

陈小荷 冯敏萱 徐润华 等著

先秦文献信息处理

计算语言学研究系列

陈小荷 主编

世界图书出版公司

先秦文献信息处理

陈小荷 冯敏萱 徐润华 等著

世界图书出版公司

北京·广州·上海·西安

图书在版编目 (CIP) 数据

先秦文献信息处理/陈小荷, 冯敏萱, 徐润华等著. —北京: 世界图书出版公司北京公司, 2012. 12

ISBN 978-7-5100-5661-1

I. ①先… II. ①陈…②冯…③徐… III. ①古籍—信息处理—中国—先秦时代 IV. ①G256. 1-39

中国版本图书馆 CIP 数据核字 (2013) 第 005489 号

先秦文献信息处理

著 者: 陈小荷 冯敏萱 徐润华 等

责任编辑: 陈晓辉

出 版: 世界图书出版公司北京公司

出 版 人: 张跃明

发 行: 世界图书出版公司北京公司

(地址: 北京朝内大街 137 号 邮编: 100010 电话: 64077922)

销 售: 各地新华书店和外文书店

刷: 三河市国英印务有限公司

开 本: 711 mm × 1245 mm 1/24

印 张: 14.5

字 数: 315 千

版 次: 2013 年 1 月第 1 版 2013 年 1 月第 1 次印刷

ISBN 978-7-5100-5661-1

定 价: 35.00 元

版权所有 翻印必究

《语言科技文库》总序

李葆嘉

当代语言学已经进入了一个科学与技术的互补时代，信息处理水平成为衡量国家现代化水平的重要标志之一。知识世界的载体是语符系统，信息处理的根本对象是语言信息处理。与计算机的出现使得语言符号有可能成为数据处理对象相似，神经科学实验仪器设备的应用，使得在大脑神经层面探讨语言机制成为可能。这些无疑都引导语言研究走向科技化，“语言科技新思维”（李葆嘉 2001）应运而生。所谓“语言科学”包括理论语言学、描写语言学、历史语言学、应用语言学等分支学科，所谓“语言技术”指语言研究的现代技术手段，包括语言信息处理、语音实验分析，以及语言的神经、心理和行为实验分析的技术手段等。就语言信息处理而言，又可以分为语料库研制技术、知识库研制技术、知识挖掘和抽取技术、句法信息处理技术、词汇信息处理技术、语音信息处理技术、语义信息处理技术、语用信息处理技术等。

2001 年 5 月，南京师范大学文学院创办了史无前例的“语言科学及技术系”，率先迈出了从传统文科教育范型向现代科技教育范型转变的步伐。“十五”期间，南京师大“211 工程”重点学科建设项目“语言信息处理与分领域语言研究的现代化”（陈小荷教授主持），以基础平台建设、资源建设和理论探索等为主，迈出了语言科技研究的一大步。

“十一五”期间，南京师大文学院、外国语学院和国际文化教育学院联袂申报“211 工程”三期重点学科建设项目。该项目以“语言科技”为引导，以“多学科交叉、跨院系整合、开放型营运”为理念，建设具有前瞻性、原创性、成长性的语言科技高级工作平台。以典型课题的工作原理为核心，进行资源开发和系统研制，拓展语

音科技、二语习得的神经机制研究、言语能力受损儿童的语言能力研究等新方向。同时造就新一代学术领军人物和培养一批高层次复合型人才，以期形成一支高水平的交叉学科团队。该项目设计，体现了工作平台建设、理论创新、应用研究、人才培养、团队建设的学科发展一体化思路。其旨趣在于，加速语言研究从传统文科范型向现代科技范型的转变，以引领 21 世纪语言科技的新潮流。

作为新兴交叉学科项目，通过教育部组织的专家匿名评审，“语言科技创新及工作平台建设”（2008～2011）获批，总投入 1000 万元。总体而言，这一“语言科技创新”团队，分支学科齐全，专业知识互补。涵盖了理论语言学、计算语言学、语义科技、语音科技、实验方言学、历史语言学、神经语言学、二语习得研究、话语行为语言学等领域。这一期间，项目组成员获批的国家级基金项目达 20 多项。该项目理念之前瞻、实力之雄厚、工程之浩大、经费之保障，为学界瞩目。

