上海交通大学国际教育学院立项教材

汉语语料库应用教程

郭曙纶 编著

上海交通大学出版社
SHANGHAI JIAO TONG UNIVERSITY PRESS

上海交通大学国际教育学院立项教材

汉语语料库应用教程

郭曙纶 编著

上海交通大学出版社
SHANGHAI JIAO TONG UNIVERSITY PRESS

内 容 提 要

本教程首先简单介绍了语料库语言学的基本知识、汉语语料库建设的基本步骤及基本问题，然后着重介绍了汉语语料库应用的方方面面：有汉语的字频、词频、句频研究，有基于语料库的汉语词汇、语法研究，有基于语料库的小学语文教材、对外汉语教材研究，有基于语料库的对外汉语教学研究，有基于语料库的对外汉语教材编写、对外汉语词典编纂等。本教程为了方便文科读者的阅读，特别注意避免繁琐的计算与推导，而着重于汉语语料库应用中基本概念、基本方法及具体步骤的介绍。

在介绍汉语语料库应用时，有的章节侧重于具体操作的直观展现，以便读者能够按照书中图文并茂的介绍，自己一步步学会、掌握具体应用的操作方法（如第5章汉语语料库中的字频研究）；有的章节侧重于应用中概念、方法的介绍，以便读者能够把握这些概念、方法自己动手尝试新的具体应用（如第8章基于语料库的汉语词语搭配研究和第11章基于语料库的对外汉语教材研究）；有的章节侧重于案例的具体统计与分析，以便读者能够学会、掌握具体应用的写作方法（如第10章基于语料库的小学语文教材研究和第12章基于语料库的对外汉语教学研究）。

总之，本教程希望通过介绍汉语语料库应用的一些实例，能够开阔读者的眼界，树立起读者应用汉语语料库的信心，从而掌握汉语语料库应用的一些基本方法。

本教程可以作为汉语类研究生相关课程的主要教材或者参考书，也适合对汉语语料库具体应用感兴趣的大学生、研究生及相关领域的研究者阅读。

图书在版编目(CIP)数据

汉语语料库应用教程 / 郭曙纶编著. —上海：上海交通大学出版社，2013
（语言学文库）
ISBN 978-7-313-09489-6

Ⅰ.①汉… Ⅱ.①郭… Ⅲ.①汉语—语料库—教材
Ⅳ.①H1

中国版本图书馆 CIP 数据核字(2013)第030380号

汉语语料库应用教程
郭曙纶 编著
上海交通大学出版社出版发行
（上海市番禺路951号 邮政编码200030）
电话：64071208 出版人：韩建民
常熟市大宏印刷有限公司印刷 全国新华书店经销
开本：787 mm×960 mm 1/16 印张：9 字数：166千字
2013年3月第1版 2013年3月第1次印刷
印数：1～2030
ISBN 978-7-313-09489-6/H 定价：28.00元

FOREWORD 前言

从我2006年上半年开始给上海交通大学国际教育学院语言学及应用语言学专业的首届硕士研究生讲授《语料库建设及应用》这门课程算起，至今已经有8届硕士研究生听过这门课程了。本教程就是在《语料库建设及应用》课程讲义的基础上进行了较大修改后写成的。

这门课程主要是为计算语言学方向的研究生开设的。7年中也曾经有过几届其他方向的研究生选修这门课程。作为任课老师，最让我感到欣慰的是，有不少同学(包括一些其他方向的同学)把从这门课程中学到的知识和方法运用到后来的硕士学位论文的写作当中。根据同学们事后的反馈，他们认为在学习这门课程的过程中，受益最大的除了语料库应用的基本知识与方法之外，就是懂得了做事时必须认真细致地做好每一步工作，尤其是最基础的工作(比如语料库建设过程中的校对与标注工作)。前面的工作做好了，后面的工作自然就水到渠成了。

