面向语言教学和辞书编纂的汉语平衡语料库建设*

邱立坤 亢世勇
鲁东大学文学院，山东烟台 264025
qiulikun@gmail.com

摘要：语言教学和辞书编纂走向现代化的一个必需的手段是借助语料库技术的辅助，基于此，语料库的建设就成为一个亟待解决的问题。参照美国当代英语语料库的框架，我们尝试构建一个汉语平衡语料库，分为口语、小说、报纸、学术期刊、时尚杂志五种体裁，分年度采集。本文介绍了该平衡语料库的语料选择和深加工等方面的构建情况。

关键词：汉语 平衡语料库 语言教学 辞书编纂

Construction of a Chinese Balanced Corpus for Language Teaching and Dictionary Compilation

Qiu Likun and Kang Shiyong
School of Chinese Language and Literature, Ludong University
qiulikun@gmail.com

Abstract: One essential means for the modernization of language teaching and lexicography is the use of corpus-based technology. From this view, the construction of large scale corpora becomes a valuable issue. According to the framework of the Corpus of Contemporary American English, we try to build a balanced Chinese corpus, which is divided into five genres, including oral, novels, newspapers, journals, fashion magazines and will be gathered annually. This article describes the balance corpus, including text selection and annotation.

Key words: Chinese, balanced corpus, language teaching, dictionary, compilation

1 引言

随着全球化和信息时代的到来，语言变化的速度愈来愈快，其中词汇上的变化尤其明显。传统的词典编纂手段因此面临巨大的挑战，难以追上语言日新月异的步伐。从 20 世纪 50 年代开始，基于计算机技术的语料库词典编纂方法开始被引入到词典编纂领域，并且随着计算机技术的发展而突飞猛进，目前已成为词典编纂的主流方法。

与传统方法相比，大规模的语料库可以为词典编纂提供大量自然的、真实的例句，克服传统方法过于依赖人的直觉经验的缺陷，从而大幅度提高词典的质量。英美等国面向辞

* 本文曾在 2013 年汉语词汇语义学会议上宣读，有修改。本文工作受到国家自然科学基金青年项目（61103089）、山东省优秀中青年科学家科研奖励基金（BS2013DX020）、鲁东大学人文社会科学研究项目（WY2013003）的资助。

编纂建立大规模的语料库，比如COBUILD的英语文库(The Bank of English)、朗文语料库网络(Longman Corpus Network)、英国国家语料库(British National Corpus，BNC)和剑桥国际语料库(Cambridge International Corpus, CIC)等，规模都在亿词量级，而且具有良好的架构，覆盖了多种语体。在这些语料库的基础上，一系列基于语料库方法编纂的词典相继出版，包括朗文当代英语词典、剑桥国际英语词典、麦克米伦高阶英语词典、朗文联想词典等。这些词典面世之后取得巨大的成功，且能以较快的频率更新版本，进一步显示了语料库方法在词典编纂中的地位和价值，反过来又推动了更大规模语料库的建设和发展。

国内的语料库词典学研究也取得了较大进展，香港科技大学与广东外语外贸大学联合建立了科技英语语料库，南京大学词典研究中心与商务印书馆合作筹建了CONULEXID语料库。但是与欧美国家尤其是英国和美国相比，我国在面向辞书编纂的汉语语料库建设方面还有较大差距。

基于此，我们计划参考当代美国英语语料库（Corpus of Contemporary American English）构建一个面向辞书编纂的大型现代汉语语料库。这个语料库应该具有以下特点：

（1） 监控语料库。分年度采取语料，通过长时间的积累反映出语言的变化。

（2） 多语体。覆盖口语、小说、新闻、杂志、学术期刊五种基本语体构成基础语料库；再辅以中小学教材、微博语料、产品评论、专利等特殊类型语料构成的扩展语料库。

（3） 深加工。其中部分语料可以提供经过人工校对的分词和词性标注结果以及句法分析结果；剩余语料使用自动分析工具产生词语切分、词性标注和句法分析结果。

2 基础语料库

基础语料库参照当代美国英语语料库（Corpus of Contemporary American English）的构建方式和框架，分口语、小说、新闻、杂志、学术期刊五种体裁，各体裁规模相等；按年度选取，每年选取 2000 万词，每种体裁不少于四百万词(约 700 万字)。

