国家出版基金项目
NATIONAL PUBLICATION FOUNDATION

中国文化典籍计算机整理与开发技术研究系列

丛书主编◇侯汉清

GUJI JISUANJI ZIDONG JIAOKAN
ZIDONG BIANZUAN YU
ZIDONG ZHUSHI YANJIU

# 古籍计算机自动校勘、自动编纂与自动注释研究

常　娥◎著

安徽师范大学出版社

中国文化典籍计算机整理与开发技术研究系列
丛书主编◇侯汉清

GUJI JISUANJI ZIDONG JIAOKAN
ZIDONG BIANZUAN YU
ZIDONG ZHUSHI YANJIU

# 古籍计算机自动校勘、自动编纂与自动注释研究

常　娥◎著

安徽师范大学出版社

责任编辑：谢晓博　　责任校对：潘　安
装帧设计：丁奕奕　　责任印制：郭行洲

**图书在版编目（CIP）数据**

古籍计算机自动校勘、自动编纂与自动注释研究/常娥著. —芜湖：安徽师范大学出版社，2013.11
（中国文化典籍计算机整理与开发技术研究系列/侯汉清主编）
ISBN 978-7-5676-1002-6
Ⅰ.①古… Ⅱ.①常… Ⅲ.①计算机应用—古籍整理—研究 Ⅳ.①G256.1-39

中国版本图书馆CIP数据核字（2013）第239150号

古籍计算机自动校勘、自动编纂与自动注释研究
常　娥　著

出版发行：安徽师范大学出版社
芜湖市九华南路189号安徽师范大学花津校区　　邮政编码：241002
网　　址：http://www.ahnupress.com/
发 行 部：0553-3883578　5910327　5910310（传真）　E-mail：asdebsfxb@126.com
经　　销：全国新华书店
印　　刷：安徽芜湖新华印务有限责任公司
版　　次：2013年11月第1版
印　　次：2013年11月第1次印刷
规　　格：700×1000　1/16
印　　张：12.75
字　　数：174千
书　　号：ISBN 978-7-5676-1002-6
定　　价：28.00元

# 出版说明

## 一

中国文化典籍是中华民族在数千年历史发展过程中创造的重要文明成果，蕴含着中华民族特有的精神价值、思维方式和想象力、创造力，是中华文明绵延数千年的历史见证，也是人类文明的瑰宝。对古籍的整理、保护与开发，是中华儿女应尽的义务和职责。

我国古籍资源数字化工作起步于20世纪80年代初期，经过几十年的发展，已取得令人瞩目的成就。第一批《国家珍贵古籍名录》和全国古籍重点保护单位的申报工作早已完成，制定古籍数字化标准列入议程，古籍整理与保护工作进入一个新的历史阶段。

古籍资源数字化最初主要是制作书目数据库，后来发展到古籍全文数据库，直至如今的网络检索系统。信息技术的发展和数字化成果的不断涌现，对古籍数字化提出了更高的要求。专家认为，数字化的古籍资源除了实现文本字符的数字化、具有基于超链接的浏览阅读环境和强大的检索功能外，还需具有“研究支持功能”。所谓“研究支持功能”，是指能够提供有关古籍内容本身科学、准确的统计与计量信息，提供与古籍内容相关的参考数据、辅助工具。这些信息、数据或工具都是古籍内容的增值或补充。北京大学计算语言研究所和古文献研究所合作开发了“古诗研究计算机支持系

统”，并取得了阶段性成果。

时值古籍数字化研究日新月异、如火如荼之际，安徽师范大学出版社于2011年精心策划、2012年成功申报、2013年落实出版国家出版基金项目“中国文化典籍计算机整理与开发技术研究”（编号：2013G2－011），在数字化古籍诸项功能特别是“研究支持功能”上给予探索。

改革开放30多年来，国泰民安、政通人和，中国传统文化日益受到政府重视，有关科研机构加大了对古籍整理研究的力度。安徽师范大学出版社能够有机会申请到国家出版基金项目的资助，本项目丛书能顺利进行，实在与国家关注出版事业、关注中国传统文化、关注文化典籍计算机整理工作密切相关。

## 二

“中国文化典籍计算机整理与开发技术研究”项目主要内容如下：

第一，探索与试验古籍知识库、模式库，将之改造为规则库。

本项目利用命名实体识别、词汇同义词关系的识别、文本主题概念的提取等技术，从各类古籍数据库抽取人名、地名、文献名、职官名、物品名、年号等；并将人名表、地名表、书名表、年代年号表等，与引书模式、异名别称模式、断句模式、分类模式等模式库整合成一个古籍整理与开发专用的知识库，以方便中文古籍整理与开发。

