目　录

第一章　语料库语言学简介

【问题导入】

以下是一线英语教师和学生初学语料库时的反馈语截图。

1　本来对语料库的使用还比较陌生
2　对语料库的了解不透彻
3　语料库并不是自己想象的那样，很多骨干老师都知道
4　首次接触了语料库资源和技术这一概念
5　之前有接触过语料库这个概念
6　上这样的课一定要先对语料库有个简明的整体介绍
7　我第一次了解语料库
8　我们曾经很好奇的视频音频处理及语料库的应用
9　我是第一次这么直观地了解语料库的运用
10　以前只对语料库有最最基本的了解
11　一直觉得语料库好繁复
12　语料库并不是第一次接触
13　语料库是什么
14　语料库对于我是一个新事物
15　这是我第一次接触到语料库
16　以前一直认为语料库就是写论文的时候才用得上的“高大上”的一个工具
17　对语料库产生了兴趣
18　对语料库感到很兴奋

【要点指引】

以上是一批中学及高校教师在初涉语料库时的想法和困惑。多数人认为语料库是一个新生事物，对其既有好奇心、感兴趣，又觉得太“高大上”，难以把握。由此可见，在当今信息社会，计算机科学和教育技术虽然为外语教师带来了前所未有的教学理念与手段更新，但是最早体现大数据时代理念的语料库在我国英语教育教学领域的应用还存在探索不足的问题。本章将通过一系列学习和体验活动，达成以下学习目标。

【学习目标】

1. 初步了解语料库语言学的定义及学科发展沿革；
2. 了解和观察各种类型语料库样品及其体例特征。

第一节　学科定义

【活动 1.1】什么是语料库和语料库语言学？

步骤一：阅读以下两组文献语段中多位语料库语言学家对语料库和语料库语言学定义的阐述。注意表述中反复出现的词和短语，看能否捕捉到语料库和语料库语言学的定义，包括语料的特点、规模、来源、储存、提取手段以及用途等等。

1. 关于语料库

- [A corpus is] a collection of naturally-occurring language text, chosen to characterize a state or variety of a language.

  The whole point of assembling a corpus is to gather data in quantity.

  (Sinclair, 1995: 21, 1999: 171)
- A corpus is a body of texts assembled according to explicit design criteria for a specific purpose.

  (Atkins & Clear, 1992: 5)
- [A corpus is] a body of naturally-occurring (authentic) language data which can be used as a basis for linguistic research.

  In the past thirty-five years, the term corpus has been increasingly applied to a body of language material which exists in electronic form, and which may be processed by computers for various purposes.

  (Leech, 1997: 1)
- [A corpus is] a collection of texts, especially if complete and self-contained: the corpus of Anglo-Saxon verse.
- [A corpus is] a body of texts, utterances, or other specimens considered more or less representative of a language, and usually stored as an electronic database.

  Currently, computer corpora may store many millions of running words, whose features can be analysed by means of tagging (the addition of identifying and classifying tags to words and other formations) and the use of concordancing programs. Corpus linguistics studies data in any such corpus.

  Bodies of natural language material (whole texts, samples from texts, or sometimes just unconnected sentences), which are stored in machine-readable form. Computer corpora are rarely haphazard collections of textual material: they are generally assembled with particular purposes in mind, and are often assumed to be (informally speaking) representative of some language or text type.

  (McArthur & McArthur, 1992: 266)

- A corpus is a collection of pieces of language that are selected and ordered according to explicit linguistic criteria in order to be used as a sample of the language.
  (Sinclair, 1996: 3)
- ... a corpus ... might ... be described as a finite-sized body of machine-readable texts, sampled in order to be maximally representative of the language variety under consideration.
  (McEnery & Wilson, 2001: 30, 32)
- [A corpus is a] collection of linguistic data, either written texts or a transcription of recorded speech, which can be used as a starting-point of linguistic description or as a means of verifying hypotheses about a language (corpus linguistics).
  (Crystal, 2008: 117)
- A corpus is a principled collection of authentic texts stored electronically that can be used to discover information about language that may not have been noticed through intuition alone.
  (Bennet, 2010: 12)

