全国高等学校外语教师教学实践系列

语料库应用教程

Using Corpora: A Practical Coursebook

梁茂成 李文中 许家金 著

外语教学与研究出版社
FOREIGN LANGUAGE TEACHING AND RESEARCH PRESS

全国高等学校外语教师教学实践系列

语料库应用教程

Using Corpora: A Practical Coursebook

梁茂成 李文中 许家金 著

外语教学与研究出版社
FOREIGN LANGUAGE TEACHING AND RESEARCH PRESS
北京 BEIJING

图书在版编目(CIP)数据

语料库应用教程／梁茂成，李文中，许家金著．— 北京：外语教学与研究出版社，2010.7
（全国高等学校外语教师教学实践系列）
ISBN 978－7－5600－9844－9

Ⅰ．①语… Ⅱ．①梁… ②李… ③许… Ⅲ．①词语—研究—教材 Ⅳ．①H03

中国版本图书馆 CIP 数据核字（2010）第 144592 号

出 版 人：于春迟
项目负责：段长城
责任编辑：郑丹妮 段长城
封面设计：覃一彪
版式设计：吴德胜
出版发行：外语教学与研究出版社
社　　址：北京市西三环北路 19 号（100089）
网　　址：http://www.fltrp.com
印　　刷：北京双青印刷厂
开　　本：787×1092 1/16
印　　张：15.75
版　　次：2010 年 7 月第 1 版 2010 年 8 月第 1 次印刷
书　　号：ISBN 978－7－5600－9844－9
定　　价：43.90 元（含 CD-ROM 一张）

*　　*　　*

购书咨询：(010)88819929　　电子邮箱：club@fltrp.com
如有印刷、装订质量问题，请与出版社联系
联系电话：(010)61207896　　电子邮箱：zhijian@fltrp.com

物料号：198440001

序

语料库语言学是语言学科中飙升得最快的学科之一。早在 1898 年，德国的 Kaeding 就建立过德语单词的频率词典；1921 年美国 Thorndike 编写过 10,000 词频率的教师词汇手册，并在他所编著的初级英语词典里，对头 3,000 个常用词做了标注；1953 年 Michael West 独辟蹊径，对 2,000 个常用词的义项频率做了统计；Quirk 于 1959 年开始了英语用法调查（Survey of English Usage）这一项目。但是，所有这些都不是用计算机来完成的。第一个能够实现机读的语料库是 1967 年 H. Kučera 和 W. Nelson Francis 在美国 Brown 大学建的 100 万词现代美国英语语料库。所以上世纪 60 年代被认为是现代计算机语料库真正诞生的年代。但是当时在美国大行其道的是 Chomsky 语言学，以理性主义为基础的 Chomsky 语言学和以实证主义为基础的语料库语言学是南辕北辙的。Francis 回忆往事时，谈到他在 1962 年建库初期，曾经碰到一个语言学教授，Francis 告诉他正在建立语料库，这位教授听到后说："你这是在浪费自己的时间和政府的金钱。你是一个英语本族语者，在 10 分钟内就能够说出一个英语语法点的许多例子，比几百万词的随机文本所提供的还要多"（见 Svartvik 2007）。