2008 年秋，本项目以南京师范大学语言科技研究所为实施单位正式启动。主要有三大任务：建设一个领先性的语言信息科技实验室、建立一个独创性的语言科技工作平台、撰著一套有特色的语言科技文库。

从实验室方案设计到设备招标采购，再到实验室用房改造，经过 8 个月的努力，2009 年 12 月，语言信息科技实验室建成，为语言研究从传统范型向科技范型的转变提供了基本保障。该实验室划分为实验工作区、科研工作区和管理服务区。实验工作区建有语音实验与计算室、神经认知实验与计算室、课堂话语实录室三个专门实验室。科研工作区建有语义科技工作室、语音科技工作室、方言实验工作室、知识工程工作室 I（先秦词汇）、知识工程工作室 II（中古词汇）、知识工程工作室 III（敦煌俗语言文字）、语言习得神经机制工作室、语言习得中介机制工作室，以及参研工作室。管理区服务包括办公室、管理室、编辑室和交流室。出席“语言科技高层论坛暨语言信息科技实验室落成仪式”（2009 年 12 月 14 日）的专家认为，该实验室体现了语言学跨学科研究的当代性和先进性，具有整体性、科技型、开放型三个特点，处于全国领先地位，是“语言科技新思维”的又一体现。同时认为，该实验室的科研工作涵

盖了四个二级学科、四个博士学位点，有稳定明确的研究方向，有合理的设计规划和很好的科研基础；整体设计合理，功能齐备。以教育部重点实验室建设标准衡量，很多方面超过了指标。

语言科技工作平台是基于工作原理（课题定位—理论方法—技术路线—关键技术—评估方式）而建设的高级平台。一方面，从语言信息、语言知识和语言机制三个层面，围绕典型课题进行设备配置、资源建设和软件开发；一方面，将典型课题研究与工作平台建设融为一体，依据典型课题建设的子平台应具有解决同类课题的功能。

建设语言科技工作平台的目标是要实现语言研究手段的技术化和模型化，总体设计包括三个二级平台和八个子系统。

一、语言信息工作平台 1. 语义科技工作系统（李葆嘉教授主持）：基于词汇语义—句法语义的一体化研究思路，开发“人—机交互语义标注工具”，研制“深度语义标注信息库”；研制“幼儿（2~6）日常话语跟踪语料库”，完成幼儿语义系统和话语行为分析研究。2. 语音科技工作系统（顾文涛教授主持）：研制“多语言、多语境、多语用的语音语料库”，基于声学信号分析、感知实验和数学建模，完善语音韵律理论与相关技术应用。3. 方言实验工作系统（刘俐李教授主持）：完成“网络版汉语方言有声语料库”，拟定系统的可操作性语音、词汇、语法实验模型和研究方法，进一步完善新兴交叉学科“实验方言学”。

二、语言知识工作平台 1. 先秦词汇统计与知识检索系统（陈小荷教授主持）：研制“先秦文献语料库”、“专名知识库”、“汉语词汇档案库”等，开发先秦文献自动分词算法、古籍版本异文自动发现算法、同指专名检索软件工具等，完成“先秦汉语词汇统计与知识检索”。2. 中古词汇统计与知识检索系统（董志翘教授主持）：研制“中古文献语料库”、“专名知识库”、“中古汉语词汇档案库”等，开发中古文献自动分词和标注工具等，完成“中古汉语词汇统计与知识检索”。3. 敦煌俗语言文字统计与检索系统（黄征教授主持）：研制“敦煌文献资料库”、“敦煌文献俗语词语档案库”，开发相应工具，完成“敦煌文献资料与知识检索”。

三、语言机制工作平台 1. 二语习得的神经机制研究系统（倪

传斌教授主持):研制“英语受蚀词汇库”等,基于行为学、脑成像和脑电三维度模型,进行中国人英语习得与磨蚀的神经机制研究,完成“基于神经机制的英语个性化学习分析系统”。2. 二语习得的中介机制研究系统(肖奚强教授主持):研制“留学生汉语口语中介语语料库”,基于中介语理论、对比分析理论、偏误分析理论以及二语习得影响因素等,完成“留学生汉语习得的中介机制研究”。