2．关于语料库语言学

- Corpus linguistics is not an end in itself but is one source of evidence for improving descriptions of the structure and use of languages, and for various applications, including the processing of natural language by machine and understanding how to learn or teach a language.
  (Kennedy, 2000: 1)
- What, then, is a “corpus linguist”? I would like to think that it is a linguist who tries to understand language, and behind language the mind, by carefully observing extensive natural samples of it and then, with insight and imagination, constructing plausible understandings that encompass and explain those observations. Anyone who is not a corpus linguist in this sense is, in my opinion, missing much that is relevant to the linguistic enterprise.
  (Chafe, 1992: 96)
- Corpus linguistics approaches the study of language in use through corpora (singular: corpus).
  ... corpus linguistics serves to answer two research questions:
  1. What particular patterns are associated with lexical or grammatical features?
  2. How do these patterns differ within varieties and registers?
  (Bennet, 2010: 2)

步骤二：将上述文献（1 和 2）的部分语段合并成可由语料库检索工具查阅的微本语料（见本书语料样品 Minifiles\corp + definition.txt），可获得如图 1-1 所示的检索界面。观察图中各行 corpus 或 corpus linguistics 后面的文字内容，看是否能归纳出语料库的基本定义和典

型特征。反思这种阅读方式与步骤一有何不同、二者分别有何优势。

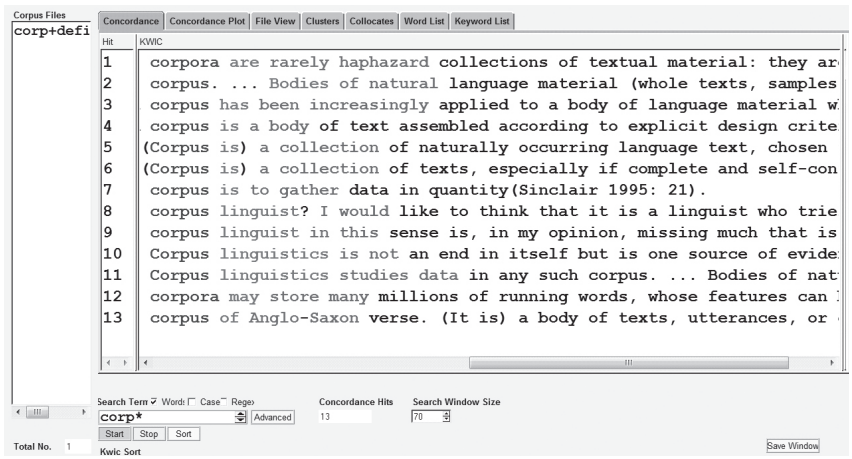


图 1-1　源自文献 1 和 2 部分语段的微本语料检索截图

步骤三：使用 Adobe Acrobat Reader 9 的高级检索工具（Search，如图 1-2 所示）可以聚焦性阅读有关语料库书籍的 PDF 版文档，如本例中 *From Corpus to Classroom*（O'Keeffe 等，2007）一书部分章节中所有带 corpus is 的语境共现行（见图 1-3），也可以点击其中任意一行直接进入该行出现在书中的上下文，进行更大语境范围内的阅读（见图 1-4），看能否找到关于语料库定义的表述。