除了正式讲授这门课程，这些年还先后受邀到上海大学、上海师范大学、华东师范大学和上海外国语大学做过与汉语语料库相关的讲座，受到了一些好评。在讲座过程中，听众提出了语料库应用中的一些具体问题，对这些问题的解答与思考也成为本教程的重要内容，如第5章汉语语料库中的字频研究就是这样写成的。感谢上述高校的老师及同学们，是他们的需求与肯定促成了本教程目前的框架。当然还有一些问题的讨论未能及时写到本教程中，希望有机会在下次修订时进行补充(在正式修订前我会在完成后把它们作为

补充材料发布到自己的博客 http://gshulun.blog.163.com/当中）

本书的出版得到了上海交通大学国际教育学院 2011 年教材资助出版立项，对此感谢学院的领导以及院学术委员会的各位委员，感谢他们对本教材的肯定以及对本人的鼓励。

本书能够顺利地出版，还应该感谢上海交通大学出版社编辑管新潮老师，他为本书的出版付出了许多辛苦的汗水。

2012 年 11 月

CONTENTS 目 录

第1章 语料库语言学概述

1.1 语料库的定义

语料库(corpus,复数 corpora 或 corpuses)研究的出现与语料库语言学(corpus linguistics)的诞生是计算语言学(computational linguistics)与语言学发展的结果,也是信息社会的需要。

根据 2011 年出版的《语言学名词》[1],语料库就是"为语言研究和应用而收集的,在计算机中存储的语言材料,由自然出现的书面语或口语的样本汇集而成,用来代表特定的语言或语言变体。"由于电脑语料库容量大,资料真实,信息提取准确,因此,语言学家借助语料库可以从多方面、多层次描写语言并验证各种语言理论和假设,甚至建立新的语言模式和语言观。

语料库并非语篇的简单堆砌或集合,它应具有以下几个基本特征:① 样本代表性,② 规模有限性,③ 机读形式化。

语料库有不同的加工层次,加工的语料库一般指标有语言学标记的语料库。称未加工的语料库为"生语料库",加工过的语料库为"熟语料库"。使用标注正确率高的熟语料库更有利于对自然语言的研究。

随着语料库研究的发展,慢慢地产生了语料库语言学。

顾曰国[2]认为"语料库语言学"这个术语其实有两层含义。

> 一是利用语料库对语言的某个方面进行研究,也就是说"语料库语言学"不是一个新学科的名称,而仅仅反映了一个新的研究手段。二是依据语料库所反映出来的语言事实对现行语言学理论进行批判,提出新的观点或理论。只有在这个意义上"语料库语言学"才是一个新学科的名称。

不过从现有的文献来看,属于后一类的研究还比较少。本书所讨论的问题基本上属于前一类的研究。

“语料库语言学”是“语言学的一个分支。把大规模的真实的自然语言数据(书面文本或言语录音的转写)作为语言学描写、验证语言假说或建立语言学统计模型的依据。也是一种以语料库为基础的语言研究方法。包括：① 对自然语料进行加工、标注；② 应用已经标注好的语料或原始语料进行语言研究和应用开发。”[1] 由此可见，目前学界已经倾向于认为语料库语言学是一个新的学科，而不仅仅是一种研究方法。

1.2 语料库的类型

根据其选择的语料内容、选择的方式以及建设目的的不同，语料库的类型可以有不同的划分方法，比如通用语料库与专用语料库，同质语料库与异质语料库，动态语料库与静态语料库，共时语料库与历时语料库，第一代语料库与第二代语料库，书面语料库与口语语料库，等等。下面列出一些常见的语料库类型，并做简要说明。[1,3,4]

通用语料库(general corpus)：又称一般语料库，是文本的集合，为了保证收集的语料具有广泛的代表性，对语料采用系统的办法进行采集，用于事先未指定的语言学研究。通用语料库应有“平衡性”(balanced)，即语料库要收集不同类型、不同领域的包括口头的或书面的文本。通用语料库也有人称为系统语料库(systematic corpus)或平衡语料库(balanced corpus)，有时还被称为“核心语料库”(core corpus)。当然，严格说来，这些不同的名称之间还是存在差异的。