语料库语言学后来在欧洲得到迅猛的发展。2006 年，Jan Svartvik 曾经通过 Google 在互联网上检索 3 个词：transformational grammar、minimalist program、corpus linguistics, 其命中数分别为 12,400、11,600、34,500。为什么欧洲对语料库语言学会有如此大的兴趣呢？这可以从两个角度看：一是在美国占统治地位的 Chomsky 语言学的研究焦点是句法，它感兴趣的是哪些句子是可能的（What is possible?），追求的是语言理论的"解释力"，例如 Colorless green ideas sleep furiously、The rat the cat the dog chased ate died 都是可能的，但语料库语言学对此却没有兴趣，因为实际上没有人会这样说。在哪个语料库里都无法检索出这类"可能的"句子。相反地，语料库语言学感兴趣的是哪些语言现象在实际使用上是很有可能的（What is probable?），这和统计学中的概率有关，不是可否的问题，而是多少的问题，也就是语言的使用问题。语料库语言学在欧洲之所以能够生根发芽，是因为欧洲（特别是北欧）的语言学家历来关心语言和社会生活间的关系，对用法特别感兴趣。欧洲语言的多元性使语料库语言学更能够大展拳脚，因为它可以用于翻译、词典编纂和对比语言学研究等多种语言研究活动中。英国的几家出版社，如 Collins、Longman、Cambridge 和 Oxford 看到语料库在词典编纂上大有作为，也对语料库建设予以大力资助，而且建立了专用的数据库来收集新词、新义和例句。有的词典，如 *Collins COBUILD* 更是在 Collins 出版社和 Birmingham 大学合作建成的 6.45 亿词的 Bank of English 的基础上开发出来的，令人耳目一新。一直到世纪之交，美国学者才意识到他们在语料库语言学方面已经大大落后于欧洲的学者，至 1999 年才在 Michigan 大学召开第一次全国性研讨会，对分散在 Massachusetts (Boston)、Northern Arizona、Rice、California (Santa Barbara)、Michigan、Pennsylvania、Illinois State、Cornell、Tempere 等大学的语料库语言学研究做了一番检阅。在此次会议上，Simpson & Swales (2001) 承认美国落后了 15 年，但他们乐观地预言："正如我们所看到的后卫星年代的空间竞赛和最近一个世纪的美国汽车工业一样，美国具有早经证实的迎头赶上的能力。"

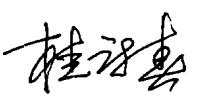
语料库语言学之所以能够发展神速和计算机的普及大有关系。Brown 语料库当初创建时需要动用到该大学的大型计算机。当使用到这套设备时，整个大学的科室和以计算机为基础的研究活

动都必须停止，给它让路。但是到今天，任何人在家里都可以借助个人电脑完成百万词级语料库的建设，并可以对语料库进行复杂检索或开展其他研究活动。那么怎样开始动手，从无到有地建立语料库呢？这里既牵涉到对语料库语言学的一些基本概念和理论的理解，又牵涉到使用一些具体的软件进行操作的问题。关于前者，我国读者已经读到一些著作，国内外的都有，但是怎样去具体实现呢？大多著作仅仅交待有哪些软件可以使用，常常给人一种“隔靴搔痒”的感觉，研究者读后仍然无从下手。

为了满足读者了解如何动手制作语料库的需要，自 2006 年以来，中国外语教育研究中心和外语教学与研究出版社连续四年举办了语料库与外语研究研修班，梁茂成、李文中、许家金几位博士都是主讲人。他们根据授课经验，收集和设计了许多案例和颇为有效的软件，精心编制了这本《语料库应用教程》，手把手地传授语料库创建和应用方面的方法和技术。此事可庆可贺，而此书则是一本值得向读者大力推荐的好书。这本书有什么好处呢？

1) 说理清楚。这本书强调的是怎样动手，但是知其然，还必须知其所以然。否则就会陷入盲目性。语料库语言学牵涉到的学科专业不少，例如文本分析、语体分析、计算机技术、统计学等。这本书删繁就简、用精取宏，把道理和概念都交代得清清楚楚。
2) 按部就班，循序渐进。要动手，必须讲究先后安排，那就是条理性的问题。这本书经过精心策划，安排得当，使学习者容易上手，然后逐步登堂入室。
3) 锐意求新，体现语料库语言学的新进展。比如，XML 语言是近些年来才得到发展和认可的语言，便于记录语料库中的结构化信息，而和它相联系、便于实现高效检索的正则表达式在语料库相关研究中正在得到广泛应用。这本书对怎样利用 XML 语言进行语料库建设、如何使用正则表达式进行高效检索等问题做了清楚而详尽的解释和说明，而现有的语料库教科书对此均讳莫如深。
4) 资源丰富。书中介绍了很多关于语料库语言学的资源，并附有作者自行编制的软件，为学习者建立语料库、利用语料库从事研究提供了极大的方便。

我接触语料库语言学已有 10 多年了，也编制过几个语料库，那些都是第一代的语料库。读了这本新书后，才感到自己的语料库知识大有更新的必要。这本书出版后，我将以首先获得它为荣。



2010 年 5 月

于广州

参考文献

Simpson, R. & J. Swales. 2001. Introduction: North American perspectives on corpus linguistics at the millennium. In R. Simpson & J. Swales (eds.). *Corpus Linguistics in North America*. Michigan: The University of Michigan Press.