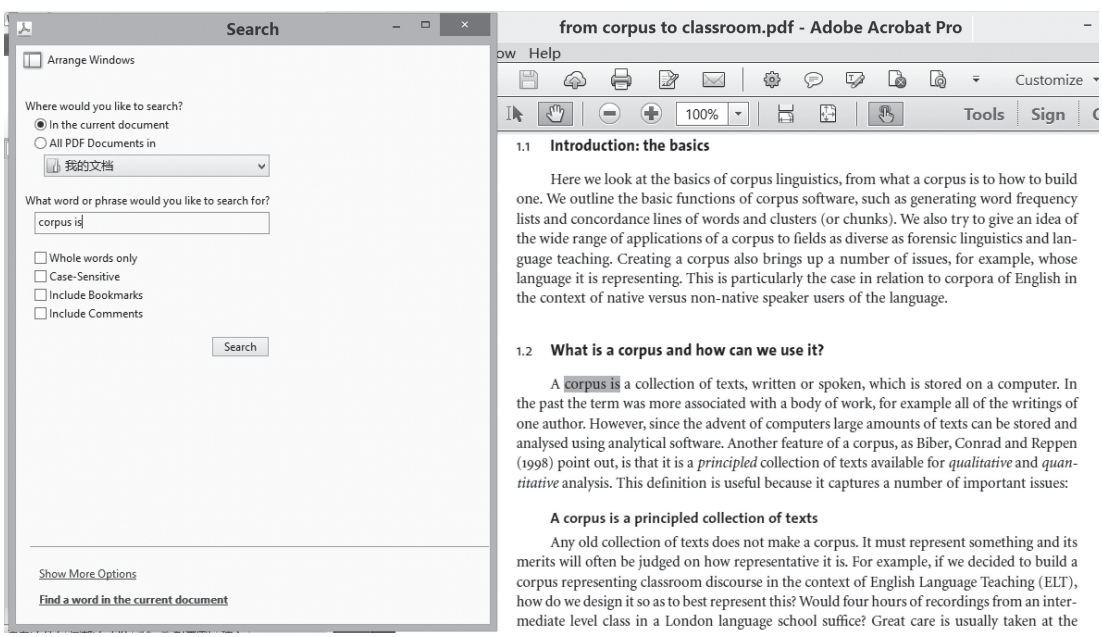


图 1-2　Adobe Acrobat 的高级搜索框

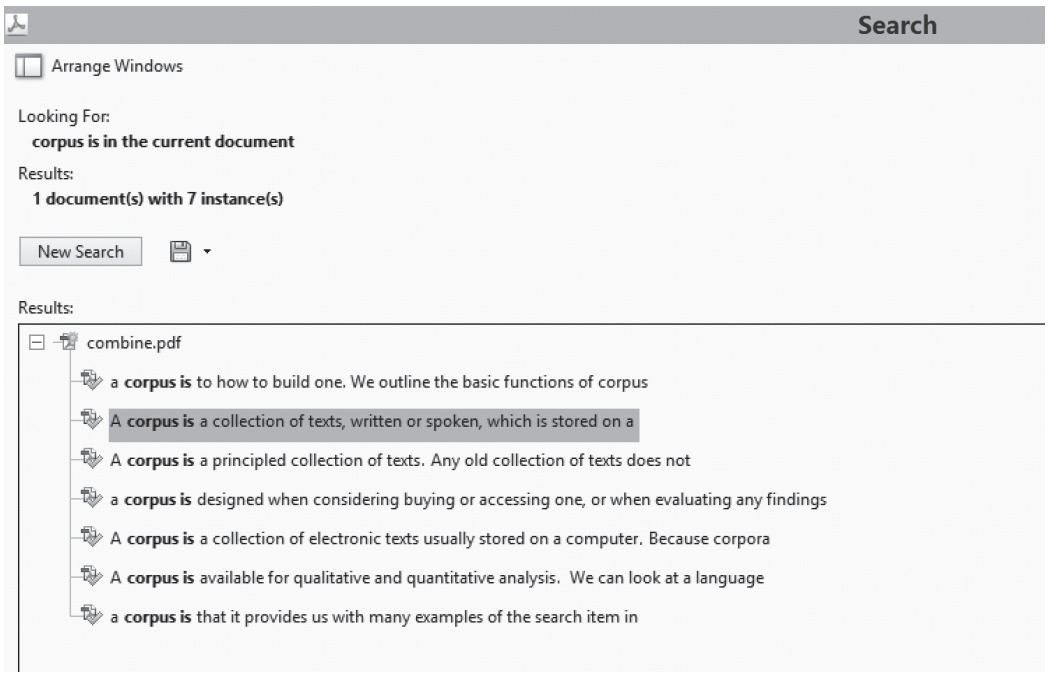
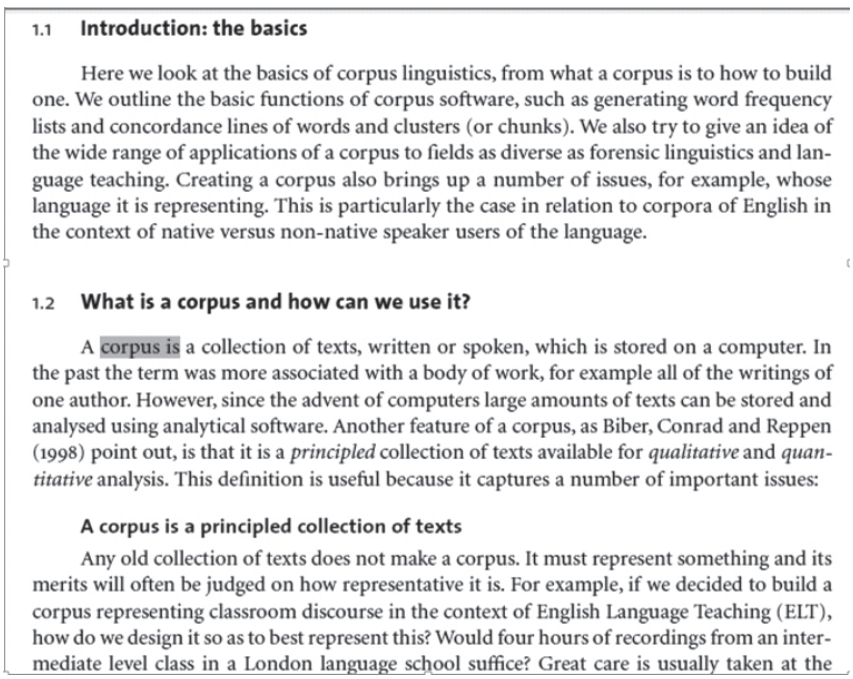