专用语料库(specialized corpus)：又称专门用途语料库(special purpose corpus)，指用于某种特殊研究的语料库。它又可分为方言语料库(dialect corpora)、区域性语料库(regional corpora)、非标准语料库(non-standard corpora)和初学者语料库(learner's corpora)等。它还可分为书面语料库(written corpora)和口语语料库(spoken corpora)。口语语料库是研究口语特征的重要工具，如语音语调的规律，其研究成果在语音合成中有重要应用。口语语料库的建设涉及口语真实语料的采集及语音转录，工作量极大。

异质语料库(heterogeneous corpus)：大量收集文字材料，尽可能广泛地接受各类材料而没有事先制定任何选材原则。收藏的文本在格式和内容上各异，而存储的格式和原来的出版物完全一样。

同质语料库(homogeneous corpus)：它是“异质语料库”的对立面。一般用于专业语料库。

动态语料库(dynamic corpora)：又称为监控语料库(monitor corpora)，用于观察现代语言的变迁。

与此相对的是静态语料库(static corpora)，只收集某一固定时期的共时语言

材料，语料库建成后，就不再扩充。

共时语料库(synchronic corpus)：指收集同一时代的语言使用样本构成的语料库。与此相对的是历时语料库(diachronic corpus)，指的是收集不同时代的语言使用样本构成的语料库。历时语料库主要用来观察和研究语言的历时变化，共时语料库则用来观察和研究某一时代的语言使用状况。对历时语料库的分解可以得到多个共时语料库。

平行/双语语料库(parallel/bilingual corpus)：把两种语言中完全对应的文本输入计算机，通过分析对比找出两者的对应关系，可用于机器翻译研究。近年来还出现了多语语料库(multilingual corpus)，如可以从网上免费下载的 Europarl Parallel Corpus(European Parliament Proceedings Parallel Corpus)就收集了多达11种欧洲议会的多语言文集。

第一代语料库指的是20世纪60年代到80年代所建成的一批语料库，这个阶段是以电子语料库的兴起为主要特征。第一代语料库规模相对比较小，大多只在百万词级。在这一阶段，语料库的发展以容量不断增加和种类的不断扩展为主要特征。

第二代语料库指的是从20世纪90年代中期开始建成的上亿词的大型语料库。

1.3 语料库的加工

语料库的加工可分为两个方面：

1) *语料库的标注*

标注就是使语料的某些单位(词、句、段落、篇章等)和表示对这些单位的某种层次的“理解”的知识信息(标记符)相关联。比如，汉语中的切词、词性标注、短语标注(树库标注)等。因此所谓标注其实就是加工者添加其对语料库中的字、词、句等的理解信息。

2) *语料库的知识获取*

指通过对语料库的处理，获得代表语料库中普遍现象的知识。它独立于语料库中某特定单位，反映了语言中的某种普遍规律。比如，组词造句的规律，具体可以表现为一些短语结构规则等。

1.3.1 语料库的加工层次

语料库有不同的加工层次。对语料库可以进行下列加工并形成不同加工层次的语料库：索引、主题标引、词的切分、词性标注、句法分析、语义分析、语用分析等。对语料库的加工还包括“预处理”。

语料库可以包含某个文本的全部，也可以从某个文本中抽取一部分构成。

下面简单介绍一下语料库加工的不同层次。

1）索引

逐词索引：提供在语料库中每个词指定词性每次出现的相关信息。逐词索引记录了每个词形在语料库中每次出现的相关位置，据此就可以提供每次出现的上下文信息。

关键词索引：提供出现指定关键词的文本、段落等信息。

就汉语而言，可以是以字为单位的逐字索引和关键字索引。

2）主题标引

主题标引是指对文本内容进行主题分析，赋予主题词标识的过程。

3）词的切分

词的切分就是从信息处理需要出发，按照特定的规范，对汉语按切词单位进行划分的过程。换句话说，就是将连续的字串按照一定的规范重新组合成词串的过程。

4）词性标注

词性标注就是对已经切词的语料中的每一个词赋予一个词性标记。词性标注与词的切分经常是由同一个系统来处理。词性标注的主要问题是兼类词的处理，还有一个问题是未登录词的处理。