Svartvik, J. 2007. Corpus linguistics 25+ years on. In R. Facchinetti (ed.). *Corpus Linguistics 25 Years On*. Amsterdam: Rodopi.

前　言

语料库语言学是一门年轻的学科。语料库应用是一种很有意义的活动。近年来，这种活动魅力四射，吸引了很多人参与。然而，语料库研究是一门交叉学科，融合了语言分析、语言教育、研究方法、统计学以及信息技术等多种学科；语料库研究中特别注重研究设计、程序和步骤的规范化，因此很多人或“虽不能至，心向往之”，或“心虽向往，力有未逮”。我们一直在想，能不能为语料库应用写这么一本书：它简易实用，直截了当，能让人边读边做，容易上手；尽量避免“学究”腔调和理论讨论，不做繁琐的考证和引用；内容基于实际案例，又具有充分的开放性，能让人举一反三，逐步深入；既能作为教学和培训教材，又能作为自学的入门读物。本教程就是在这样的思路下创作出来的。

自 2006 年以来，中国外语教育研究中心和外语教学与研究出版社连续四年主办语料库研究研修班，本书的作者都是主讲人。事实上，本书是作者根据研修班的教学实践以及个人研究经验，切磋合作的结晶。我们认为，本书有以下特征：

问题驱动：问题是一切研究的开始。本书不是一本语料库技术说明书，各章节都围绕具体的研究问题和目的展开。方法、工具、步骤及技术应用，都是为了回答研究问题、达到研究目的服务的，不能为了使用语料库而使用语料库。

基于实例：本书使用的案例都来自真实的研究实践，其中一些选自作者已正式发表的论文，一些则是作者正在思考和探索的话题。我们鼓励读者参与我们的探索，并与我们分享思路，开展自己的研究。当然，我们不能保证书中每个思路都能带来有价值的发现。

注重操作：本书不仅仅是用来“读”的，更重要的是用来“做”的。最好的做法是，根据本书具体章节中呈现的方法和步骤，真正动手操作。本书所附光盘提供了所有需要的资源，用来练习是足够了。

模块化设计：尽管本书在结构上各章节循序渐进、由易入难、相互参照，但其主要内容仍然为相互独立的模块。比如不需要建库的读者可以跳过相关内容，直接阅读索引分析或主题词分析章节。但在个别情况下，考虑到内容表述的连贯性与模块的完整性，部分技术环节在相应章节有所重述。

尽管本书强调应用和操作，避开理论上的讨论，但这并不意味着我们否认语料库语言学理论及其他语料库研究实践的重要性和价值。本书也并未包括所有的语料库应用，如平行语料库、自然语言处理等。“有不虞之誉，有求全之毁。”但我们仍然欢迎真诚的批评和建议。

本书通过酝酿、讨论和合作到最终成稿，有不少思路是在啤酒桌上形成的，其过程令人愉快。其中 1.1 节、1.2 节、2.4 节、3.1 节、3.2 节、5.1 节、5.4 节、6.1 节为梁茂成所撰写，他还负责统筹全书；2.1 节、2.3 节、3.4 节、5.3 节、6.2 节为李文中撰写；2.2 节、3.3 节、3.5 节、4.1 节、4.2 节、5.2 节、7.1 节、7.3.3 节、7.4 节为许家金撰写，并负责全书校对。