图 1-3　PDF 展示的批量 corpus is 语境共现行



1.1 **Introduction: the basics**

Here we look at the basics of corpus linguistics, from what a corpus is to how to build one. We outline the basic functions of corpus software, such as generating word frequency lists and concordance lines of words and clusters (or chunks). We also try to give an idea of the wide range of applications of a corpus to fields as diverse as forensic linguistics and language teaching. Creating a corpus also brings up a number of issues, for example, whose language it is representing. This is particularly the case in relation to corpora of English in the context of native versus non-native speaker users of the language.

1.2 **What is a corpus and how can we use it?**

A corpus is a collection of texts, written or spoken, which is stored on a computer. In the past the term was more associated with a body of work, for example all of the writings of one author. However, since the advent of computers large amounts of texts can be stored and analysed using analytical software. Another feature of a corpus, as Biber, Conrad and Reppen (1998) point out, is that it is a *principled* collection of texts available for *qualitative* and *quantitative* analysis. This definition is useful because it captures a number of important issues:

A corpus is a principled collection of texts

Any old collection of texts does not make a corpus. It must represent something and its merits will often be judged on how representative it is. For example, if we decided to build a corpus representing classroom discourse in the context of English Language Teaching (ELT), how do we design it so as to best represent this? Would four hours of recordings from an intermediate level class in a London language school suffice? Great care is usually taken at the

图 1-4　点击图 1-3 第 2 行打开该书章节的相关语段

【活动 1.1 述评】

完成上述三个步骤后，可发现语料库（corpus）有多个角度的界定。首先，无论从哪个步骤的文本展示都可以看出，行文中总有一些反复出现的、意义相同或相近的词和短语，如 a body of、a collection of、language texts、samples、authentic、naturally-occurring、representative、computer、electronic 等等，由此可归纳出语料库的四大特征：一是语言信息量巨大；二是语料来源真实且具代表性；三是要靠计算机来储存和检索；四是用于语言调查和分析。这些特点综合起来便可将语料库界定为：一个由大批量真实使用的语言信息汇集而成、用计算机储存和提取并用作语言调查和分析的巨型语言资源库。这样巨大的资源库如今已经成为大数据、信息爆炸时代和学习型社会中不可或缺的语言研究及语言教学的资源和手段。

再看语料库语言学（corpus linguistics），它是当代语言学的一门新兴学科。其特点是运用计算机对大型语言资源库进行高速检索，并准确展示语言使用的批量实例，从而揭示语言使用的倾向性规律及其所传递的意义、功能乃至思想意识。由于语料库是在一定原则下收集的批量语言素材并且以电子版本的形式储存在电脑中，所以它可以用来对语言进行量化调查和质性分析，从而深化对语言结构和语言使用的描述并且开拓多个方面的研究与应用，包括大规模语言调查、语言教育教学和人工智能语言研究等等。许多语料库语言学者的语言观是将语言视为一种社会行为和做事方式，故将实际运用中的语言作为研究对象，通过计算机技术对海量的语言事实进行调查和分析，从而获得对语言的本质和使用规律更深刻、更全面的认识。语料库语言学反映了经验主义和实证研究相结合的语言哲学思想。

对本活动展示的三种阅读方式还可作进一步反思。步骤二和步骤三均与步骤一有所不同，前两者的特点是搜索关键词的语境进行聚焦性的阅读，本质上是训练学生“跳读”、“选读”或“找读”的阅读微技能，而且阅读界面也从以往的从左到右转变为从上到下。这是语料库语言学带来的一种能够高速捕捉关键信息的聚焦性阅读方式，具有革命性的意义。

第二节　发展简史

【活动 1.2】了解语料库语言学发展史

步骤：阅读不同时期语料库语言学学者对该学科发展沿革的归纳性陈述。[1]