这一工作平台,既是科技研究平台,也是人才培养平台,即一个现代化的科学研究和人才培养工作体系。

作为本项目的文本成果,《语言科技文库》包括计算语言学研究、语义语法学研究、汉语方言学研究、古代汉语学研究、语言教学与研究、语言新专题研究六个系列。其总体特征为:领域的开拓性、理论的原创性、选题的新颖性、方法的交叉性、考据的精审性、成果的应用性。在研究过程中,除了数据采集分析、资源建设和软件开发,更重要的还是要有新思路、新理论和新材料。陈小荷提出的先秦文献信息处理新方法,从先秦典籍注疏文献中挖掘出用于自动分词和词义消歧的知识,再注入已开发的古汉语分词和词性标注工具中去,所取得的先秦古籍版本异文自动发现、先秦词汇知识自动挖掘等成果均具开拓性。李葆嘉提出的语义语法学理论和话语行为理论,基于研制专用语料库或语义信息库和技术手段,开拓了语义网络建构、深度语义分析和话语行为研究等新的领域。刘俐李建构的实验方言学理论和方法,为方言学向现代科技方法的转型研究提供了新路,并取得了一系列新成果。黄征多年来从事敦煌文献及其俗词语文字研究,古代汉语学研究系列中的敦煌文献校录整理,以及敦煌写本字词考释、以古佚和疑伪经为中心的敦煌佛典词语和俗字研究、两汉声母系统研究等新见迭出。肖奚强基于汉语中介语语料库的二语习得研究,在对外汉语教学研究界已经产生了影响。钱玉莲的汉语介词与相应英语形式比较研究等专著各有亮色。倪传斌依据语言测试和认知实验等数据,从行为学、生理学和语言学三个层面分析影响中国英语学习者外语磨蚀的相关因素。刘宇红基于隐喻的理论探讨,对各类隐喻形式的结构、特性和解读规律进行了多视角的深入探讨。

《语言科技文库》所收论著,由作者在2008年12月申报选题,

2011 年始逐步完稿。系列主编审读了书稿，主要就其学术价值、章节安排、内容关联、行文表述、图表绘制等方面，提出审阅意见。此后，作者们对书稿又进行了修改和润色。《语言科技文库》的作者，大多数是具有博士学位的年轻教师。对于我们这些 20 世纪 80 年代走进语言学研究领域的而言，出版论著可能已不足为道。然而，对于年轻学者而言，其论著的出版既是几年来研究的结晶，也是对其继续探索的促进。换言之，“211 工程”重点学科建设的目的之一，就是为年轻教师搭建一个可持续发展的科研和教学平台。学科带头人的主要任务之一就是提携后进。

尽管从根本上来说，科学或学术研究是一种个人的探索行为，然而复杂问题的研究，无疑需要群体协作。“学科建设”或团队合作模式，是 20 世纪 90 年代后期出现的一个新概念。这种模式涉及总体规划、多方协调，是需要付出精力和心血的。2008 年，通过投票方式推举我担任该项目总负责时，就意识到自己成了一个“劳动班委”。2009 年，前往安徽大学拜访黄德宽教授时，曾谈到“学科负责人的任务就是规划设计，争取项目经费和提供科研设备设施”，得到黄教授的赞许。2010 年，申报江苏省高校哲学社会科学重点研究基地时，评审专家柳士镇教授提问的“作为一个交叉学科项目，各学科之间的协调是怎么考虑的，有什么做法”，可谓一语中的。作为后学，深知交叉研究之艰、学科整合之难。相关学科之间的整合协调需要借助行政机制，但凭借行政方式并非就能完成。当时的回答是，目前做到的是建成了一个可以合作研究的场所，至于学科之间的进一步沟通合作应有较长过程。有一点很明确，只有通过交叉项目，相应学科才能渗透，合作者才能逐步磨合。我们只是在一步步探索。

十一五期间的“211 工程”建设项目即将完成，但是学科建设的任务并没有结束。2010 年，“语言信息科技研究中心”被评审为江苏省高等学校哲学社会科学重点研究基地，为“语言科技”这一交叉领域注入了新的建设活力。重点研究基地建设，除了“跨院系整合、多学科交叉、开放型运行”理念，需要凸显“合作性攻关”。围绕交叉性项目，实施计算语言学、语音科技、神经语言学、语义科技等力量的联合攻关计划。只有通过全面开放以及和与国内外同