书中部分内容是在作者早先出版物的基础上改写的，这些出版物包括：

梁茂成，2009，词性赋码语料库的检索与正则表达式的编写，《中国外语教育》(2)。

梁茂成，2009，微型文本及其在外语教学中的应用，《外语电化教学》(3)。

陈国华、梁茂成、Adam Kilgarriff，2005，语料库和词典编纂的接口，《广东外语外贸大学学报（增刊)》。

李文中，2009，CIA 方法评析，《外语电化教学》(3)。

许家金、熊文新，2009，基于学习者英语语料的类联接研究：概念、方法及例析，《外语电化教学》(3)。

许家金，2009，词汇中心教学法的交际观：理论溯源与反思，《中国外语教育》(4)。

许家金，2010，从词语到话语：通过语料库开展话语研究，《中国英语教育》(1)。

本书的撰写得到以下基金资助：

- 国家社科基金项目“中国学生英语树库建设与研究”(09BYY033)
- 国家社科基金项目“基于语料库的英语中国本土化研究及应用”(07BYY022)
- 教育部人文社会科学研究项目“基于语料库的中国大学生英语口语话语特征研究”(08JC740002)
- 教育部人文社会科学重点研究基地北京外国语大学中国外语教育研究中心基金

最后，非常感谢外语教学与研究出版社高等英语教育出版分社社长常小玲女士，是她最初的提议和大力支持，使这本教程得以顺利成书。出版社的段长城、郑丹妮等编辑人员为书稿的编校工作付出了细致而专业的努力，在此我们也表达我们最衷心的感谢。

书中如有错误和问题，概由作者负责。

梁茂成
北京外国语大学中国外语教育研究中心

李文中
河南师范大学外国语学院

许家金
北京外国语大学中国外语教育研究中心
2010 年春

目录

第一部分　语料库语言学基本知识与语料库基本操作

第一章 语料库语言学基本知识

1.1 语料库语言学基本概念

本节主要介绍语料库语言学领域的部分常用术语和概念。鉴于本书的目的在于介绍语料库基本方法及其实际应用，我们并不试图为语料库语言学领域的概念和术语给出无懈可击的精确定义，也不会对众多学者在概念理解上的细微差异进行综述和甄别，更不会试图去澄清那些尚存争议的理论问题。我们将主要通过实例来介绍那些与实际操作密切相关的基本概念，为进一步介绍语料库应用扫清可能的障碍。

1.1.1 *语料库和语料库语言学*

语料库（corpus，复数为 corpora）一词来源于拉丁语，本意为 body。如今我们谈到语料库时，指的往往是一个“电子文本集”（a collection of texts stored in an electronic database）。一个小型文本集并不是真正意义上的语料库。真正意义上的语料库是一个按照一定的采样标准采集而来的、能够代表一种语言或者某语言的一种变体或文类的电子文本集。可以说，一个语料库由若干个电子文本构成，而这些电子文本作为一个整体可以代表某语言或者某语言的某种变体或文类。因此，以一个语料库为数据源（data source）进行的研究可以看作是对该语料库所代表语言、语言变体或文类的研究，研究所得到的结论可以推广到整个语言、语言变体或文类。

在一些人看来，语料库语言学（corpus linguistics）是一个独立的学科，它有自己独到的理论体系和操作方法。由于语料库语言学立足于大量真实的语言数据，对语料库做系统而穷尽的观察和概括所得到的结论对语言理论建设具有无可比拟的创新意义。而在另外一些研究者看来，语料库语言学并非语言学的又一个分支学科，而是一种研究方法，这种方法基于大量的真实语言，可以用来回答通过其他途径很难回答的问题，从而极大地丰富已有的研究方法。语料库语言学以大量精心采集而来的真实文本（authentic texts）为研究素材，主要通过概率统计的方法得出结论，因此语料库语言学从本质上讲是实证性的（empirical）。

1.1.2 语料库的主要类型

因研究目的的不同，语料库也有多种类型，以代表各种各样的语言、语言变体或文类。常见的语料库类型主要有：

- **通用语料库（general corpus）：**

  广泛采集某语言的口、笔语形式，取样时尽可能考虑口、笔语的主要社会变体、地域变体、行业变体等各种变异及语言使用的各种场合之间的平衡，力求最好地代表一种语言的全貌而建成的语料库。通用语料库一般较大，常常达到数亿词次，代表性的英语通用语料库有英国国家语料库（British National Corpus, BNC）、英语文库（Bank of English, BoE）、美国国家语料库(American National Corpus, ANC)等。通用语料库是描述语言全貌、编制工具书、核查语言用法等最理想的语料。此外，通用语料库还常常被用作参照语料库，以方便我们发现某些专门语料库的语言特点。

