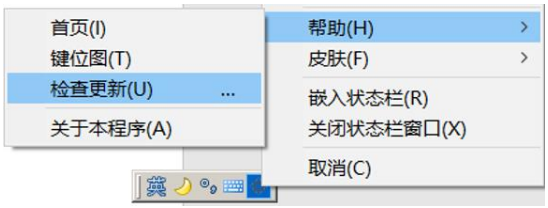

图 40　集外字输入法检查更新栏

4.5.3 集外字输入

当输入法界面正常显示时，表示处于集外字输入状态，此状态下输入所需集外字对应的四角号码编码，即会出现字符输入候选栏。候选栏中有“1-8”键，PgUP 和 PgDn 表示翻页，按键功能与传统输入法无差异，按数字键或空格键选择所需集外字后会自动清屏，等待用户下一次输入。例如想要输入集外字“”，则可直接键入其所对应的四角号码编码“30212(soiei)”进行选取，输入状态如下所示，如图 41 所示：

图 41　集外字输入法字符候选列表

4.6 小结

本章首先从 IMM-IME 框架输入法体系方面，详细论述了建立集外字输入法所用的框架及其主要功能函数；其次，从输入法码表方面在论述了其主要功能函数及数据要求的基础上，对集外字的检字方案进行论述，从而建立了适应于小组出土文献语料库的集外字子库输入法码表。

最后，为了实现集外字在语料库中的输入、输出、显示及检索，本文尝试利用现有的输入法生成器，导入建立好的集外字输入法码表及字库，配置并设定输入法相关参数及字库注册表信息，最终获得“出土文献集外字输入法”。该输入法的优点是，第一，检字便捷迅速，方便出土文献集外字的输入与输出；第二，可编辑性高，字库可持续更新且可导入简牍词表并调节上屏字词；第三，可共享，在第三方电脑端安装字库及输入法安装包即可录入出土文献集外字。

生成集外字子库的输入法不是本研究的最终目的，这也是为了集外字能够参与计算机利用数字人文技术处理出土文献全文本。以下，将对出土文献的数字人文应用进行讨论。

第五章　集外字应用于文本信息处理

本文所讨论的“出土文献语料库”是涵盖目前所有出土文献文本信息的语料库。语料库中包含文本中的词表、词性标注、语义信息、语法信息等内容，建立语料库的目的不止于文本显示及检索，同时还需将语料库库中的数据应用到文本信息处理过程中。因此，在建立集外字字库之初，就应当考虑能够使得集外字具有可被文本信息处理的功能。

但要实现此目的，需考虑集外字在语料库中的存储形式。目前研究者们通常采用“图片+文字”的形式数字化出土文献文本，但此方式不能直接被计算机进行文本信息处理，是因为在图片与文字混合的语料库中，两者无法同时满足语料库的计算机应用：其一，图片与文字在语料库中的储存形式有根本的不同，图片在语料库中是以二进制类型的形式存在，而文字是以字符类型的形式存在于语料库中；其二，两者在语料库中具体使用的设置参数方法及数据调取方法不同，所以图片与文字虽然可同时存在于语料库中却无法同时参与到文本信息处理过程中。

为了使集外字可应用于文本信息处理过程中，并且可适用于功能性语料库的需求，需要将集外字转换为以字符类型存在的形式。本文所尝试建立的集外字字库不仅使得集外字转换为字符类型存在的字符串，以满足统一语料库中数据存在格式的要求；还在集外字与集内字都以字符类型储存在语料库内的基础上，实现全文本的计算机应用。

在本论文第四章节中已经论述过使用本方法所建立的字库，其中的字符能够被应用于信息处理。本文还尝试了将“单个字符维度”的处理手段上升至“词组维度”，甚至“段、句、全文本维度”，再利用计算机进行文本处理。检验这种处理方式的手段之一是“中文分词”。

中文分词（Chinese Word Segment，简称 CWS）是将文本切分成多个具有语义信息的最小的单位“词组”的过程。分好词可更好地分析文本语义，同时分词是实现命名实体识别、文本分类、文本挖掘等中文信息处理工作的基础任务，所以将中文分词作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是非常合适且重要的。若集外字可在计算机中实现分词这样的基础工作，则其同样可应用于文本挖掘、文本分类、聚类等类似文本信息处理过程中。

