文章编号:1005－0523(2005)05－0159－04

古代汉语标注语料库的建设与应用

徐紫云

(华东交通大学 人文学院社会科学学院,江西 南昌 330013)

摘要:近年来,许多国家都非常重视语料库的建设.我国陆续建成了一些规模不同、功能各异的现代汉语标注语料库,但古代汉语标注语料库的建设却十分欠缺,使得古代汉语研究及教学手段明显落后.本文从语料库的语料选取、标注及应用等几方面阐述古代汉语标注语料库的建立方法及意义.

关 键 词:古代汉语;语料库;标注;选取;应用
中图分类号:F062.2　　**文献标识码**:A

0 引言

任何民族的发展都不能没有继承.我国大量优秀的传统文化均以古代汉语作为载体.中华文化要传承、要变革,都得要有一批人读文言文,整理古籍,研究历史.[1]但是,随着现代科技的飞速发展,人们的工作方式、学习方式与生活方式正在发生深刻变化,少部分人利用传统方式整理古籍,难以把大量优秀的文化成果传递给现代人.为了适应信息化的要求,古代汉语学习与研究手段的更新势在必行.

语料库是指按照一定的语言学原则,运用随机抽样的方法,收集自然出现的连续的语言运用文本或话语片段而建成的具有一定容量的大型电子文库[2].自上个世纪 60 年代世界上第一个计算机语料库——布朗语料库诞生以来,计算机语料库已成为语言研究的一种普遍资源,语料库的建设日益受到各国语言学家的重视,我国从 70 年代末开始,也陆续建立了一批大规模计算机语料库,不仅用于汉语计量分析研究、语言研究,还可用于自然语言处理研究.但是,目前所建立的最典型的几个大型汉语语料库,除台湾中央研究院平衡语料库收集了超过百万字次的古汉语语料外,大部分为现代汉语语料库.并且目前已有古代汉语语料库中,绝大多数为仅供检索用的未经深加工的语料库.如深圳大学的《红楼梦》检索系统及市场出售的古籍检索光盘等.尽快建立符合标准的用途广泛的大型古代汉语计算机标注语料库对于古籍的整理、古汉语辞典的编纂、古代汉语的研究及教学都将起到非常积极的作用.

1 古代汉语标注语料库的建设

任何语料库研究均开始于语料库体的建立,而语料库的质量又将影响以后所要做的一切基于语料库的研究工作,所以语料库语料的选择及标注都是极为重要的工作.

1.1 古代汉语计算机语料库语料的选取

一般而言,语料的数量越多,语料库就越有代表性.计算机大容量的存储功能及良好的性能为大规模语料库的构造提供了可能;电子出版物的广泛普及和 OCR 的应用也使语料的获取更为方便(清华大学所建现代汉语生语料库已达 7－8 亿字).但目前语料库的规模受到加工能力的限制,现有语料加工还不能完全实现自动化,信息处理主战场还在“词”的处理,自动分词中,歧义字段的切分处理结果尚不尽人意.至于对句法及语义特征标注的研究就更少.一些特殊用途的语料库(如清华大学智能技术与系统国家重点实验室与北京语言文化大学语言信息处理处研究所联合研发的 HuaYu 语料库、上海师范大学建立的 300 万字的标注语料库和 100 万字《作家文摘》标注语料库)均为手工标注语料库.古代汉语语料库的建设尚处起步阶段,标注目前无法实现自动化.面对丰富的古代典籍,我们必须确立一个合理的语料选取原则:

收稿日期:2005－10－18
基金项目:2005 年江西省高校人文社会科学规划基金,编号:YY0504
作者简介:徐紫云(1973－),女,江西德兴人,华东交通大学人文学院讲师,主要研究方向:古代汉语,计算语言学.