本项目构建的各类知识库，具体有：古代官名、人名和地名表；避讳字、异体字和繁简字对照表；常用古籍名称库；专业术语词典，按专业分为历史、天文、农业、医学、宗教等多个专业词典；主题术语词典，按主题分为动物、植物、矿物等若干主题词

典；古代关联词语表，用语义相似度计算和基于词典释义的同义词识别算法，开发古代关联词语表；禁用词典。

本项目构建的各类模式库有：异名别称模式库，包括别称词、避忌特称、地域特称、文献特称等；断句标点模式库，包括句法特征词法、同义语标志词法、反义复合词、引书标志、时序、数量词、重叠字词、动名结构及比较句法等多种模式识别库；古籍分词模式库。大多数古籍文本无标点，分词的长度及方法需要单独构建。

这些知识库与模式库，采用拿来主义，并经过计算机检验与筛选，最终形成适用于计算机处理古籍的规则，合成为一个综合的规则库，从而为计算机处理古籍提供有力的规则支撑。

第二，重点探索与试验下列古籍智能整理与开发的关键技术。

自动校勘技术：采用对校法，借鉴中文文本自动校对和模式匹配技术，通过比对程序校勘古籍。

自动断句标点：对现有部分标点本古籍进行数理统计，归纳、总结其断句和标点模式。同时结合语言学方法，进一步优化断句和标点模式，从而实现计算机辅助断句与标点。

自动分词和标引：利用汉语现代文本的分词理论和方法，探索古籍文本的自动分词技术，并利用统计学方法（N-gram 等），从古籍数据库中筛选出有一定表达意义的实词词汇。同时利用异名别称模式，创建并完善古籍用词同义词典。在此基础上，引入文本数据挖掘、主题提取和自动分类技术，探索基于知识库的古籍文本的自动标引与分类。

自动编纂：让计算机模拟人脑从大量古籍文本中判断、选择出与编纂主题相关的资料，实现古籍专题资料的自动编纂工作。

自动注释：收集已有古籍专业词汇及其注解，构建古籍语词注解知识库。

第三，在上述基础上，将它们整合为计算机整理与开发古籍的

“一条龙”服务，即构建出古籍整理与开发的专家系统或智能处理系统。

将以上各种词汇、知识、模式整合起来，构建成一个内容丰富、功能多样的古籍规则库，再与自动校勘、自动断句标点、自动分词标引、自动编纂、自动注释等各项技术结合，从而实现文化典籍整理与开发的“一条龙”服务，提出并设计一种集成各种古籍整理与开发智能技术的原型系统。该系统集知识与模式于一身，集规则与技术于一体，具有合成性，既适用于古籍数据库的建设，又适用于古籍数据库的开发使用。

第四，在上述基础上，本研究进行四项个案研究，在实践中探索上述集成的古籍整理与开发智能技术原型系统的可行性与应用性。

农业历史文献数字化：构建农史文献资源库，对农史文献进行自动标引和自动分类，提供农史文献的浏览与检索服务。

建立农史文献门户：构建农史门户网页智能搜索引擎和农史网页自动标引与自动分类实验系统，构建农史门户实验网站。

探索民国农业文献自动索引：在民国农业文献数字化整理中的具体应用，研究索引自动编纂、电子图书编纂、电子索引编纂、数据库建设和主题网关构建等技术方法。

地方志中农业资料的挖掘：从《方志物产·广东》中选取比较实用的全文数据库、物产索引、引书索引、物产分析和引书分析等几个方面进行研究。

总之，本项目充分利用目前在现代汉语文本已经取得成功的中文信息处理技术成果，并根据此成果中的模式识别技术、聚类技术、信息自动提取、信息检索及其他自然语言处理技术等，对照现已建成的大量数字化文化典籍数据库，归纳并修订各类知识库与模式库，研究古籍的自动校勘、自动断句标点、自动分词标引、自动

编纂、自动注释等技术，合成古籍整理与开发的专家系统或智能处理系统，从而为大规模建设新的更多古籍数据库作准备。

## 三

本项目成果的推广和运用，不但对于探索数字时代古籍文本自然语言处理的理论和方法具有一定意义，而且对推动古籍整理和研究的自动化和智能化、促进我国文化典籍资源的建设和开发以及弘扬传统文化等方面，均具有重大的现实意义和很高的应用价值，可以为继承与发扬中华古籍文化、为建设中国特色社会主义文化服务。