5.1 常用中文分词方法

5.1.1 基于规则的分词方法

规则分词是在词组维度对文本进行分词的分词方法，是文本信息处理中最基础的分词方法，它的工作原理是：首先建立一个词典并在其中加入可以被切分的词组，其次按照规则的方法将待切分文本与词典中的词组进行匹配，若该文本中有连续的字符串与词典中的词组相同，则认定该字符串为词组并进行截取切分。接下来去除已被切分的字符串，并对文本中的剩余字符串开始新的匹配切分工作，直至剩余字符串为空，结束切分。分词方法根据截取字符串方向的不同，可分为正向最大匹配、逆向最大匹配、双向最大匹配[54-57]：

正向最大匹配法（Maximum Match Method，简称 MM）

算法思想：对于一个待切分文本，指定其最大词组长度为 L，将待切分文本从左至右选取字符长度为 L 的文本，与词典中的词组进行匹配，若匹配到最大长度词组，则将该词组切分并保存，并从该词组最后一个字符串后面的字符作为初始字符，截取长度为 L 的文本，重复上述匹配工作，直至文本字符长度为零时结束。

逆向最大匹配法（Reverse Maximum Match Method，简称 RMM）

算法思想：对于一个待切分文本，指定其最大词组长度为 L，将待切分文本从右向左选取字符长度为 L 的文本，与词典中的词组进行匹配，若匹配到最大长度的词组，则将该词组切分并保存，并将该词组首个字符前面的字符作为初始字符，截取长度为 L 的文本，重复上述匹配工作，直至文本字符长度为零时结束。对于现代汉语文本而言，逆向最大匹配算法的比正向最大匹配算法的分词准确率高。

双向最大匹配法（Bi-directction Matching method，简称 BMM）

算法思想：对于正向最大匹配算法和逆向最大匹配算法的切分结果进行比较，若正向匹配的分词结果词数最少，则输出该结果，反之则输出逆向最大匹配的切分结果。绝大多数情况下两者的切分结果相同，只有较少数情况下切分结果不同，需要选择分词数最少的结果输出。

基于规则分词的优点是：第一，不需要大量的语料库支撑，代码简单；第二，只要有完整的词表，所切分的准确率接近百分之百；第三，对于有歧义的切分句式，可以添加对应的词表以及切分规则，较好的解决歧义切分问题。

基于规则分词的缺点是：第一，无法识别未登录词，假如该词不存在于现有的词典中，则无法对该词组进行正确切分；第二，需要完整的词表支撑，而构建完整的词

表需要花费大量的人力、物力，例如出土文献中有大量的人名、地名，要将其全部统计并分类，整理为可用的词表工作量巨大；第三，处理歧义准确率低，例如现代汉语中：“南京市/长江大桥”和“南京/市长/江大桥”就是两种歧义切分结果，这类切分在出土文献语料中也同样存在。

5.1.2 基于统计的分词方法

统计分词是基于语料库的全文本维度的分词方法，语料库是在人工对全文本进行标记再加工而来的，统计分词是对已有的语料进行训练得到参数，根据参数去预测新的语句的分词概率，语料库的准确性影响分词结果的准确性。现有的公开语料库有微软语料库、1998 年的人民日报语料库等。

基于统计的分词方法包含 N 元模型（N-gram）、隐马尔可夫模型（Hidden Markov Model，简称 HMM）[58]、条件随机场模型（Conditional Random Field，简称 CRF）等。本实验的目的在于验证所造集外字可以参与到基于统计的分词过程中，因此选择了隐马尔可夫模型进行分词实验。

在 HMM 模型中，语料库中对每个字词都会有一个标识，标识当前字处于语句中的位置，标识可分为四种分别是：词语首端(B)；词语中间(M)；词语尾端(E)；单独成词(S)[59]。HMM 分词的工作原理是通过对现有语料库中语料数据进行训练，通过模型训练得到语料数据三大概率矩阵的参数，分别是初始概率矩阵、转移概率矩阵、发射矩阵。其计算方式如下：