1) 以共时为主

古代汉语大致有两个系统:一是以先秦口语为基础而形成的上古汉语书面语言以及后来历代作家仿古的作品中的语言,也就是通常所谓的文言;一个是唐宋以来以北方话为基础而形成的古白话.[3]作为中国文学的源头,先秦的作品,无论是思想还是语言对后代的影响都非常深远,古代汉语学习和研究的对象主要是这一时期的作品.所以,语料库中语料的选取对象重点应是先秦的典范作品,少量吸收汉代及其他朝代仿古的作品.出于对语言发展动态研究的需要,以后再逐步建立不同时期作品的语料库.

2) 语料的平衡性

一个语料库必须在某种程度上能够代表某一种语言,这就需要考虑语料库的平衡性.古代汉语语料库的平衡性是相对而言的:首先,由于客观原因,我们不可能搜集到口语语料,而只能录入书面语;其次,由于研究目的与现代汉语不同(现代汉语研究有规范与应用的要求,古代汉语的研究更主要的在于对古代文化精髓的掌握),本语料库所涉及的语料绝大部分为正式语言、文学语言.古代汉语语料库的平衡主要体现在题材及体裁的平衡上,所选取的语料题材包括神话、宗教、卜筮、政治、文学、历史、地理、哲学、语录、军事等方面;体裁则包括诗歌、散文、议论文、应用文、寓言故事、赋等.

1.2 语料的标注

未经任何处理的电子文本语料库被称为生语料库,其中没有包括词法及语法等信息,应用价值非常有限,要实现语料库的多种价值,必须对语料库进行多层次的标注.

1) 标注规范

无论是现代汉语还是古代汉语,不同语法体系对于语法的描述都是不完全相同的.在对现代汉语语料库进行标注时,很多专家采用的是中学语法教学中长期使用的《暂拟系统》,也有些(如教育部语言文字应用所研制《信息处理用现代汉语词类标记集规范》)是把具有代表性的吕叔湘、朱德熙、张斌、胡裕树等先生的语法体系与中学教学语法系统相结合.对于古代汉语来说,中学语文教学中,文言文的比例较小.所以,这个语料库并不采用暂拟语法系统,而是采用具有代表性的马建忠、王力等语法学家的语法系统.

(1) 词类标记

目前,古代汉语语料库还没有统一的标注规范,各个语料库之间,不能实现数据共享.考虑到现代汉语与古代汉语的相继承性及古今对译的需要,我们认为,古代汉语语料库的词类标记可以在《信息处理用现代汉语词类标记集规范》[4]与《信息处理用现代汉语词类及标记集》[5]的基础上,结合古代汉语的实际情况加以修改.具体的分词标准并不与现代汉语一致,但相同词类可以采用同一标记.

依照《规范》分为20个大类,小类基本采用《标记集》中的分法,但有所改变:代词、形容词与连词三大类中增加更为具体的小类标记;助词类中删去结构助词与时态助词二类(结构助词中的"所"归入特殊代词一类),语气助词增加句首语气助词与句中语气助词二类,由于许多古代典籍并无标点,各种语气主要靠语气词来实现,所以句末语气助词又分为疑问、判断、感叹和祈使四类;在部分词类下增加某些功能标记:动名词(名词活用为动词)、动形词(形容词活用为动词)、动数词(数词款用为动词).

具体标记如下:

1.名词 N	9.数词 M
1.0 普通名词 NG	9.1 序列词 MI
1.1 方位名词 ND	9.2 动数词 MV
1.2 时间名词 NT	10.量词 Q
1.3 地名 NS	10.1 名量词 MQ
1.4 处所名词 NL	10.2 动量词 QV
1.5 人名 NH(姓 NHF、名 NHG	11.代词 R
1.6 族名 NN	11.1 人称 RR
1.7 其他专名 NZ	11.2 指示 RZ
1.8 非量名词 NO	11.3 疑问 RY
1.9 动名词 NV	11.4 无指 RW
2. 副词 D	11.5 特殊(者、所)RT
3. 拟声词 O	11.6"诸"RZP、"焉"RZY
4. 叹词 E	12.介词 P
5. 动词 V	12.1 把(将)PBA
5.1 一般动词,非谓语中心动词且不带宾 VG	12.2 被(让、叫)PBEL
谓语中心且不带宾 VGO	12.3 在 PZAI
带名宾 VGN	13.连词 C
带动宾 VGV	13.1 句子连词 CS
带形宾 VGA	13.2 词与词组连词 CW
带小句宾 VGS	14.简略语 J
带双宾 VGD	15.助词 U
带兼语 VGJ	15.1 语气助词 UY
5.2 能愿动词 VA	句首 UYF
5.3 趋向动词 VQ	句中 UYM
5.4 是(表判断)VY	句末 UYE
5.5 有 VH	疑问 UYEY
6.形容词 A	判断 UYEP
6.1 性质形容词 AQ	感叹 UYEG
6.2 状态形容词 AS	祈使 UYEQ
6.3 动形词 AV	16. 习用语 I
7.	17. .非语素字 X
7.1 标点符号 WP	18.前接成分 H
7.2 非汉字字符 WS	19.后接成分 K
8.区别词 F	20.语素 G