行的合作交流，才有望建成具有影响的语言科技研究、人才培养和学术交流基地。

十年前，我（2001）曾写道：“语言科技”的内涵是以理论研究为指导，以描写研究为基础，以应用研究为枢纽，促使语言研究向计算机应用、认知科学和现代教育技术领域等延伸，沟通文理工相关学科以实现语言研究过程及其成果的技术化。“语言科技”的外延为语言工程科技、语言教育科技和语言研究科技。其中，“语言研究科技”是将语言研究活动与资源建设、软件开发相结合，其目标是实现语言学自身的科技化。还应包含语言实验、数据处理这些实验语音学、神经语言学研究的科技手段。

虽然语言学家不可能也不必要都转向语言计算或实验研究，尽管描写、考据和内省始终是最基本的方法，但是具有一定的语言科技意识却非常必要。语言学家只有了解有哪些可供利用的资源、软件或仪器，才能提高其研究深度、精度和效率。语言学家也只有了解到信息处理的语言研究需求，才有可能为之提供可资应用或参考的基础成果。“语言科技”是21世纪语言学研究的潮流。

此为出版缘起。是为总序。

2011年8月谨识于南都

目 录

第一章 绪 论	1
第一节 先秦文献	1
第二节 先秦文献信息处理	4
第三节 先秦文献信息处理的特点	9
第四节 基于注疏文献的处理方法	10
第二章 注疏文献处理	18
第一节 注疏文献的选取	18
第二节 注疏文献的内容分析	27
第三节 先秦文献与其注疏文献的句子对齐和注释对齐	36
第三章 词语切分	56
第一节 先秦汉语分词规范	57
第二节 基于统计模型的自动分词方法	69
第三节 利用注疏文献的自动分词方法	81
第四章 词性标注	92
第一节 词类体系	93
第二节 基于统计的分词标注一体化方法	98
第三节 基于注疏文献的词性校正	107
第五章 古今字、通假字资源建设	126
第一节 古今字、通假字研究现状	126
第二节 古今字、通假字数据库的构建	130
第三节 通假字标注实验	136
第四节 实验结果及其分析	137
第六章 词汇概貌	146
第一节 语料来源和说明	146
第二节 先秦文献的汉字统计	148
第三节 先秦文献的词汇统计	154

第四节	计算每种文献的特色词	161
第五节	成语来源统计	166
第七章	词义消歧	169
第一节	古代汉语词义特点和消歧难点	170
第二节	义项区分的颗粒度	171
第三节	基于 CRF 模型的消歧实验	177
第四节	分类器集成的消歧实验	191
第八章	专名标注	216
第一节	关于“专名”和“命名实体”的辨析	217
第二节	先秦时期的人名与地名	218
第三节	基于统计的专名识别	224
第四节	基于注疏文献的专名识别	231
第五节	人名所指歧义消解	242
第六节	基于专名标注的事件划分	254
第九章	版本异文发现	261
第一节	异文和版本异文	261
第二节	个案和研究方法	263
第三节	基于相似度计算的句珠配对	265
第四节	基于同文排除的异文配对	269
第五节	基于双序列比对的一体化处理	272
第六节	实验结果及其分析	275
第十章	《左传》中的春秋社会网络分析	283
第一节	社会网络分析的基本方法和软件	284
第二节	社会关系数据的获取和网络的定义	286
第三节	人物-事件网络分析	291
第四节	人物关系网络分析	296
第五节	人物关系网络的深入分析	303
后 记		321

第一章 绪 论

第一节 先秦文献

先秦是指秦朝统一中国之前的历史阶段，可以上溯至三皇五帝乃至更远。不过，就先秦文献的产生年代而言，则基本上是商周两代，且以东周（春秋战国）时期为主。现在所能见到的先秦文献可分为两类：先秦传世文献与先秦出土文献。前者经历过较长历史时期的流传，后者则是现当代发掘出土、尚未广泛流传的文献。