初始概率矩阵：根据对语料库中的语料统计，计算初始文本的 B、M、E、S 四个标识在整个语料库中出现的频次，一般初始频次只计算 B 和 S 两项标识，M、E 标识的频次均为 0。

转移概率矩阵：是当前状态到下一状态的概率，例如以标识 B 为例，计算 B 状态到 B 状态、B 状态到 M 状态、B 状态到 S 状态、B 状态到 E 状态的概率，其中 B 状态到 B 状态和 B 状态到 S 状态的概率恒为 0。

发射矩阵：统计某种状态下所有字出现的频次，例如在 B 状态下语料库中所有字出现的频次，在 M 状态下语料库中所有字出现的频次，在 E 状态下语料库中所有字出现的频次，在 S 状态下语料库中所有字出现的频次，将语料库中每个字的统计结果转换为概率矩阵。

在通过训练得到以上参数后，根据得到的参数统计待切分文本每个字词的初始矩

阵、转移矩阵、发射矩阵的概率，计算出待切分文本的所有路径概率，通过维特比算法选择最优的概率结果对待切分文本进行分词。

基于统计的分词方法优点是：第一，通过数学概率模型计算，可消除句子切分歧义；第二，切分准确率与语料库规模成正比，只要语料库规模足够大，准确率就会更高。

基于统计的分词方法缺点是：第一，现代汉语语料库、传世文献语料库、出土文献语料库不能通用，出土文献文本分词需要对应的出土文献语料库；第二，构建相应的语料库除了需要构建数字化文本，还需对文本进行标注，工作量巨大；第三，对于复杂的统计模型以及规模较大的语料库，模型训练时间久，分词时间长。

5.1.3 主流分词工具

(1)jieba 分词工具

jieba 分词工具是基于 python 语言可实现汉语自动分词。该分词工具的优点有：第一是该分词系统允许用户根据实际所需添加用户自定义词典，这对于以生僻词组居多的出土文献文本而言有较高的分词准确率；第二是该分词系统算法中既包含有规则分词又包含统计分词，用户可手动开启或关闭隐马尔可夫链。隐马尔可夫链的作用是，当句子中出现词典中未包含的未登录词时，会根据概率计算该词与下一个词的匹配概率，对该词进行分词；第三点是该分词工具是 python 中的第三方库，调用简便且可提供三种分词模式：

精确模式：对文本进行精确切分，jieba 的默认切分模式；

全模式：对文本进行全扫描，切分出该句所有可能成为词组的词；

搜索引擎模式：对文本进行精确切分之后，在对切分的结果进行切分。

以现代汉语“我在西安西北大学上学”这句话为例，经过三种模式的切分后对应的分子结果如下：

精确模式：我/在/西安/西北大学/上学

全模式：我/在/西安/安西/西北/西北大学/北大/大学/学上/上学

搜索引擎模式：我/在/西安/西北/北大/大学/西北大学/上学

在分词实验过程中根据实际需要，本文选择精确模式对文本进行切分。

(2)Hanlp 分词工具

Hanlp 是哈工大开发的分词工具，基础语言为 java，有适用于 python 调用的第三

方工具包，其分词结果默认附带词组词性，且语料实时更新，对于词性标注、命名实体识别等需要文本分词、文本挖掘的场景非常友好。Hanlp 分词工具自身内含词典，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。

5.2 实验数据及实验设计

5.2.1 实验数据

里耶秦简牍于 2002 年及 2005 年两次出土，出土范围是里耶古城一号井及护城壕 11 号坑，两次共出土了三万八千余枚简牍。简牍内容主要是秦始皇廿五年（前 222 年）至二世二年（前 208 年）的秦朝洞庭郡遷陵县的官文书[60]。文书内容丰富，可被大致分为簿籍、録课、书传、符券、律令及检楬类。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 6050 条简，总字数约为 12.25 万字。本文在小组建立里耶秦建语料库的基础上，建立集外字字库，并应用于里耶秦建语料库的建设中。

词表数据为里耶秦简两卷的所有二字词、三字词、四字词等词组。本词表采用出土文献小组成员刘铭所制作的词表，作为自定义词典导入规则分词算法、jieba 分词工具中，本词表数据作为分词词典应用于分词实验。