2) 字标记

a.异体字 ZY

b.同音借用字 ZJ

c.古字 ZG

d.不可通用的简繁字 ZH

(3) 音标记

a.别义音或破读音 YY

b.生僻字读音 YS

c.人地名特殊音 YT

(注汉语拼音的同时,加注反切)

(4) 短语结构标记

主谓结构 ZW、动宾结构 DB、定中结构 DZ、状中结构 ZZ、数量结构 SL、宾语前置 TB、谓语前置 TW

在完成以上标注的同时,还对部分语篇信息及词义进行了标注(BC),另外以单独文本的方式对结构不明的单复句进行注释.

2) 标注

古代汉语语料库主要采取人工标注的方式,语料标注示例如下:

<bc[当初]>初|nt,|wp <bc[郑国第二代君主,姬姓,名掘突,谥武公]>郑武公|nh 娶|vgo 于|p<bc[姜姓小国]>申|ns,|wp 曰|vgn<bc[武公正妻姜氏.“武”表丈夫的谥号]>武姜|nh,|wp 生|vgn<bc[郑国第三代君主,谥庄公]>庄公|nh 及|c<bc[“段”为名,“叔”表排行第二,后出奔共,称“共叔段”]>共叔段|nh.|wp 庄公|nh[zz<bc[难产,分娩时脚先出]<zj[ ]bc[倒逆]>寤|sa 生|vgo],|wp 惊|avs 姜氏|nh,|wp 故|c 名|vgo 曰|vgo 寤生|nh,|wp<bc[於是]>遂|ds<yy[wu4 乌各切] bc[厌恶]>恶|vgn 之|rr.|wp 爱|vgn 共叔段|nh,|wp 欲|va<bc[确定地位] >立|vgn 之|rr.|wp<bc[屡次] ns[qi4]>亟|dq 请|vgo 於|p 武公|nh.|wp 公|nh 弗|df 许|vgo.|wp

2　古代汉语语料库的应用

构建一个古汉语语料库,能够为人们提供一个古代汉语研究平台,其作用体现在以下几个方面:

2.1　检索

检索功能是任何一个语料库最基本的功能.古代汉语标注语料库蕴含着词信息及部分语法功能信息,通过索引,不仅能进行传统的训诂研究,还能从各个角度将索引内容与所在的语言环境一起加以考察分析,以更科学地研究古汉语的语法现象以及与相关的语言现象进行对比研究.传统的基于关系的全文检索系统基本上不利用语言知识,因而极易出现漏检和误检,并且索引空间太大.比较理想的基于计算语言学的全文检索系统需要对文本资料进行语言学意义上的理解,因而,标注语料库就显得非常重要.