1　编者注：陈述中个别文字格式稍有改动。

1. Aijmer & Altenberg（1991: 1）谈三大经典语料库的起始

[T]here has been a rapid development of corpus linguistics in the last three decades. This development stems from two important events which took place around 1960. One was Randolph Quirk's launching of the Survey of English Usage (SEU) with the aim of collecting a large and stylistically varied corpus as the basis for a systematic description of spoken and written English. The other was the advent of computers which made it possible to store, scan and classify large masses of material. The first machine-readable corpus was compiled by Nelson Francis and Henry Kučera at Brown University in [the] early 1960s. It was soon followed by others, notably the Lancaster-Oslo / Bergen (LOB) Corpus, which utilized the same format as the Brown Corpus and made it possible to compare different varieties of English. In 1975 Jan Svartvik and his colleagues at Lund University undertook the task of making the spoken part of SEU corpus available in machine-readable form. The result [is] the London-Lund Corpus (LLC).

2. Renouf（2007: 28）谈语料库发展的五个阶段

- 1960s onwards: the one-million word (or less) Small Corpus
  - — standard
  - — general and specialized
  - — sampled
  - — multi-modal, multi-dimensional
- 1980s onwards: the multi-million word Large Corpus
  - — general and specialized
  - — sampled
  - — multi-modal, multi-dimensional
- 1990s onwards: the "Modern Diachronic" Corpus
  - — dynamic, open-ended, chronological data flow
- 1998 onwards: the Web as corpus
  - — web texts as source of linguistic information
- 2005 onwards:
  - — the Grid
  - — pathway to distributed corpora
  - — consolidation of existing corpus types

3. McCarthy & O'Keeffe（2010: 4-6）谈计算机技术推动语料库发展

- The first computer-generated concordances had appeared in the late 1950s, using punched-card

technology for storage.

- [F]rom as early as 1970, library and information scientists had developed a keen interest in KEY Word In Context (KWIC) concordances as a way of replacing catalogue indexing cards and of automating subject analysis.
- It was the revolution of hardware and software in the 1980s and 1990s which really allowed corpus linguistics as we know it to emerge.
- [T]he seemingly unstoppable increases in desktop computing power in the 1990s enable small team[s] and individuals to take on quite ambitious corpus projects. The parallel growth of the internet and fast download speed meant that data and results could be transferred easily from scholar to scholar.
- Technology also enabled the creation of multi-modal corpora, in which various communicative modes (e.g. body-language, writing) could all be part of the corpus, all linked by simple technologies such as time-stamping and all accessible at one go.

【活动 1.2 述评】

语料库语言学作为语言学研究的一门新兴学科，起源于 20 世纪 60 年代的英美国家。英语语料库语言学的发展历史有以下几个里程碑：

（1）从 20 世纪 60 年代起，小型语料库的规模通常是 100 万词（或者更少）。如美国英语版的 The Brown Corpus 和英国英语版的 The LOB Corpus 笔语语料库以及后来的 The London-Lund Corpus of Spoken English (LLC)，以上三者可称为早期三大经典语料库。

（2）从 20 世纪 80 年代起，大型语料库的规模是以往的数十倍，如 730 万词的 COBUILD 语料库已发展成 6.5 亿词的 BoE 语料库。

（3）从 20 世纪 90 年代末起，动态语料库的特点之一是对早期语料库实行后期的内容扩充，如 90 年代的 Frown 和 F-LOB；特点之二是建立开放性的、滚动式发展的历时语料库，如截至 2015 年一直在扩展的英国《独立报》报刊语料库（*The Independent* Corpus）。

（4）从 2005 年起，网络语料库的特点是在互联网上设置检索引擎，将互联网上的语言信息作为一个巨大的、动态的、开放的语料库，如 WebCorp、杨百翰大学 Mark Davies 教授开发的 COCA 等大型免费在线语料库（Davies, 2010），其中最具影响力的 COCA 语料库自 1990 年至 2019 年每年扩充超过 2500 万词，新增的文本取自口语、小说、流行杂志、报纸、学术文本、电视和电影字幕、博客和其他网页，这使得 COCA 语料库成为唯一一个大规模历时平衡的美国英语语料库。

（5）自 21 世纪以来，多模态语料库的特点是融视频、录音、图像和语料检索为一体，可用于语言研究和语言教学，如美国密歇根大学的 MICASE、欧盟投资建设的七国青少年外语学习平台 SACODEYL、杨百翰大学 Mark Davies 教授 2018 年建成的总词数超