全文本数据包含了里耶秦简两卷的所有文本内容，其中所有以图片形式存在的集外字，都被替换为所造字库中的字，并选择一部分未切分的里耶秦简语料作为测试语料。其中有 7 条含有方块集外字的简文及 3 条临摹集外字的简文，简文节选数据如下表 7 所示：

表 7　分词实验语料筛选

集外字字形	集外字	简号	释文
方块字	戔	5-5	☒戔夌公叹忐告誘☒(正) ☒戔夌行士事昌戈☒(背)
方块字	戉	5-10	☒以為戉具箸至日□☒
方块字	葢	8-190	卅一年後九月庚辰朔甲☒ 吏僕養者皆屬倉☒ 署倉非弗智殿葢☒ 亘曰上真書狀何□☒(正) 後九月甲申旦食時☒(背)
方块字	宽	8-458	遷陵庫真□ 甲三百卌九

			甲廿一 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)
方块字		8-765	卅四年九月癸亥癸酉三 卅四年八月丁未 蔓柏丞屠廷卅四年八月丙申下(正) 卅五年 三月朔日卻壬辰卻壬寅卻 正月甲辰夫 二月壬午夫 二月辛丑卻癸卯卻 正月甲辰┗丙午乙┗戊申┗(背)
方块字		8-1540	粟=五斗 卅一年五月癸酉倉是史感稟人 堂出稟隸妾嬰揄 令史尚視平 感手
方块字		9-563	……□□…… 書病敬季丈人家室毋恙也┗夫爲縣官田芋 當粟= ……(正) □□□□□□□□□□□□□季重□ 賤=臣=死骨不死謁報急敢言之(背)
临摹字		8-300	鄉守履貲十四甲☒ 鄉佐就貲一甲 ☒ 鄉佐貲六甲☒
临摹字		8-777	從人論報 擇免歸致 書具此 中
临摹字		8-1570	坐一斗酒面節弗平幸告使者

简 5-5 中的“  ”被释为“”，与“夌”构成词组。

简 5-10 中的“  ”在校释中未释，属于单字词。

简 8-190 中的“  ”被释为人名，属于单字词。

简 8-458 中的“  ”被释为防御装备，与“甲”构成词组。本条简主要内容记

录了遷陵库的武器装备数量。

简 8-765 中的“犴”被释为人名，属于单字词。

简 8-1540 中的“  ”被释为“兒”，与“嬰  ”构成词组。

简 9-563 中的“”被释为“願”，属于单字词。本条简内容为问候信件。

简 8-300 中的“  ”在校释中未释，疑似人名，属于单字词。

简 8-777 中的“  ”在校释中被删去，属于单字词。

简 8-1570 中“  ”在校释中未释，属于单字词。

该测试语料中，包含了“方块字”与“临摹字”两类字形所造的集外字，且简文形式多变，长短不一，内容丰富。因此，这部分随机选择的数据具有普适性，对于说明全体集外字的分词效果是完备的。

5.2.2 实验设计

本文将对含有集外字的文本做三个实验，并进行对比研究，具体实验描述如下：

实验一是基于“词组”维度的规则分词实验：在词组维度上，实验语料选择上一小节筛选的含有两种类型集外字的里耶秦简文本，分词算法选择规则分词中的双向最大匹配算法，该算法最终的输出结果优于正向与逆向匹配的分词结果。同时导入自定义词典，观察结果是否对集外字进行有效切分及显示。

实验二是基于“全文本”维度的统计的分词实验：在全文本维度上，统计分词模型选择 HMM 模型，实验的训练语料选择人工标注并校对的里耶秦简分词语料，在训练模型过程中将实验数据中的十条语料去除，最后用训练好的模型对带切分语料进行分词。观察其结果是否对集外字进行有效切分及显示。

实验三是主流分词工具的分词实验：实验分为两个方案。方案一，基于 jieba 分词工具和 Hanlp 分词工具的默认切分模式下，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。