裘燮君介绍了几种重要的先秦传世文献的断代研究。^{[1].[2]}关于这些文献的成书年代，较为流行的说法是：

《周易》，西周末年。

《左传》、《论语》，战国时期。

《孟子》，战国中期，孟子卒年之前。

《荀子》，战国后期。

此外，《诗经》风、雅、颂产生有早有晚，同为国风，各篇年代也不能一概而论，许多诗篇的产生年代甚至不可考。

一、先秦传世文献

据李零^[3]介绍，先秦传世文献共 60 种，其中经传类 11 种，小学类 2 种，史书类 6 种，子书类 26 种，诗赋类 1 种，兵书类 5 种，数术类 4 种，方技类 5 种。

表 1-1 先秦传世文献目录

文献	分类	文献	分类	文献	分类	文献	分类
诗经	经传	战国策	史书	鹖冠子	子书	楚辞	诗赋
尚书	经传	穆天子传	史书	慎子	子书	司马法	兵书

续表

文献	分类	文献	分类	文献	分类	文献	分类
仪礼	经传	竹书纪年	史书	中子	子书	六韬	兵书
礼记	经传	世本	史书	商君书	子书	孙子	兵书
大戴礼记	经传	论语	子书	韩非子	子书	吴子	兵书
周礼	经传	曾子	子书	邓析子	子书	尉繚子	兵书
周易	经传	子思子	子书	尹文子	子书	甘石星经	数术
春秋	经传	孝经	子书	公孙龙子	子书	连山	数术
左传	经传	孟子	子书	鬼谷子	子书	归藏	数术
公羊传	经传	荀子	子书	尸子	子书	山海经	数术
穀梁传	经传	墨子	子书	吕氏春秋	子书	黄帝内经太素	方技
尔雅	小学	老子	子书	燕丹子	子书	黄帝内经素问	方技
史籀篇	小学	庄子	子书	鬻子	子书	黄帝内经灵枢	方技
逸周书	史书	文子	子书	管子	子书	黄帝八十一难经	方技
国语	史书	列子	子书	晏子春秋	子书	黄帝甲乙经	方技

2008 年起,我们参加了南京师范大学 211 工程三期重点学科建设项目“语言科技创新及工作平台建设”,承担子课题“先秦词汇统计与知识检索系统”的研究任务,选择了 25 种比较重要和可靠的先秦传世文献作为处理对象:

表 1-2 25 种先秦传世文献的篇幅(万字)

文献	分类	篇幅	文献	分类	篇幅	文献	分类	篇幅
尚书	历史	2.81	论语	政论	1.59	吕氏春秋	政论	10.15
左传	历史	17.98	孟子	政论	3.54	晏子春秋	政论	4.08
国语	历史	7.24	荀子	政论	7.26	诗经	诗歌	3.10
仪礼	典章制度	7.13	墨子	政论	7.94	楚辞	诗歌	1.53
礼记	典章制度	9.80	老子	政论	0.59	孙子兵法	军事	0.67
周礼	典章制度	4.09	庄子	政论	6.48	吴子	军事	0.47
孝经	典章制度	0.18	商君书	政论	2.00	周易	其他	2.12
公羊传	语言文字	4.44	韩非子	政论	10.58			
穀梁传	语言文字	4.09	管子	政论	12.79			

这 25 种先秦传世文献合计 1 334 780 字（纯汉字）。

二、先秦出土文献

先秦出土文献按出土时间主要有以下几个批次：

银雀山汉简：1972 年于山东临沂银雀山两座汉墓中发掘出土。共计有完整简、残简 4 942 枚，此外还有数千残片。其内容包括《孙子兵法》、《孙臆兵法》、《六韬》、《尉繚子》、《晏子》、《守法守令十三篇》、《元光元年历谱》等先秦古籍及古佚书。这些古籍均为西汉时手书，是较早的写本。^{[4],[5]}

定州汉简：1973 年河北省文物管理处和定县博物馆从定县（今定州市）八角廊 40 号西汉墓中发掘出土的竹简。其中先秦文献有《论语》、《儒家者言》、《哀公问五义》、《保傅传》、《六韬》、《文子》等。^[6]