过百亿的 iWeb 大数据语料库。

推动语料库发展的动因至少有以下三点。一是自 20 世纪中叶以来，计算机科学的迅猛发展带动了大规模语料库的收集和检索；语料库工具的使用从大型主机（mainframe）发展到手提电脑。二是语言社会使用的需要，如大数据语言分析调查和语言教学的需求。三是现代多媒体信息技术的涌现。总而言之，语料库语言学的发展反映了人类对知识的渴望和对语言使用的需求，也折射出现代科学技术的强大推动。

国内的语料库语言学起步于 20 世纪 80 年代，当时上海交通大学建立了国内首个百万词的科技英语语料库 JDEST（Yang, 1986）。进入 21 世纪以来，语料库语言学在国内逐步推开，近年发展尤为迅猛，并呈现出以下特点：

(1) 注重建设外语学习者的中介语语料库。如先后建成并发行流通的中国学习者英语语料库（CLEC）（桂诗春、杨惠中，2003）、中国学生英语口笔语语料库（SWECCL）（文秋芳等，2005，2009）等。

(2) 注重建设汉语语料库以及汉语与外语匹配的双语或平行语料库。例如，国家语委现代汉语语料库[1]、中国英汉平行语料库（王克非，2004）、中日对译语料库（徐一平，2002）等等。

(3) 注重建设外语教学语料库。例如，汇集中小学英语教材、课堂教学与学生口笔语语料的中学英语教育语料库（光盘）（华南师范大学外语系，2000）；“一针三库”智能教研平台（https://languagedata.net）（金檀等，2023）；基于各语域或各专业学科的学术或教学语料库如高中英语教材语料库（何安平、郑旺全，2009）、英汉医学平行语料库（管新潮等，2011）等等。

展望未来，语料库的发展方向则很可能是由后互联网时代网络技术支持的，更为即时、同步、多模态的以及全球或区域性整合型的巨量语料库。

第三节　语料库类型

【活动 1.3】用语料库检索工具展示各类语料库文本

语料：本书语料样品 Minifiles 目录下的各种微本[2]（除特别说明，这些样品均源自华南师范大学外国语言文化学院自 1997 年以来建设的英语教育教学语料库，此后全书不再重复

1　可访问 http://www.china-language.edu.cn/#/languageResources/corpus。

2　在外语教学与研究出版社“高等英语教学网”（https://heep.fltrp.com）首页的“教材支持”内查找本书——《语料库辅助英语教学入门》（第二版），即可免费下载本书语料样品 Minifiles。全书同。

说明）。

工具：AntConc 3.2.1w（2007）（除特别说明，本书的语料库截图均来自于该版本工具，此后全书不再重复说明）。

步骤：观察以下用语料库检索工具打开的各类语料库文本（见图 1-5 至图 1-12），尝试能否分辨出各种类型特征。例如，是笔语语料还是口语语料，是原始语料还是附码语料，是英语本族语者语料还是英语学习者语料，等等。

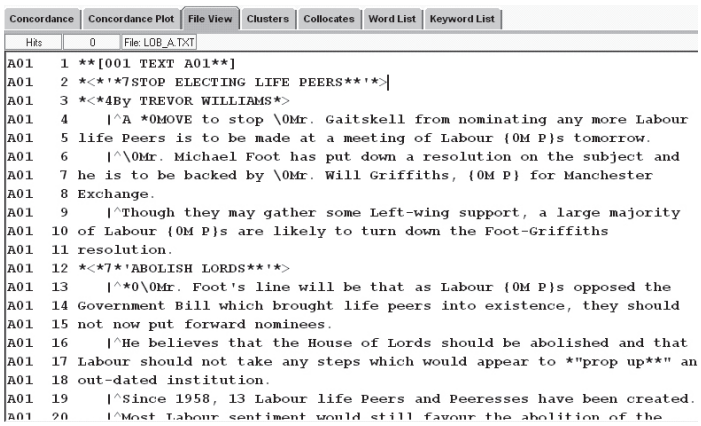


图 1-5　原始语料库片段

观察指引

图 1-5 是 100 万词的英国笔语语料库（The LOB Corpus，1978）片段。其中 A01 栏表示语料类型及文本号，如 A01 表示该语料库中新闻类语料的第一号样品。A01 之后的阿拉伯数字为该样品的行号序列，如第 1 行表示文本序号 A01；第 2 行为该文本的标题；第 3 行为该文本的作者；从第 4 行开始是文本正文，其中 | 号表示段落开头、^ 号为句子开头，等等（详见 Johansson 等 , 1978: 12；何安平，2004a: 141-143）。