5.3 实验结果及分析

5.3.1 实验一与实验二的分词结果及其对比分析

实验一采用双向最大匹配算法，其中需要导入里耶秦简词表；实验二采用 HMM 模型的分词方法，其训练语料为人工分好词的里耶两卷文本。两个实验的实验结果如表 8 所示：

表 8　实验一与实验二的分词结果

简号	实验一分词结果	实验二分词结果
5-5	戉夌/公/叡忎/告/誘/戉夌/行/士/事/昌/戈	戉/夌/公/叡忎/告誘/戉/夌/行士/事/昌/戈
5-10	以/為/戉/具/箸/至/日	以為/戉/具/箸/至/日
8-190	卅一年/後九月/庚辰/朔/甲/吏僕/養/者/皆/屬/倉/署/倉/非/弗/智/殴/葢/亘/曰/上/真書/狀/何/後九月/甲申/旦食時	卅一年/後九月/庚辰/朔甲/吏僕/養者/皆/屬倉/署倉/非/弗/智/殴/葢/亘/曰/上/真書/狀何/後九月/甲申/旦食時
8-458	遷陵/庫/真/甲/三百丗九/甲宪/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	遷陵/庫真/甲/三百/丗/九/甲宪/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
8-765	卅四年/九月/癸亥/犴/癸酉/三/卅四年/八月/丁未/犴/蔓柏/丞/届/廷/卅四年/八月/丙申/下/卅五年/三月/朔日/卻/壬辰/卻/壬寅/卻/正月/甲辰/夫/二月/壬午/夫/二月/辛丑/卻/癸卯/卻/正月/甲辰/丙午/乙/戊申	卅四年/九月/癸亥/犴/癸酉/三/卅四年/八月/丁未/犴/蔓柏/丞/届/廷/卅四年/八月/丙申/下/卅五年/三月/朔日/卻/壬辰/卻/壬寅/卻/正月/甲辰/夫/二月/壬午/夫/二月/辛丑/卻/癸卯/卻/正月/甲辰/丙午/乙/戊申
8-1540	粟/粟/五/斗/卅一年/五月/癸酉/倉/是/史/感/稟人/堂/出稟/隸妾/嬰臭/揄/令史尚視平/感手	粟/粟/五斗/卅一年/五月/癸酉/倉/是/史/感/稟人/堂/出稟/隸妾/嬰/臭/揄/令史/尚/視平/感手
9-563	書/病/敬/季/丈人/家室/毋恙/也/夫/爲/縣官/田/芋/當/粟/粟/顯/季/重/賤臣/賤臣/死骨/不/死/謁報/急/敢言之	書/病/敬季/丈人/家室/毋恙/也/夫/爲/縣官/田芋當/粟/粟/顯/季/重/賤/臣/賤/臣/死骨/不/死/謁報/急/敢言之
8-300	鄉守/履/貲/十四/甲/鄉佐/就/貲一甲/鄉佐/茉/貲六甲	鄉守/履/貲/十四/甲/鄉佐/就/貲/一甲/鄉佐/茉/貲六甲
8-777	從人/論報/擇/免歸/致/書/具/此/中/𠂤	從人/論報/擇/免歸/致/書/具此/中/𠂤
8-1570	坐/一/斗/酒/諴/面/節/弗/平/幸/告/使者	坐/一斗/酒/諴/面節/弗平/幸告/使者

实验结果分析如下：

第一，实验一与实验二都对集外字进行了分词，说明集外字精准地参与了分词每个过程。这是因为集外字保存在出土文献集外字字库中，已将其转换为集内字，有唯

一对应的 Unicode 编码，计算机在分词前已将 Unicode 编码转换为二进制的机内码，故可成功对集外字进行分词。

第二，实验一对集外字的切分准确率高于实验二。其主要原因是实验一中人工切分好的里耶秦简词表包含了所有集外字的词组。因此，在规则分词匹配过程中，含有集外字的词组都会准确被切分为词。以简 5-5 分词结果为例，“戉”在实验一中切分为词组“戉夌”，这是因为实验一中加入了含有集外字的里耶秦简词表，词表中有“戉夌”一词。正向最大匹配分词的过程为，首先算法截取文本“戉夌公叹忎”并在词表中查找该词，未查找到；则去除最后一个字“忎”，对剩余文本“戉夌公叹”进行匹配，未找到该词组；重复上述操作，直至匹配到词组“戉夌”，结束本轮匹配。并开始匹配下一文本“公叹忎告誘”，重复匹配工作，最终得到结果“戉夌/公/叹忎/告/誘/戉夌/行/士/事/昌/戈”。