  由于通用语料库容量庞大，包含有多种不同属性的文本，我们常常可以对通用语料库进行分解，得到一个个专门用途的语料库。比如，我们可以从英国国家语料库中抽取所有的新闻语言文本，构成一个新闻英语语料库。

- **专用语料库（specialized corpus）：**

  又称专题语料库（special purpose corpus）。与通用语料库相反，出于某种特定的研究目的，人们常常只收集某特定领域(domain)的语料样本建成语料库，此类语料库称为专用语料库。在实际研究中，人们常常将专用语料库与通用语料库进行对比，来分析特定领域内语言的特点。此外，专用语料库也可作为编制专门领域工具书的理想语料。

- **共时语料库（synchronic corpus）：**

  由同一时代（主要是当代）的语言使用样本构成的语料库称为共时语料库。共时语料库是相对历时语料库而言的。基于不同时代的语言所建成的多个共时语料库可以构成一个历时语料库。

- **历时语料库（diachronic corpus）：**

  收集不同时代的语言使用样本构建而成的语料库称为历时语料库。历时语料库是观察和研究语言变化时常用的语料库。对历时语料库进行分解可以得到多个共时语料库。赫尔辛基英语文本语料库（Helsinki Corpus of English Texts）是一个典型的英语历时语料库。

- **口语语料库（spoken corpus）：**

  口语语料库常常包括由口语转写而来的文本，有时也包括语音文件。因为取样和转写的困难，口语语料库的文本容量很难达到笔语语料库的规模。将口语语料库与通用语料库进行对比，可以有效地发现口语特征。为了方便口语研究，人们常常对口语语料库中的语音、语调、停顿、重复、修正等口语特征进行标注。

● **笔语语料库（written corpus）：**

笔语语料库取材于书面语，常常包括书籍、报刊、书信、学术论文等常见笔语形式。由于笔语文本较容易收集，笔语语料库的容量一般较口语语料库的容量更大。

● **本族语者语料库（native speakers' corpus）：**

本族语者语料库中所收集的语言使用样本，全部源自于本族语者。本族语者语料库区别于非本族语者语料库和学习者语料库，在分析非本族语者或学习者语言使用特点时，经常以本族语者语料库作为参照。

● **学习者语料库（learner corpus）：**

由非本族语学习者语言使用样本构成的语料库。学习者语料库又可分为口语语料库和笔语语料库。国际上影响较大的学习者语料库有比利时学者 Sylviane Granger 等人于上世纪 90 年代初建立的英语学习者国际语料库（International Corpus of Learner English, ICLE）和鲁汶英语中介语国际数据库（Louvain International Database of Spoken English Interlanguage, LINDSEI）等。国内较有影响的学习者语料库有中国学习者英语语料库（Chinese Learners' English Corpus, CLEC）（桂诗春、杨惠中 2003）、中国学生口笔语语料库（Spoken and Written Corpus of Chinese Learners, SWECCL 1.0 & SWECCL 2.0）（文秋芳等 2005；文秋芳等 2008）、中国学习者英语口语语料库（College Learners' Spoken English Corpus, COLSEC）（卫乃兴等 2005）、中国大学生英汉汉英口笔译语料库（Parallel Corpus of Chinese EFL Learners, PACCEL）（文秋芳、王金铨 2008）、CEM（Corpus for English Majors）语料库（中国高校外语专业多语种语料库建设和研究项目组 2008）等。

● **单语语料库（monolingual corpus）：**

单语语料库中的语料来自于同一种语言，如英语语料库、汉语语料库等。

● **平行/双语语料库（parallel/bilingual corpus）和多语语料库（multilingual corpus）：**

双语 / 平行语料库中的语料来自于两种语言，而且相互对应，即一种语言是另外一种语言的译文。双语语料库建设中的一个重要环节是两种语言间的对齐（alignment）问题。目前，大多数双语语料库都进行了句子间的对齐，也有人尝试词语间的对齐和意义单位之间的对齐。双语语料库对翻译研究和机器翻译研究具有重要价值。北京外国语大学王克非教授主持建立的英汉双语平行语料库是国内较有影响的英汉汉英平行语料库。