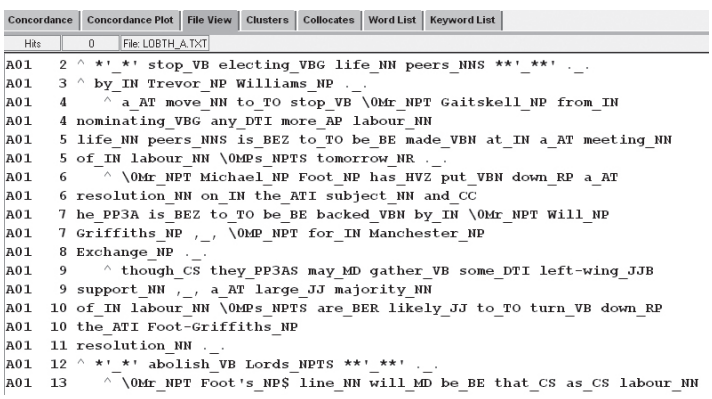


图 1-6　含词性附码的语料库片段

观察指引

图 1-6 是图 1-5 语料片段内容加注了词性附码的版本（来源同上），例如，第 2 行的 stop_VB 为动词，electing_VBG 为动词的现在分词或动名词，life_NN 为名词，peers_NNS 为复数名词（详见 Johansson, 1978: 10-13, 141-149；何安平，2004a: 143-150）。

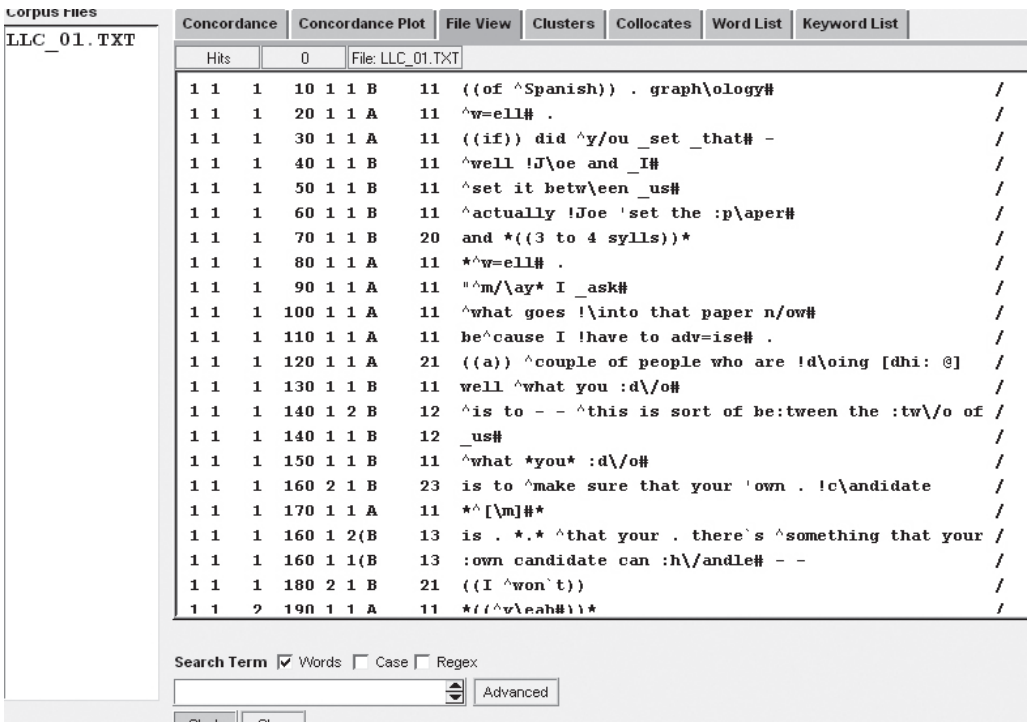


图 1-7　含语音标注的语料库片段

观察指引

图 1-7 是 50 万词的英国口语语料库（LLC，1980）的语音附码片段。以第一行为例，前 6 栏阿拉伯数字分别为文本号及行号等，随后的大写字母 B 为说话者的代号，B 话语中 ((of ^Spanish)) 表示双括号内的话语听得不太清楚，其中的 ^Spanish 表示该部分为语调组（通常一行为一个语调组）的调头重音（表示该词是组内第一个重读词），随后的实心点表示短暂停顿，graph\ology# 表示该调组的降调调核词（表示该词的语调为降调），# 号是该语调组结尾标记（表示该语段结束）（详见 Svartvik & Quirk, 1980: 893; 何安平，2004a: 177-178）。