马王堆帛书：1973 年长沙马王堆三号墓出土的帛书，共 28 种，计 12 万余字。其中先秦文献有《周易》、《老子》等。帛书中还有一部《战国纵横家书》（整理者定名），共 27 章，1.1 万余字，其中计有 16 章不见于《战国策》和《史记》。^{[7],[8]}

阜阳汉简：1977 年安徽省文物工作队等单位在安徽阜阳发掘出土的西汉前期汝阴侯墓简牍。内容包括《仓颉篇》、《诗经》、《周易》、《庄子》、《吕氏春秋》等，各种文献所得残片数量不一。^{[9],[10]}

郭店楚简：1993 年 10 月，在湖北省荆门市郭店村，郭店一号楚墓发掘出竹简，共 804 枚，为竹质墨迹。其中有字简 730 枚，共计 13 000 多个汉字。郭店楚简包括 16 篇先秦时期的文献，大多为两千多年前的先秦佚籍。^[11]

上博楚简：1994 年前后湖北江陵地区楚墓中被盗掘出的战国楚简，流转于香港文物市场，后为上海博物馆购回。这批楚简有 1 200 余枚，3.5 多字，涉及先秦古籍 80 多种（一说 100 余种），其中大多数是先秦佚籍。^[12]

清华简：清华大学于 2008 年 7 月收藏的一批战国竹简，共 2 388 枚，由清华校友赵伟国捐赠。经碳 14 测定证实，这批竹简的年代是战国中晚期，从文字风格来看主要是楚简。竹简的内容多与历史有关，其中包含一部失传两千多年、记录在 138 枚竹简上的编年体

史书。^[13,14]

出土文献的可贵之处在于：第一，出土文献未经流传，因此而避免了流传过程中的改写，能保留原貌；第二，可以用现代科技手段鉴定其成书年代；第三，出土文献常常伴有其他出土文物，可以相互印证其共生关系。出土文献中，有的是先秦佚籍，其文献价值自不待言，有的是可与传世文献对照的更可靠的版本。有学者称出土文献重写先秦文学史^[15]、重新建构先秦哲学的脉络^[16]。

第二节 先秦文献信息处理

中文信息处理是一门有三十多年发展历史的综合性学科，是语言信息处理的一个分支^①。语言信息处理是用计算机对自然语言在各个层次（语素、词、短语、句子、段落、篇章）上的各种表现形式（图像、声音、文本）的信息进行处理：输入、输出、转换、存储、压缩、检索、抽取和提炼。

先秦文献信息处理应该是中文信息处理的一个分支。三十多年以来，中文信息处理的研究对象实际上主要是现代汉语普通话，对中国境内的各少数民族的语言文字也颇为重视，但古典文献信息处理和汉语方言信息处理水平则远远落后于现代汉语普通话^②。目前，现代汉语普通话已经完成了字处理任务，词处理、句处理乃至篇章处理都有相当可观的研究成果。然而，古典文献信息处理基本上还处于字处理阶段，还缺乏一些基础资源和实用技术，例如用于超大汉字字符集的输入法、古籍汉字识别等等。

① 也有人认为是自然语言处理的一个分支。我们认为，自然语言处理的研究主要是那些需要较多智能的任务，例如自动问答、机器翻译等等。早期的自然语言处理被归入人工智能学科，就是一个明证。语言信息处理的研究范围较广，不仅包括智能处理，也包括那些不一定需要使用计算模型、但可为智能处理提供基础资源和基本技术手段的任务，例如字符编码、文本信息的输入输出等等。

② 汉语方言信息处理方面，已有一些可贵的探索，例如，通过相关系数计算和聚类分析来确定方言之间的亲属关系^{[18],[19]}，通过同源词统计确定语言或方言的形成年代、亲属关系^[20]，但总的来说还不成体系。汉语方言差异在很大程度上反映汉语的历时变化，因此汉语方言信息处理与中国古典文献信息处理是密切相关的。

一、先秦文献信息处理的研究目标

“中国古典文献学是综合运用版本、校勘、目录、注释、考证、辨伪、辑佚、编纂、检索等方面的理论与方法，科学地分析、整理、研究中国古代文献，进而探讨古代文献的产生、分布、交流和利用的规律，并总结对古代文献进行分析、整理、研究工作的规律与方法的学科。”^[17]