5）句法成分标注（句法分析）

句法成分标注就是平时常说的树库加工，对已经标注了词性的文本标注上句法成分的信息，也就是标注上主语、宾语、谓语、定语、状语、补语等是什么，一般同时标注上这些句法成分是由什么样类型的短语（如名词短语、动词短语、形容词短语、介词短语等）充当的。

6）语义信息标注（语义分析）

语义信息标注，可以有不同的理解。一种是标注词义，一般在标注词性之后进行，给每个词语标注上词义信息，往往是义项标注，也就是通常所做的词义消歧。一种是语义角色标注，一般在句法成分标注之后进行，给每个句法成分标注上语义信息，如施事、受事等。

7）语用信息标注（语用分析）

语用信息标注，就是对文本标注上相关的语用信息，如话题、述题、话轮、省略成分等，为语用分析服务。它可以在生语料的基础上进行，也可以在熟语料的基础上进行。

8）特定语言模式的标注

特定语言模式的标注，就是根据研究需要，标注上研究者所需要的相关信息，如未登录词的标注、专有名词的标注、最大名词短语的标注等。

其实这就是说，研究者可以根据自己的研究需要进行几乎任意的语言信息标注。比如，可以标注一个句子的长度、一个句子的类型（包括句法类型和功能类型等），或者标注出一个句子的主要动词以及它的主语、宾语（或者施事、受事）、状语类型等。

1.3.2 语料库的标注原则

加工、标注语料库时应遵循一些基本的原则，对此，G. Leech 曾提出了有标记的语料库应满足的 7 条基本原则[3,5]：

1）所作标注可以删除，恢复到原始语料

这主要是为了保证语料的充分利用。原始生语料库的建设也需要花费大量的人力、物力和财力。只有保证原始语料的可恢复性，才能保证生语料库的复用性。因为语料库可用于不同的目的，可能需要采取不同的标注方法。

2）所作标注可以单独抽出，另处存储

这一原则实际上与第 1 条原则基本一致。由此可知，语料库中语料的标注应该最大限度地增加语料使用的灵活性。因此，如果有可能的话，最好把原始语料和标注信息分开存储在不同的文件中，然后通过专门的软件来进行阅读、编辑和管理。

3）语料库的最终使用者应该知道标注原则和标注符号的意义

因此，大多数语料库都配有详细介绍标注原则和标注符号意义的手册，供使用者参考。手册内容一般应该包括下列内容：

a）标注规范，即标注所用标准的描述和解释性文档。

b）记录标注者、标注地点和怎样标注的文档。

c）由于标注通常会出现差错、不一致或歧义现象，因此应当有关于标注质量的说明。例如，语料库的标注结果被校核到什么程度、它的精确率有多高（被判断为正确标注的百分比），以及标注的一致性达到什么程度等。

不过，实际上，目前公开的语料库中很少能见到有这么全的文档公开。

4）在语料的使用说明中，应该说明标注是什么人用什么方法做的

比如，是人工标注还是计算机标注、是一人标注还是多人标注。

5）应向用户声明，语料标注并非绝对无误，它只是一种可能有用的工具

不论是人工标注，还是计算机自动标注，还是两者的结合，都有可能产生标注的分歧甚至错误，因为标注的过程实际上是对语料中语言单位的特征进行解释的过程，不同的人可能会有不同的解释，同一个人在不同时期对同一语言单位的特征也可能会有不同的理解。

6）标注模式应不依赖于某一家之言，尽可能中立

在标注的过程中，为了方便语料库的使用，标注应该采用综合的、使用范围广

泛的语法理论，而不是采用某一特定的、使用范围狭窄的语法理论。

7）任何标注模式都不能作为第一标准，即使有，也只能通过实践在大量比较中得到

目前，世界上还没有一种被普遍接受的标注模式。

这7条原则，概括起来就是最大可能地方便加工者和使用者。“语料的标注和语料的利用是一对矛盾。从用户的角度，语料标注得越详尽越好，而标注者则还需考虑标注的可行性。因此，任何标注模式都是二者之间求得的一种妥协的产物。”[5]