本项目丛书主编由南京农业大学信息科技学院博士生导师侯汉清教授担任。侯先生是中国古籍整理专业第一个硕士研究生，早年在北京大学任教，现执教于南京农业大学，系中国古籍整理专家、中国索引学会副理事长。中图分类法就是侯先生主创起来的。2008年，侯先生主持国家社会科学基金重点项目“文化典籍整理与开发智能技术研究”（编号：08ATQ002），本套丛书即此项目的纸质成果。

本丛书分为六册，各册的内容及其撰写者简要介绍如下：

《古籍计算机自动断句标点与自动分词标引研究》，侧重于自动断句标点、自动分词标引研究，兼顾古籍计算机整理与开发系统的构建与集成。作者黄建年，博士，研究馆员，现就职于南京财经大学。

《古籍计算机自动校勘、自动编纂与自动注释研究》，侧重于自动校勘、自动编纂与自动注释研究，兼顾古籍计算机整理与开发系统的构建与集成。作者常娥，博士，现就职于东南大学，硕士生导师。

《古籍计算机自动索引研究——以民国农业文献自动索引为例》，侧重于自动索引研究，并以民国农业文献自动索引为样本。作者王雅戈，博士、博士后，中国索引学会理事，现就职于常熟理工学院。

《古籍计算机全文数据库及内容挖掘研究——以〈方志物产·广东〉为例》，侧重于数据库内容挖掘研究，并以《方志物产·广东》之物产、引书等内容挖掘研究为样本。作者衡中青，博士，中国索引学会理事，现就职于佛山科学技术学院。

《古籍计算机信息门户自动构建与应用——以农史学科为例》，侧重于信息门户自动构建与应用，并以农史学科信息门户构建与应用为样本。作者刘竟，博士，现就职于江苏大学。

《农业历史文献数字化建设研究》，侧重于农史文献数字化实践——中国农业遗产信息平台建设，并介绍其实际应用。作者曹玲、薛春香，均为博士，分别就职于南京信息工程大学、南京理工大学。

本项目丛书的出版发行，可为正在有志于从事本领域研究和工作的人员提供一个可资借鉴的文本。我们期待本丛书能为中国从文化古国向文化大国、文化强国迈进尽绵薄之力。

# 目　录

# 1 绪 论

## 1.1 古籍校勘概述

### 1.1.1 古籍校勘的意义

校勘，是指利用古籍不同的版本和其他补充资料，通过比较核对和分析推理，发现并纠正古籍在流传过程中发生的文字错误[1]。古籍流传至今，绝大多数已不是原稿本、原抄本或原刻本，而是经过多次传抄翻刻之后的抄刻本。古籍在抄刻的过程中，由于操作人员的主观原因或者技术、时代文字差异等客观原因，往往会出现各种错误，一本古籍流传数百年甚至数千年之后，可能产生多个版本，各个版本之间的内容差异很大，这种现象屡见不鲜。在古籍的整理过程中，必须对古籍进行一次或多次校勘，以弥补这种由于版本之间的差异而引起的对读者的误导。古籍校勘的意义主要体现在以下三个方面：

第一，校勘是古籍整理的基础。古籍整理主要包括校勘、标点、注释、翻译、影印、汇编、辑佚、编制目录和索引等各项工作。校勘工作是古籍整理其他工作的基础，这是因为上述其他各项工作开展之前都需要有一个正确的底本。如果底本有误，其他工作可能出现相应的错误，或令人难以理解原文，或导致理解错误。因

此，必须先对古籍进行校勘，恢复原本的面貌，再通过其他方法进行整理，充分发挥古籍的利用价值，更好地为社会服务。

第二，校勘是古典文献研究的前提。要做好古典文献研究工作，也必须有一个文字正确无误的底本，如果底本有很多错误，那么针对某种古籍的文献研究也会出现相应的错误。只有通过校勘，得到比较正确的底本，相关的古典文献研究也才能得出正确的结论，取得令人满意的结果。

第三，校勘是阅读古籍的先导。校勘是古籍整理的基础，古典文献研究的前提是古籍阅读的先导。阅读古籍，参考古书，欣赏古代作品，必须选择文字正确无误的底本，才能获取较为真实的知识。

### 1.1.2 古籍校勘的对象

古籍在流传过程中所发生的错误，在形式上多表现为误、脱、衍、倒等文字差异，在内容上则表现为理解分歧。一般而言，校勘的对象主要针对各版本之间文字的差异，也就是指出和改正古籍在流传中发生的各种文字错误，以恢复古籍的原貌，而对于原书内容上的错误，一般不作为校勘的主要对象[2]。