古典文献信息处理并不是要取代古典文献学。它的首要任务是要为古典文献学提供数字化处理的技术手段。中国古典文献有几千年的历史积累，卷帙浩繁，熟悉文献而又能融会贯通的专家不多，仅凭人力实在难以完成对古典文献的分析、整理和研究工作。电子计算机的问世以及汉字在计算机上的输入输出的实现，给古典文献处理带来了福音。例如，以现在的眼光来看，给文献编制逐字索引是一件很简单的事情，一部电子版的《左传》几秒钟即可完成索引。如果用以前抄卡片的方法编制逐字索引，一个人需要做好几年，还容易出错。

如果仅仅是为古典文献处理提供现代化的技术手段，古典文献信息处理还不足以成为一门学科。作为一门学科，它必须有自己的理论、方法和研究目标。作为语言信息处理（具体到中文信息处理）的一个分支，古典文献信息处理的理论基础是信息论和概率论，基本的研究方法是语料库统计，基本的工具是语言计算模型。它的研究目标也与古典文献学有所区别：

第一，古典文献学主要关注文献内容，古典文献信息处理除了关注文献内容之外，还应该特别关注文献的语言形式。先秦汉语是汉语发展的源头，先秦文献信息处理的一个重要的研究目标就是要从先秦文献中探寻这个源头的根本面貌，包括先秦汉语的语音、词汇和语法面貌，为汉语史的研究提供可靠的统计数据。鉴于课题的性质，我们主要是做了先秦汉语的词汇处理：对 25 种先秦文献全面地进行了词语切分和词性标注，在此基础上自动生成了先秦汉语词表，以先秦文献为参照考察了成语来源以及古今字和通假字的分布。（见第五章“古今字、通假字资源建设”，第六章“词汇概貌”）

第二，古典文献学对文献内容的关注，主要体现在对文献真实

性的关注,包括文献产生的时间、作者或编者、整篇或某一部分是否系后人伪托而成、各个版本之间的关系等等。古典文献信息处理对文献内容的关注,除了上述方面之外,还可体现为基于内容的计算。拿先秦文献信息处理来说,可以做《左传》和《国语》的事件抽取、人物关系抽取、先秦诸子思想比较等等。我们的检索系统不仅可以做字词检索,还可以检索《左传》中的人物、地理实体和事件,并以可视化界面展示社会网络分析的结果。(见第十章“《左传》中的春秋社会网络分析”)

二、先秦文献信息处理的任务

先秦文献信息处理的首要任务是文献数字化,数字化的基本含义是把文献变成机器可读的文本。方正超大字符集有6万多汉字,Unicode5.0及其以上版本有7万多汉字,已经基本够用。目前急需的是:(一)能支持超大字符集的各种方便快捷的汉字输入法:拼音输入(音韵学家使用)、字形输入(以笔画或部首为码元)、手写输入^①。(二)能尽量避免图形操作的造字工具。理想的造字工具应该是,用户只需输入若干字根以及字形的结构类型(上下结构、左右结构等),即可完成造字过程。(三)能支持繁体直排版的字符识别软件。对于古籍而言,理想的字符识别软件应能得到不同字号和字体的输出(区分正文和注释),保留专名标记,甚至嵌入造字模块。

现在从网上下载的先秦文献电子文本存在许多问题:录入较早的是GB2312-80编码的简体文本,繁体文本不多。有些繁体文本是从简体文本自动转换过来的,简繁转换错误未经校对。冷僻字通常用两个汉字组合表示。此外还有脱字、讹字等等。这些电子文本大多是热心网友的无私奉献,应该好好加以利用。人工校对是不可避免的,但应该研究辅助校对工具,尽量减少人工劳动。辅助校对工具应能运用本校法(对于篇幅较大的文献)和他校法(对于有多个版本的文献)自动发现可能的错误。

跟校对相关的工作是校勘。在古典文献信息处理中,校对的目标是做到电子文本与纸质文本一致,校勘是发现同一种文献的各个

^① 逍遥笔6.5版本支持超大字符集的手写输入,但字符集设置需手工改写启动文件。