1.3.3 语料库的加工技术

在语料库加工的过程中，运用到的主要技术手段包括：

① n-gram模型；② 马尔可夫模型；③ 概率上下文无关文法模型；④ 统计机器翻译模型；⑤ 互信息；⑥ 熵；⑦ 聚类；⑧ 共现统计；⑨ 分类；⑩ 平滑方法（解决数据稀疏）；⑪ EM参数估计方法；⑫ 韦特比（Viterbi）参数估计方法；⑬ 动态规划求解最优解的方法；⑭ 有限状态自动机理论和模型。

这些技术手段的具体操作都需要比较专门的知识与技术，感兴趣的读者可以阅读姚天顺等的《自然语言理解——一种让机器懂得人类语言的研究（第2版）》[6]、王建新的《计算机语料库的建设与应用》[7]、Christopher D. Manning和Hinrich Schütze的《统计自然语言处理基础》[8]、Daniel Jurafsky和James H. Martin的《自然语言处理综论》[9]、宗成庆的《统计自然语言处理》[10]等，尤其是后者提供了许多技术细节的讨论。

语料库加工的主要困难有三个方面：一是数据稀疏问题，二是歧义问题，三是语言模型本身的精确度问题。围绕这些问题的详细讨论，也请参考上述著作。

1.4 语料库的应用

语料库在语言学研究的作用是多方面的，最主要的就是提供丰富多样的语言实例，然而其应用却远不止于此。这里先简单地提及一些，具体的应用将会在后面的章节中展开。

1.4.1 频率统计

频率统计主要分为字频统计和词频统计两个方面。

早在语料库概念还没有产生之前，在我国就已经有学者通过语料库统计的方法来研究汉字的频率，其目的在于研制基础汉字的字表。著名教育学家陈鹤琴统计了6种包含554 478个汉字的语料，得到不同汉字4 261个，在此基础上，编写了

《语体文应用字汇》，于1925年完成，于1928年由商务印书馆出版。[11]正式应用语料库技术来统计汉语字频的成果有贝贵琴、张学涛汇编的《汉字频度统计——速成识读优选表》[12]和国家语言文字工作委员会、国家标准局编的《现代汉语字频统计表》[13]。《汉字频度统计——速成识读优选表》通过对不同出版物的2 100多万字的用字统计，用数字揭示了每个汉字的使用频度（按常用程度分级排队）及笔画规律。这是我们目前所见中国最早的利用计算机来大规模统计汉字得到的字频统计结果。《现代汉语字频统计表》则是专门为了统计字频而进行的研究成果，该研究成果是从1977年到1982年间社会科学和自然科学的1.38亿字的材料中抽样1 108万余字利用计算机进行统计的，共有13个汉字频度表，分别按降频次序和汉语拼音字母次序排列。

词频统计方面的成果有刘源编的《现代汉语常用词词频词典（音序部分）》[14]。

目前网络上还能找到不少可以免费下载的字频和词频数据。

此外也有对现代汉语基本句型进行频率统计的[15]。

1.4.2 词汇研究

语料库可以为语言研究者提供大量真实准确的例句。这方面的研究有很多，比如为词典编纂家提供实例、为研究某个句法现象提供实例等。这样可以更为准确地把握某个词语的语义及用法。

1.4.3 语言教学

语料库中的语料是人们实际运用的语言，所有的材料均取自真实的书面语和口语文本，提供的是语言实际使用的客观例证。对这种材料进行分析，有时可以发现现有的语言教学材料中存在的问题。因此汉语语料库在语言教学中可以有两个方面的应用：一是为学习者提供丰富的汉语学习实例，二是从汉语学习者语料（汉语中介语语料库）中发现教学中存在的一些问题或者需要注意的方面。

思考与练习

1. 请参考相关文献，谈谈你对语料库的认识。
2. 请选择一个感兴趣的应用领域，查找相关文献，简介一下语料库的应用。