2. 1 口语

使用对话的转写形式，一般已经具有相应的电子文本。因为可选择来源较少，无法按年度每年选取 400 万词。

主要包括两类：

（1）电视访谈节目：杨澜访谈录、鲁豫有约、新闻今日谈（凤凰卫视，百度文库）、新闻 1+1、锵锵三人行（凤凰卫视，百度文库，“锵锵三人行+文字”）、非你莫属、实话实说

（2）都市家庭情景喜剧的对白：我爱我家、编辑部的故事、家有儿女、炊事班的故事。

2. 2 小说（按年度选取）

小说类以文学类期刊为主，以小说、散文为主，适当选取少数影响较大的网络小说和电影电视剧本。具体分为以下几类：

（1）青少年读物: 故事会、读者、上海故事、北京青年、知音、青年文摘。

（2）文学刊物:《人民文学》、《收获》、《当代》、《十月》、《作家》，湖北的《长江文艺》、浙江的《江南》、江苏的《钟山》、云南的《大家》、湖南的《芙蓉》、宁夏的《黄河文学》；

散文类刊物：《散文》《散文选刊》《散文百家》；小说类刊物：《北京文学》、《小说月报》；戏剧类刊物：《剧本》。

（3）小说、传记的第一章，包括：

长篇小说

1. .金宇澄：《繁花》，《收获》(长篇专号)2012 年秋冬卷
2. 李佩甫：《生命册》，《人民文学》2012 年第 1、2 期
3. 鲁敏：《六人晚餐》，《人民文学》2012 年第 3 期
4. 陈亚珍：《羊哭了，猪笑了，蚂蚁病了》，北京燕山出版社 2012 年 7 月
5. 杜光辉：《大车帮》，作家出版社 2012 年 3 月

中篇小说

1. 格非：《隐身衣》，《收获》2012 年第 3 期
2. 陈谦：《繁枝》，《人民文学》2012 年第 10 期
3. 计文君：《白头吟》，《人民文学》2012 年第 7 期
4. 陈应松：《无鼠之家》，《钟山》2012 第 2 期
5. 余一鸣：《愤怒的小鸟》，《人民文学》2012 年第 6 期
6. 刘建东《羞耻之乡》，《山花》2012 年第 9 期
7. 弋舟：《等深》，《乌江》2012 年第 5 期
8. 尤凤伟《岁月有痕》，《十月》2012 年第 3 期
9. 曹寇：《塘村概略》，《收获》2012 年第 4 期
10. 王手：《贴身人》，《收获》2012 年第 6 期

短篇小说

1. 王祥夫《归来》，《天下》2012 年第 2 期
2. 姚鄂梅《狡猾的父亲》，《人民文学》2012 年第 2 期
3. 朱山坡《灵魂课》，《收获》2012 年第 1 期
4. 王璞《捉迷藏》，《收获》2012 年第 1 期
5. 魏微《胡文清传》，《花城》2012 年第 1 期
6. 南翔《绿皮车》，《人民文学》，2012 年第 2 期
7. 斯继东《你为何心虚》，《上海文学》2012 年第 10 期
8. 阿乙《阁楼》，《当代》2012 年第 3 期
9. 董立勃《杀瓜》，《作家》2012 年第 1 期
10. 裘山山《意外伤害》，《长江文艺》2012 年第 9 期

2. 3 新闻(按年选取)

从全国选择若干有代表性的报纸（十家，含日报、周报、晚报、都市报等，覆盖多个地区）：南方周末，人民日报，中国青年报，新民晚报，楚天都市报（武汉），华西都市报（成都），广州日报（广州），华商报（西安），南方都市报（广州），齐鲁晚报（济南）

把十家报纸的版块组合起来，以人民日报、广州日报、中国青年报三家为主。

版块：国际（人民日报、广州日报），国内（中国青年报），本地（楚天、华西、南方、齐鲁、新民、华商等都市报和晚报），财经（南方周末、中国青年、广州日报），生活、体育、社论、军事（人民日报、广州日报），科技（含互联网、IT），（广州日报、中国青年报），

房产（中国青年报、广州日报、新民晚报），汽车（广州日报、中国青年报），娱乐（广州日报、新民晚报），教育（人民日报、广州日报），社会（广州日报、南方都市报）。

2. 4 杂志(按年选取)

公开出版的时尚杂志 100 种，覆盖新闻/观点/评论、财经、科技、文学艺术、宗教、体育、娱乐、家庭/健康、男人女人、育儿、汽车、IT 等各个领域（参考龙源期刊排行榜选取）。该排行榜中的 100 种期刊为：