多语语料库中的语言使用样本取自于多种语言。如 Europarl Parallel Corpus (European Parliament Proceedings Parallel Corpus) 收集了欧洲议会的多语言文集，将 11 种语言进行对齐处理，该语料库可以从网上免费下载。

1.1.3 文本

如上文所述，语料库是由大量的文本构成的，那么什么是文本呢？

文本（text）是一个十分有争议的术语。在语料库语言学中，文本常常可以理解为代表真实的连续话语（口语或笔语）的、以 ASCII 或 Unicode 呈现的、可以由计算机读取的电子文档。

一般说来，文本文件指纯文本（plain text）文件，可以用 Microsoft Windows 中的记事本程序打开，我们经常使用的 Microsoft Word 文档并非纯文本文件。由于 ASCII 格式的纯文本所能保存的信息较为有限，有时我们采用 Unicode 格式的纯文本来保存各种文字信息。由于普通纯文本文件难以记录语料中的各种结构化信息，有时我们采用可扩展标记语言（XML/Extensible Markup Language）记录语料中的结构化信息。

文本中可以保存生（raw）语料，即未经任何标注的语料，称为生文本（raw text）（如图 1.1 所示）。

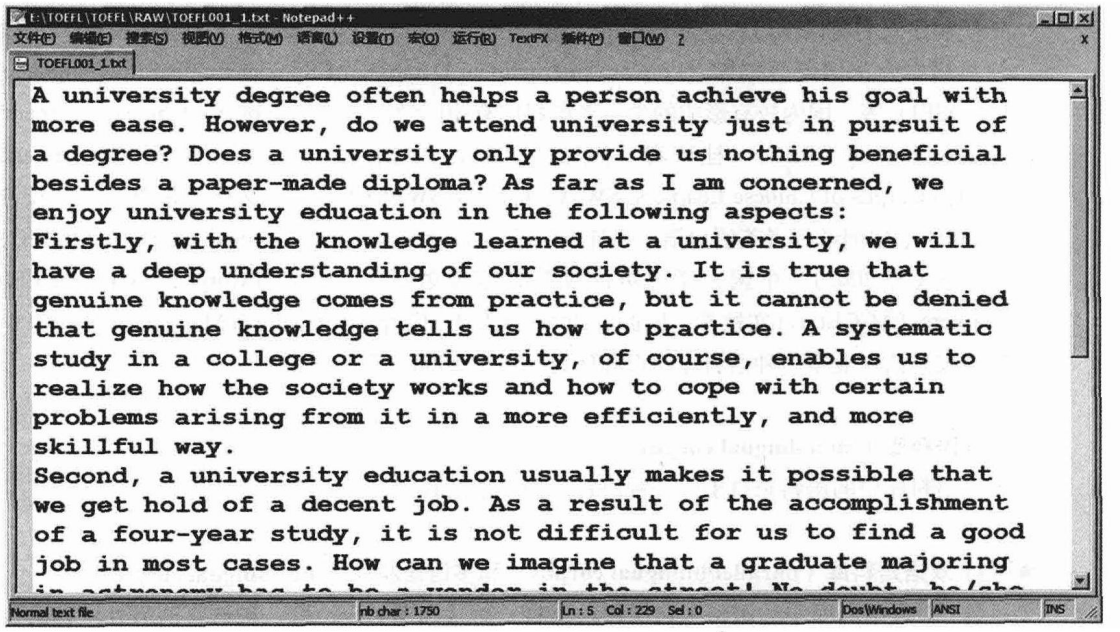


图 1.1　生文本

文本中也可以保存经过人工或自动标注的语料，这种文本称为标注文本（annotated text）。文本中的标注信息可以标示语料的来源、文本的内部结构、文本中的语言单位等多种语言信息和非语言信息。图 1.2 所示为经自动词性赋码后的文本，文本中每个词之后的下划线（"_"）是词与词性赋码之间的分隔符，而下划线之后的符号则标示词性信息（如，第一个词 A 之后的词性赋码 AT1 表示不定冠词。详见附录一）。