第三，由于实验二首先通过对分好词的里耶秦简两卷文本进行学习，由模型计算出集外字与其他文字组成词组的概率后，经过概率计算对测试文本进行分词。其分词准确率与训练语料的相关性及体量大小呈正相关。因此，实验二的分词结果的准确率较实验一低。

5.3.2 实验三：主流分词工具的分词实验

由于简牍语言以单字词居多且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的建立的必要性。

实验数据选择里耶秦简第一卷的简 8-458 为例，简文中包含集外字“”，其结果如下：

表 9 jieba 分词工具两个方案的分词

分词	分词结果
方案一分词	遷陵庫/真甲/三百/卌/九甲//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
方案二分词	遷陵/庫/真/甲/三百卌九/甲/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
标准分词	遷陵/庫/真/甲/三百卌九/甲/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

表 10　Hanlp 分词工具两个方案的分词

分词	分词结果
方案一分词	遷陵/庫/真/甲/三百/卌/九/甲/兜/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
方案二分词	遷陵/庫/真/甲/三百卌九/甲兜/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
标准分词	遷陵/庫/真/甲/三百卌九/甲兜/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 9 和表 10 的分词结果可见，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“兜”参与了分词。这说明了“甲兜”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，“甲兜”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，该词被切分为两个单字词是因为 jieba 自带词典中未检测到有“甲兜”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“兜”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“兜”字的 Unicode 编码，故可成功对集外字进行分词。这说明了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，为之后工作提供可行性基础。

第四，方案一与方案二的分词结果略有出入，主要的问题还是在 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，其词库及语料库的建设还未完备，故分词的准确率较低。

5.4 小结

本章主要选择了“中文分词”作为出土文献文本信息处理示例，其主要原因在于分词对于建立出土文献语料库的熟语料有基础作用，并对后续的文本信息处理会产生极大的影响。实验主要分为三步进行，每步均利用了自然语言处理技术中的常用分词方法或者模型，如双向最大匹配算法、HMM 模型，及 jieba、Hanlp 分词工具。

对于实验的训练数据，采用的是小组校对后的里耶秦简词表及分好词的里耶秦简

语料库，其准确性较高；测试数据是随机选择的 10 条包含两类不同方法所造的集外字简文数据，因此，这 10 条数据的选择具有全面性及普适性。

基于以上的实验基础工作，该实验结果具有可分析性及价值。其可分析性具体体现在不同方法与模型对同一简文进行切分时的不同结果，显示出这些模型对于集外字分词处理的不同效果；其价值主要体现在，本文前述所建立的集外字字库及其输入法对于文本信息处理具有适配性及可行性，且对出土文献数据库建设工作有推动作用。

结论

本文首先根据出土文献语料库文本中内含的众多集外字整理并以里耶秦简为例建立了集外字字库，并根据集外字的分类，采用了专用字符编辑程序对方块集外字及Font Creator 程序对临摹集外字的两种方法进行造字处理。其次，利用 Unicode 字符集对造字处理后的集外字进行编码，并根据其字形特征建立输入法码表的检字方案。最后，通过输入法生成器及字库注册表生成出土文献集外字输入法。本研究尝试提出一种适用于出土文献文本的信息处理程序化方法，在实现集外字在语料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，首先采用了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，提高了语料库的完整性；其次，结合四角号码，基于多多输入法生成器生成的集外字输入法，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，本方法不仅可以直接应用于建立简帛文献集外字字库及文本语料库，同时还可以为金文等其他出土文献语料库的建设提供新思路。本方法不仅提高了出土文献数据库中数据的完整性，还可使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。

其次，本研究应用了目前自然语言处理技术中主流的模型及工具对含集外字的文本进行分词实验。实验结果显示其在出土文献的文本信息处理中是有效的且高效的。在将计算机技术应用于出土文献文本的过程中，本研究不仅可完成计算机对出土文献文本语义理解的精确性；同时，也为文本分类及文本挖掘等自然语言处理技术在出土文献文本上的应用提供数据，并提升其准确性，为数字人文视域下的出土文献文本研究提供新视角的基础上，创造新方法以为更多的人文学者提供支持。

参考文献

[1] 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.