三联生活周刊，电脑爱好者，看天下，新民周刊，大众摄影，南都娱乐周刊，南方人物周刊，环球宝贝，轻兵器，意林，故事会，中国经济周刊，中国新闻周刊，第一财经周刊，电脑迷，民间故事选刊，中国周刊，文明，南都周刊，时尚内衣，财经，读者，博客天下，时代影视，看世界，城市建设理论研究，商界，考试周刊，党建，读写算，管理观察，都市丽人，IT 经理世界，南风窗，伴侣 ，恋爱婚姻家庭，今古传奇武侠版，兵器知识，创业家，环球企业家，中国实用医药，读书，计算机应用文摘，理财周刊，故事林，诗刊，上海故事，农村百事通，北京青年，航空知识，人生与伴侣，作文与考试初中版，环球人物，大众电影，微型计算机，钱经 ，《新世纪》周刊，看历史，中国社区医师，当代体育足球，法制与社会，领导文萃，世界知识，视野，收藏，价值工程，为了孩子，今日文摘，炎黄春秋，瞭望东方周刊，中国国家旅游，父母必读，恋爱婚姻家庭 ，商业时代，健康必读，做人与处世，作文与考试高中版，现代商贸工业，经济师，中国医药导报，短小说，电脑知识与技术，经济研究导刊，海外文摘，成才之路，大众医学，青年博览，中国计算机报，电影文学，37°女人，数学学习与研究，消费导刊，中国新技术新产品，中国教育技术装备，文史天地，中国中医药咨讯，百科知识，大观周刊，会计之友，中国市场

2. 5 学术期刊(按年选取)

每个学科领域选择一种代表性的学术刊物，比如各一级学会的学报。累计 100 种：

1. 政治学研究 2. 马克思主义研究 3. 中共党史研究 4. 哲学研究 5. 哲学动态 6. 宗教学研究 7. 社会学研究 8. 民族研究 9. 中国人口科学 10 法学研究 11. 中国法学 12. 中外法学 13. 科学学研究 14. 科研管理 15. 科学管理研究 16. 经济研究 17. 金融研究 18. 中国经济史研究 19 教育研究 20. 高等教育研究 21. 中国高等教育 22 体育科学 23. 中国体育科技 24. 体育学刊 25. 中国语文 26. 外语教学与研究 27. 语言文字应用 28. 语文研究 29. 文学评论 30. 外国文学评论 31. 文学遗产 32. 外国文学 33. 中国音乐学 34. 美术 35. 音乐研究 36 历史研究 37. 世界历史 38. 中国史研究 39. 中国图书馆学报 40. 大学图书馆学报 41. 图书馆杂志 42. 国家图书馆学刊 43 编辑学报 44. 新闻与传播研究 45. 数学年刊 A 辑 46. 计算数学 47. 数学进展 48. 物理学报 49. 力学学报 50. 光学学报 51. 计算力学学报 52. 高等学校化学学报 53. 分析化学 54. 应用化学 55. 地理研究 56. 地理科学 57. 地学前缘 58. 海洋与湖沼 59. 中国环境科学 60. 海洋地质与第四纪地质 61. 生物工程学报 62. 遗传 63. 中国生物工程杂志 64. 系统工程理论与实践 65. 信息与控制 66. 仪器仪表学报 67. 机械设计 68. 机械科学与技术 69 中国电机工程学报 70. 电力系统自动化 71. 电工技术学报 72. 化工学报 73. 石油学报 74. 精细化工 75. 自动化学报 76. 电子学报 77. 通信学报 78. 计算机学报 79.

软件学报 80. 计算机研究与发展 81. 岩石力学与工程学报 82. 土木工程学报 83. 建筑结构学报 84. 汽车工程 85. 中国造船 86. 中国公路学报 87. 宇航学报 88 材料研究学报 89. 高分子材料科学与工程 90. 材料工程 91. 水产学报 92. 作物学报 93. 林业科学 94. 园艺学报 95 管理科学学报 96. 管理世界 97. 中国管理科学 98. 中国中药杂志 99. 中国药学杂志 100. 中国中西医结合杂志

