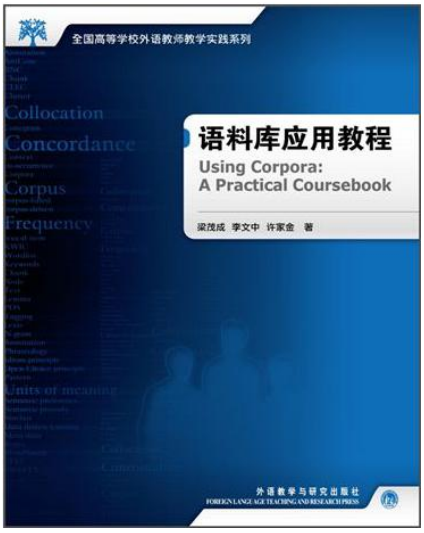

导读书目：《语料库应用教程》

作者：梁茂成、李文中、许家金

出版社：外语教学与研究出版社

出版时间：2010 年 7 月 1 日

专家导读

尊敬的读者，

语料库语言学既有其学科内部的自有体系，又可作为研究工具应用于语言研究的各个方面。然而，其看似复杂的相关技术以及规范化的操作步骤令许多研究者望之却步。本书是一本适合初学者阅读的实用教程，可操作性较强，对语料库建库、检索、正则表达式、词表生成、主题词表生成等语料库基本技术有较为详细的介绍，并附有具体清晰的操作步骤，适合读者边读边做。同时，各操作均围绕具体的研究问题和目的展开，所用案例来自真实的研究实践，向读者呈现了语料库在外语教学、词汇分析、话语研究等多领域的实际应用。除此之外，书中还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便利。

本书分为三大部分。第一部分主要介绍语料库语言学的基本知识与语料库的基本操作。包括索引、搭配、类联接、多词序列、语义韵等基本概念，文本采集、整理、标注、分词等建库步骤，以及检索、词表、主题词等基本技术。第二部分

以具体实例介绍了语料库在外语教学与外语学习中的应用。第三部分探讨了语料库在外语研究中的应用。本次活动将分两阶段进行，第一阶段将重点阅读第一部分中的基本概念与操作，第二阶段将集中在第二、三部分中的语料库应用实例。祝大家阅读与操作愉快！

刘国兵、吉洁

第一阶段活动安排 8 月 1 日—8 月 15 日

〕阅读章节

1. 语料库语言学基本知识
2. 文本采集与加工
3. 语料库基本技术

〕 章节要点

1. 第一章 语料库语言学基本知识

本章重点介绍语料库语言学的基本概念，可大致归为以下五组：（1）语料库（corpus）和语料库语言学（corpus linguistics）：前者可视为具有特定标准的文本集合，后者为基于前者的一门独立学科。（2）文本（text）和标注（annotation）：文本分为生文本（raw text）和标注文本（annotated text），后者可被附加元信息（metadata）、词性（POS）、句法（parsing）等多种标注。（3）形符（token）、类符（type）、类符/形符比（TTR）、频率（frequency）和频数（frequencies）：其中前三者针对全文本的容量而言，后两者针对某一词（组）在文本中的概率而言。（4）索引/关键词/节点（concordance/KWIC/node）、索引行（concordance lines）和正则表达式（regular expression）：作为语料库最基本的操作技术之一，对索引工具的使用及索引行的分析将是第三章中的重要内容，而正则表达式堪称语料检索的得力助手。（5）搭配（collocation）、类联接（colligation）和语义韵（semantic prosody）：三者像是针对节点词周边环境的三级窥镜，分管词语、句法和语义三

个层面。

本章思考题：

（1）形符与类符的区别？类型比（TTR）与标准类型比（STTR）的区别？频率和频数的区别？

（2）能否举出一些例子，来展示搭配、类联接与语义韵的区别？

2. 第二章 文本采集与加工

如果找不到合适的已有语料库，研究者通常需要自建一个语料库。本章介绍了建库的四大步骤：（1）文本采集：需要考虑语料来源（如 word, pdf, html 等）并选用相应的采集方式，确定语料格式（ANSI 或 UTF-8），制作备检文件等。（2）文本整理：采集好的文本通常包含一些非法格式，例如全角符号、中文标点、缺少空格、不正常断行等，这些会影响到后续的检索与标注，需要作相应的整理。（3）分词（tokenization）与词形还原（lemmatization）：自建的中英文语料库都需要进行分词，一般可直接通过自动分词软件完成，而词形还原仅针对特定研究。（4）标注（annotation）：一般自建语料库可进行基本的元信息与词性标注即可，前者利于语料分类对比，后者便于使用正则表达式进行检索。

本章操作题：

（3）请分别从 html 网页（如新闻）和 pdf 书籍（如学术）中选取部分语料，建立两个 txt 文档，从中找出这两种语料来源中常见的“非法”格式，并尝试进行整理。

（4）请根据 2.4.3 中介绍，分别使用 CLAWS4 网络试用服务及 TreeTagger 软件，对你刚才建立的语料进行词性赋码，并试对比这两种软件的赋码异同。

3. 第三章 语料库基本技术

本章主要介绍了语料库操作的三组基本技术：（1）检索与索引行分析：通过常用的 AntConc 或 PowerGREP 软件可以轻松实现对某一词（组）的检索，若借助正则表达式或 PatternBuilder 软件，还可实现对词性或句式结构的检索。检索后可输出包含该节点词（组）的索引行，并可在设定的跨距内对其前后语境进行观察。（2）词表（word list）及主题词表（keyword list）的生成：前者为单个语

料库的词频列表，后者为两个语料库（观察语料库及参照语料库）的词表对比。通过 AntConc 软件可实现这两种词表的生成，其应用将在第五章中介绍。(3) 数据统计：包括标准化（normalization）、差异检查（如卡方 chi-square 和对数似然比 log-likelihood）、搭配强度计算等。

本章操作题：

（5）通过上一题中对 CLAWS 和 treetagger 词性赋码的观察，尝试编写一个简单的正则表达式，在你之前建立的新闻和学术语料中，使用 AntConc 检索出所有的情态动词，并观察这两个语料的情态动词使用是否有不同之处（由于语料大小及取样因素，也许结果并无特别发现，仅练习检索过程即可）。

（6）使用 AntConc，分别生成自建新闻和学术这两个文档的词表，并尝试将两者对比生成主题词表（通常情况下参照语料库为通用语料，但此处仅作练习，不做细究。）