[2] 蒋玉斌.谈谈甲骨文摹写中值得注意的一些重要字形信息[J].出土文献,2021(01):27-41+154-155.

[3] 贾常业.西夏文字的字形结构组合形式与造字方法[J].西夏研究,2014(01):25-35.

[4] 尼玛扎西,李志蜀,拥措,群诺,普顿.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.

[5] 周珩珩.藏语言文字信息化处理与未收录字符的动态实现[D].中国人民解放军信息工程大学,2002:5-8.

[6] 张为.汉字专字研究[D].福建师范大学,2017:1-3.

[7] 李丹.异体字字库开发研究[D].华中科技大学,2011:2-6.

[8] 李均明,刘国忠.当代中国简帛学研究 1949-2019[M].北京:中国社会科学出版社,2019:160.

[9] 王灿辉,张敏,马少平.自然语言处理在信息检索中的应用综述[J].中文信息学报,2007(02):35-45.

[10] 王嘉灵.以《汉书》为例的中古汉语自动分词[D].南京师范大学,2014:41-44.

[11] 王大盈.《中国基本古籍库》和《瀚堂典藏》两大古籍数据库比较研究[J].情报杂志,2011,30(S1):157-158+161.

[12] 范子烨.古籍电子化与中国古代文史研究——以文渊阁《四库全书》电子版原文及全文检索版为中心[J].东南大学学报(哲学社会科学版),2004(02):111-114+128.

[13] 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.

[14] 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.

[15] 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.

[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] 刘永革,李强.甲骨文输入法综述[J].殷都学刊,2020,41(3):43-46.

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

[19] 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.

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

[21] 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.

[22] 王正平.基于藏文国际编码字符集的输入法研究[D].南京师范大学,2008:3-5.

[23] GB16959-1997.信息技术信息交换用藏文编码字符集基本集.国家技术监督局,1997,9.

[24] 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.

[25] 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.

[26] 刘志祥,尹奎英,刘晓戎.六码数字甲骨文输入法[P].中国发明专利：CN101702101B,2011.04020.

[27] 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.

[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(4):324-328.

[29] 高玉军,尹江红.藏梵文嵌入式字库、在线输入法在藏文网络信息化中的应用[J].信息技术与标准化,2007(08):20-24.

[30] 仁青诺布,高红梅,王国宏,杨鹏,索朗旺堆.藏文在线输入法的设计与实现[J].西藏大学学报(自然科学版),2013,28(01):65-69.

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

[32] 李运富.楚国简帛文字资料综述[J].古汉语研究,1995(03):45-54.

[33] 刘志基.简析古文字识别研究的几个认识误区[J].语言研究,2019,39(04):89-95.

[34] 刘志基.先秦出土文献字频状况的古文字研究认识价值[J].中国文字研究,2013(02):13-21.

[35] 陈伟主编.里耶秦简牍校(第一卷)[M].武汉:武汉大学出版社,2012:1-10.

[36] 陈伟主编.里耶秦简牍校(第二卷)[M].武汉:武汉大学出版社,2018:1-12.

[37] 许慎.说文解字[M].北京:中华书局,2020(2013 版重印):35,51,72,83,105.

[38] 蓝德康.关于完成国家图书馆古代“地方志”全文数字化一期工程的要点[C]//.第四届中国古籍数字化国际学术研讨会论文集,2013:73-82.

[39] 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.

[40] 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.

[41] 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.

[42] WH/T 91-2020,汉文古籍集外字描述规范[S].文化和旅游部科技教育司,2020,11.

[43] 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.

[44] 廖继莉.中文超大字符集输入法的研究和开发[J].语言研究,2002(S1):291-292.

[45] GB2312-80.中华人民共和国国家标准信息交换用汉字编码字符集基本集[S].北京:语言出版社,1988.

[46] 杨巧.输入法字库切换功能与分级编码研究[D].华中科技大学,2018:36-38.