2.2　统计与分析

目前主要有字数统计、字频统计、词频统计、句数统计、句长统计等.统计量的使用并不只是对语料进行简单的计数.通过统计,还可以用来说明文体、写作风格和语言之间的差异以及定义一个词的义项以及它们的使用环境,这些信息对于词典编纂、自然语言处理以及语言教学都非常重要.[4]

2.3　古代汉语辞典的编纂

基于语料库的词典编纂,在义项的整理与例证的搜集方面有着传统编纂法无法比拟的优势.比如,计算机可以提供便捷、可靠的语料检索,统计工具等等.和人工相比,计算机可以轻易地在一个几千万词次的语料库中无遗漏地找出任何一个词项在整个语料库中的全体实例,并穷尽地生成这个词项的引文或例句.此外,利用语料库,还可以进行各种古代汉语专门辞书的编纂.

2.4　古代汉语辅助教学

近年来,语料库在外语及现代汉语教学方面得到了直接或间接的应用,古代汉语则相对落后.古代汉语缺乏系统的语法教学,教材中的通论与文选两部分各成体系,关联并不紧密,文选的选择与排列并不以通论的编排为据,通论中的语法要点所涉及的句子散居各个单元之中.利用传统教学方法,很难使学生形成系统的概念.我们在已有语料库的基础上,编制程序,把整个系统分为四个模块:系统维护模块、统计模块、检索模块(这三个模块可以同时用于研究与教学)、辅助教学模块.利用辅助教学系统进行语言信息综合与分类,以有序的通论引导无序的文选,浓缩教学内容,极大地增加了信息量,在帮助学生从更为全面的文选材料中掌握古汉语的语法语用规律方面取得了较好的效果.

2.5　机器翻译

我国传统文化典籍可谓浩如烟海,目前已有的今译本微乎其微,并且今后也很难再投入巨大的人力物力对这些典籍进行古今对译.如果能实现机器自动翻译,就不仅能满足各种人不同的阅读需要,也能不断地适应语言体系的发展.经过深加工的语料库是进行真实文本自动处理和机器翻译的重要基础.语料库技术可以从词典信息(词法、语法和语义信息)、文法规则、目标语的生成及测试与评价四个方面支持机器翻译系统的开发与研究.

3　结语

本文主要介绍了古代汉语标注语料库的建设方法及其应用.建库过程中,我们的标注规范尽可能地向现代汉语语料标注的国家标准靠拢,但因为古代汉语与现代汉语语法体系的不同,所以在具体的词类划分方面还是有所不同.应用方面,由于语料库还在建设之中,因而主要还是用于检索、常用词的统计与教学方面.随着语料库规模的不断扩大,应用研究也将不断深入.

参考文献:

[1] 李如龙.文言 白话 普通话 方言[J].语言文字应用,2003,11.

[2] 李文中.语料库,学习者语料库和外语教学[J].外语界,1999,1.

[3] 王　力.古代汉语[M].北京:中华书局,1999.

[4] 靳光瑾,郭曙纶,肖　航,章云帆.语料库加工中的规范问题——谈《信息处理用现代汉语词类标记集规范》

[J]. 语言文字应用,2003,11.
[5] 刘开瑛. 中文文本自动分词和标注[J]. 商务印书馆, 2000,5
[6] 黄昌宁,李涓子. 语料库语言学[J]. 商务印书馆,2002.

Tagged Archaic Chinese Corpus Creation

XU Zi-yun

(School of Humanities and Social Sciences , East China Jiaotong University, Nanchang 330013, China)

Abstract: Corpus creation is being paid great attention to in many countries recent years. Some contemporary Chinese corpora with different scale and usage have been built, but few tagged archaic Chinese corpora are built. This paper presents ways of processing and significance of archaic Chinese corpus from following aspects: selections, tagging, uses.
Key words: archaic Chinese; corpus; use; selection; tagging

(上接第 154 页)

Prof. Xu Yuanchong's Pursuing of Aesthetics in Translation of Poems

HE Shan-xiu

(School of Foreign Languages, East China Jiaotong University, Nanchang 330013, China)

Abstract: Tlis article briefly summaries Prof. Xu's translation theory, that is, art of beautifulization and creation of the best as in rivalry. By appreciating two poems of Meng Haoran's English version, the author tries to prove that Prof. Xu's version can best express the original idea compared with the other two versions.
Key words: Xu Yuanchong; translation of poems; aesthetics