误字，是指古籍在流传过程中出现的错字，亦称“误文”或“讹文”。例如《齐民要术》：“凡秋收之后，牛力弱，未及即秋耕者，谷、黍、穄、粱、秫、芟方末反之下，即移羸速锋之，地也恒润泽而不坚硬。”，其中“芟方末反”明抄本中作“芟古末反”，湖湘本中作“芟古未反”，均为误文。

倒错，是指古籍原文位置的颠倒错乱。例如《孟子》：“赵岐注曰：‘言仕之为急若农夫不可不耕。’”明抄本《齐民要术》中引作：“赵岐注曰：‘言仕之为急若农夫不耕不可。’”此处“不耕不可”、“不可不耕”视作位置颠倒错乱。

异文，是指古籍在流传过程中出现的各种版本之间的文字差异，由于时代变迁、字体演变、书写形式不同造成的古今字、异体字、繁简字、通假字和避讳字。例如，渐西本《齐民要术》：“高诱注曰：‘菖，菖蒲，水草也。’”清抄本《齐民要术》中作：“高诱注曰：‘昌，昌蒲，水草也。’”此处“菖”和“昌”互为异文。

脱文，是指古籍在其流传过程中原文脱落遗漏了的文字，又称“夺文”。从脱文的形式上看，有脱一字、数字，有脱一句、数句，更有脱一行、数行等。从脱文的原因上看，有抄脱和意删等。例如《吕氏春秋》曰：“冬至后五旬七日菖始生。菖者，百草之先生者也，于是始耕。”渐西本《齐民要术》中引作：“冬至后五旬七日菖生。菖者，百草之先生也，于是始耕。”可见“菖生”作“菖始生”，“先生也”作“先生者也”。《齐民要术》在引用过程中脱漏了两处文字，分别是“始”和“者”。

衍文，是指古籍在流传过程中比原文多出的文字。从衍文的形式看，有衍字、衍句、衍行、衍页等，大抵与脱文一致，兹不详述。导致衍文的原因，则有无意抄刻致衍及有意妄加致衍等情况。例如《说文解字》曰：“黍，以大暑而种，故谓之黍。”段玉裁注：“大，衍字也。”古书多说夏至种黍，大暑太晚，应是衍字。又如《齐民要术》：“崔寔曰：‘六月，大暑中伏后，可收芥子。七月、八月，可种芥。’”而《四民月令》无“伏”字。缪启愉先生认为大暑正属于“中气”，故称“大暑中”，再拖个“伏”字，没有意义，故为衍文。

### 1.1.3 古籍校勘的方法

古籍校勘的方法包括对校法、本校法、他校法、理校法和综合校法。

第一，对校法，是指选定底本，用不同的版本相互校勘的方

法。对校的前提，在于广泛收集现存的各种版本，然后分析版本的源流，鉴别版本的优劣，确定一个合用的底本，再用其他异本逐页、逐行、逐字地同它对校，凡有不同之处，一一记录在“校勘表”中。该方法最简便、最稳当，也是校勘工作最基本的方法。其主旨是校版本之间的异同，不校内容上的是非，故其缺点在于，虽祖本或别本有讹，亦照样录之，其长处在于，不参己见，避免了主观臆断、妄改旧文的弊病，而且校者若将所有的异文汇集在一处，编为校勘记，则读者手执此编，就等于掌握了许多版本，如阮元的《十三经注疏校勘记》即是这种情况。

第二，本校法，是指以本书的体例、音韵、文字、语法等为依据，在没有版本异文和其他有关材料的情况下进行校勘。主要通过比较此书的前后文字来判断正误，包括目录和正文标题之间的相互校勘，前后文句之间的相互校勘，正文和注文之间的相互校勘，以及用文义、文例进行校勘，用文辞、音韵的对应进行校勘等。这种方法的优点是：在无对校本和他校本可据的情况下，仍能校正书中的文字错误；缺点是：该校法具有一定的局限性，对于校出的误、脱、衍等异文，除可确认抄刻致误者外，一般不宜改动原文，只能写入校勘记中。

第三，他校法，是指以他书校本书。因为各种农业古籍都会引用许多古书，可以依据引文校所引之书。如《齐民要术》中常引《尔雅》、《广志》、《四民月令》、《说文解字》、《释名》等书中内容，可以根据《尔雅》、《广志》等古籍中的文字来校对《齐民要术》中的引用是否准确。但由于一书在撰写中常常引用大量的他书，又常常被后世著作所引用，所以他校法具有“范围最广，用力最劳”的特点，这成为他校法的一个缺点。

第四，理校法，是指按照理论推测正误的校勘方法，是校勘工作的补充方法。校者发现古书中的确存在错误，可是又没有足够的