[47] 史维坤,潘玉坤.近二十年出土文献字词对应考察研究综述[J].东方论坛,2021(03):146-156.

[48] 赵达雄,冯子玲.四角号码的由来及其字频输入码研究[J].图书情报工作,2001(09):48-50.

[49] 陈振文,何锦川.四角号码检字法考略[J].重庆工商大学学报(社会科学版),2014,31(01):130-136.

[50] 乔永.四角号码检字法应该普及——由《辞源》说开去[J].出版科学,2015,23(03):8-13.

[51] 黄天禄.对《四角号码新词典》取号规则的探讨[J].重庆三峡学院学报,2011,27(01):144-146.

[52] 黄天禄.四角号码的取号规则应有统一标准[J].重庆文理学院学报(社会科学版),2012,31(02):54-56.

[53] 常铁英.再论四角号码检字法的传承[J].黔南民族师范学院学报,2011,31(02):91-93.

[54] Chen K J，Liu S H.Word identification for Mandarin Chinese sentences[C].Proceedings of the Fifteenth International Conference on Computational Linguistics,Nantes:COLING-92,1992:101-107.

[55] NADEAUD,SEKINES.A Survey of named entity recognition and classification[J]Lingvisticae Investigationes,2007,30(1):3-26.

[56] Danil M.Bikel, Schwarta R and Weischedel R. An Algorithm that Leams what's in A Name[J]. Machine Leanring Jounral Special Issue on Natural Language Leanring, 1999,34:211-231.

[57] Hideki Isozaki,Hideto Kazawa.Efficient Support Vector Classifiers for Named Entity Recognition[C].Proceedings of the 17th Intenrational Conference on Computational Linguistics,Taipei,Taiwan,2002:390-396.

[58] Rabiner L R.A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition[J].proc.IEEE,1989,77.

[59] Lafferty J D,MccallumA,Pereira F C N.Conditional Random Fields:Probabilistic Models for Segmenting and Labeling Sequence Data[J].Proceedings of Icml,2001:1.

[60] 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.

[61] 武汉大学简帛研究中心.楚地出土战国简册合集(三).北京.文物出版社,2019:47,53,113.

[62] 银雀山汉墓竹简整理小组.银雀山汉墓竹简.北京.文物出版社,2010:59,135.

攻读硕士学位期间取得的科研成果

发表学术论文

[1] 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).

致谢

辛丑立题，壬寅落笔，相识在春，别离在夏。白驹过隙，恍惚间求学二十余载。引用一句去年博士答辩很走心的致谢，“我走了很远的路，吃了很多的苦，才将这份硕士论文送至您面前”，感谢自己求学路上虽然一路艰辛但没有放弃。提起笔，脑中思绪万千；落笔时，只怪自己文采堪忧。

师恩难忘，犹记两年半年前的初春选择了陈镱文教授成为了我的授业恩师，她在我努力求学的路上一直给予我关心和帮助。硕士三年，其中大概有少一半的时间均受疫情所困，陈老师在我选题、开题、撰写小论文及大论文的过程中，无论是节假日还是深夜抑或清晨，无论是文章格式规范还是结构逻辑问题，陈老师都能不厌其烦地教导我、叮嘱我。每次和陈老师讨论完，她总能给我醍醐灌顶的感受。疫情期间，陈老师不顾自己的腿伤，坚持指导我的研究进展，组织同门对我的工作汇报进行讨论，而最终使得我的论文能够顺利完成。谁言寸草心，报得三春晖。

其次，要感谢的是曲安京教授、唐泉教授、袁敏副教授、王昌教授及胡鹏老师对我的谆谆教导与关心。山高水长有时尽，唯我师恩日月长。

父字当头，母字结尾，兜兜转转的二十余载求学之路，你们对我的无私付出是我得坚强后盾，本科毕业后工作两年还能坚持选择求学之路。愿我的父母健康顺遂，喜乐安康。

平生感知己，方寸岂悠悠。感谢我的师兄师弟们和室友们与我共同钻研科研困难，是我科研与生活中不可缺少的一部分。

山水相逢，终于一别，咱们后会有期！