3 扩展语料库

包括从人民日报创刊以来的所有文本， 1998 年和 2000 年已进行分词和词性标注的人工校对，其中有 400 万字左右进行了句法标注。

4 原始文本存储格式

原始文本以 XML 格式存储，每篇文本包括类别、来源、作者、时间、标题和正文等信息。示例如图 1 所示。

```
<Article>
<Category>财经</Category>
<Source>中国青年报</Source>
<Author>白雪</Author>
<Time>2012-01-09</Time>
<Title>"菜篮子"推广首重无公害</Title>
<Content>
    安徽和县蔬菜北京推介会近日在京举办，该县年产近百万吨无公害蔬菜，
    是"长江中下游最大的菜篮子"。2002年和县被评为全国无公害农产品（
    蔬菜）生产示范基地先进县，2009年被确立为全国农业标准化示范区。
</Content>
</Article>
```

图 1 原始文本存储格式

5 语料库的加工

词语切分和词性标注以北京大学 2003 版词语切分和词性标注规范（俞士汶等，2003）为基本依据，在该规范的词类体系上进行了适当的归约，总计包括 26 个词性标签：

18 个基本词性：名词 n、时间词 t、处所词 s、方位词 f、数词 m、量词 q、区别词 b、代词 r、动词 v、形容词 a、状态词 z、副词 d、介词 p、连词 c、助词 u、语气词 y、叹词 e、拟声词 o

5 个名词小类：人名 nr，地名 ns，团体机关单位名称 nt，其他专有名词 nz，英语等其他非汉字的字符串 nx。

2 个名物化词性标记：动词和形容词的特殊用法标记，即名动词 vn（具有名词功能的动词），名形词 an（具有名词功能的形容词。

1 个标点符号标记：w。

句法标注部分采用邱立坤（2012）提出的依存句法标注体系（如表 1 所示），图 2 是一个句子句法标注结果的可视化示例。

表 1 依存句法标注体系

依存关系	符号	依存关系	符号
核心	HED	数量	QUN
主语	SBV	前附加	LAD
话题	TPC	后附加	RAD
强调	FOC	介宾	POB
宾语	VOB	的字	DE
间接宾语	IOB	地字	DI
行为宾语	ACT	得字	DEI
连动	VV	重叠	RED
补语	CMP	独立结构	IS
状语	AD	小句	IC

依存关系	符号	依存关系	符号
时体	MT	标点	PUN
数量补语	QUC	并列	COO
定语	ATT	共享并列	COS
数字	NUM	同位	APP
并列式独立成分	ISC	跨小句标点	PUS

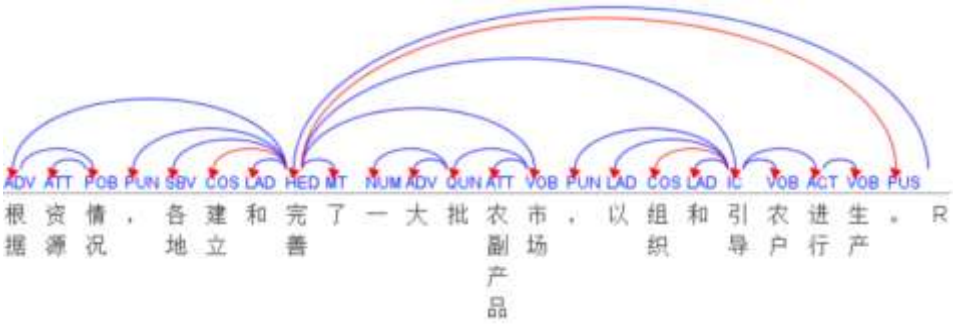

图 2 句法树标注实例

6 小结

目前，我们已经采取了口语、小说、新闻、杂志、学术期刊五种语体的语料各 700 万字左右（其中后四种语体均为 2012 年的文本），形成了一个 3500 万字左右的平衡语料库。在这一种框架下，我们计划以后每年选取同等规模的语料，形成一个具有监控功能的语料库。此外，还建立了人民日报历时语料库等扩展语料库。

在此基础上，我们计划：（1）建立计算机辅助的辞典编纂系统，可以根据需要选择不同的语料库；（2）加入微博、产品评论、专利等子语料库；（3）进行辞典编纂的实践；（4）进行辅助语言教学的实践。

参考文献

[1]Davies, Mark. The Corpus of Contemporary American English: 450 million words, 1990-present. Available online at http://corpus.byu.edu/coca/.

[2]邱立坤． 多视图汉语树库构建的理论研究与实践． 北京大学博士后研究工作报告． 2012.

[3]俞士汶、段慧明、朱学峰、孙 斌、常宝宝． 北大语料库加工规范:切分·词类标注·注音． 汉语语言与计算学报，2003，13(2)：121—158．