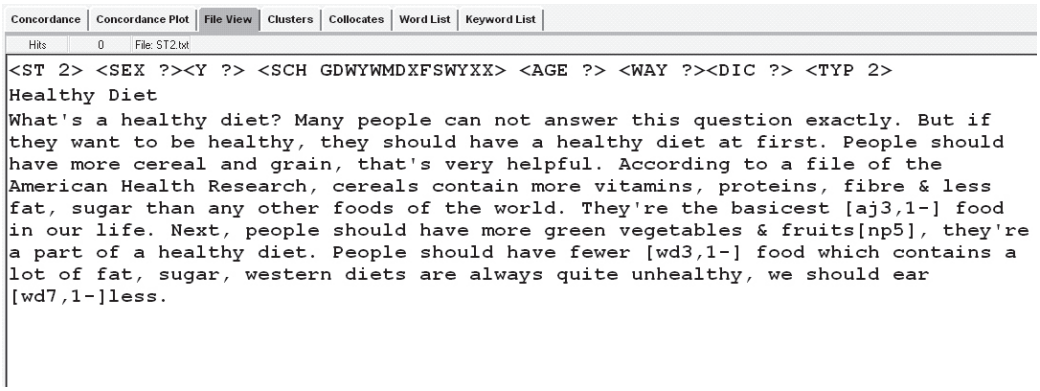


图 1-8　含偏误码的中介语语料库片段

观察指引

图 1-8 显示的是加注了偏误标码的中国学习者英语语料库（CLEC）（桂诗春、杨惠中，2003）片段。顶行的尖括号 <　> 中分别注释了该段书面写作语料的学习者英语水平 <ST>、性别 <SEX>、学英语有几年 <Y>、所在学校 <SCH>、年龄 <AGE>、是考试作文还是自由作文 <WAY>、是否使用字典 <DIC>、文章的类型（议论文还是叙事文）<TYP> 等等。正文中的方括号 [　] 内为偏误类型标码，如 [aj3,1-] 表示形容词类的偏误，[wd3,1-] 则表示用词方面的偏误，1- 表示偏误范围见该标码的前一个词（详细说明见桂诗春、杨惠中，2003: 3-8）。

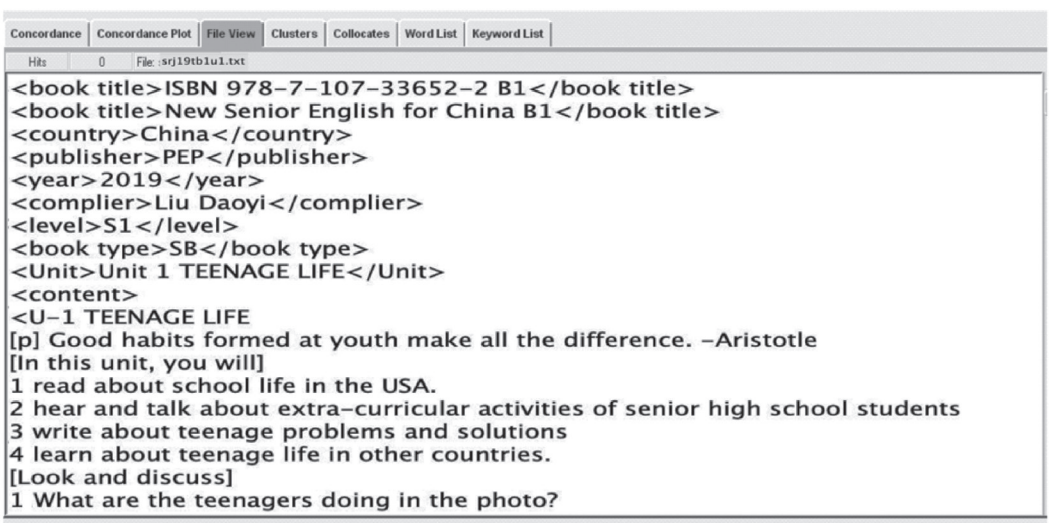


图 1-9　英语教材语料库片段

观察指引

图 1-9 是华南师范大学外国语言文化学院建设的中国英语教育教学语料库（CEEC）中的高中英语教材语料库的一个片段。文件名 srj19tb1u1.txt 表示该语料来自人民教育出版社 2019 版的高中英语学生用书必修一第 1 单元的文字部分；开头尖括号 <　> 之间标注该教材的书号、名称、出版地、出版社、出版年份、主要编者、册数序号、类型（SB 为学生用书，WB 为练习用书）和单元序号及单元名称等信息；语料中的方括号 [　] 内是教材的教学指引性文字和练习指令语。

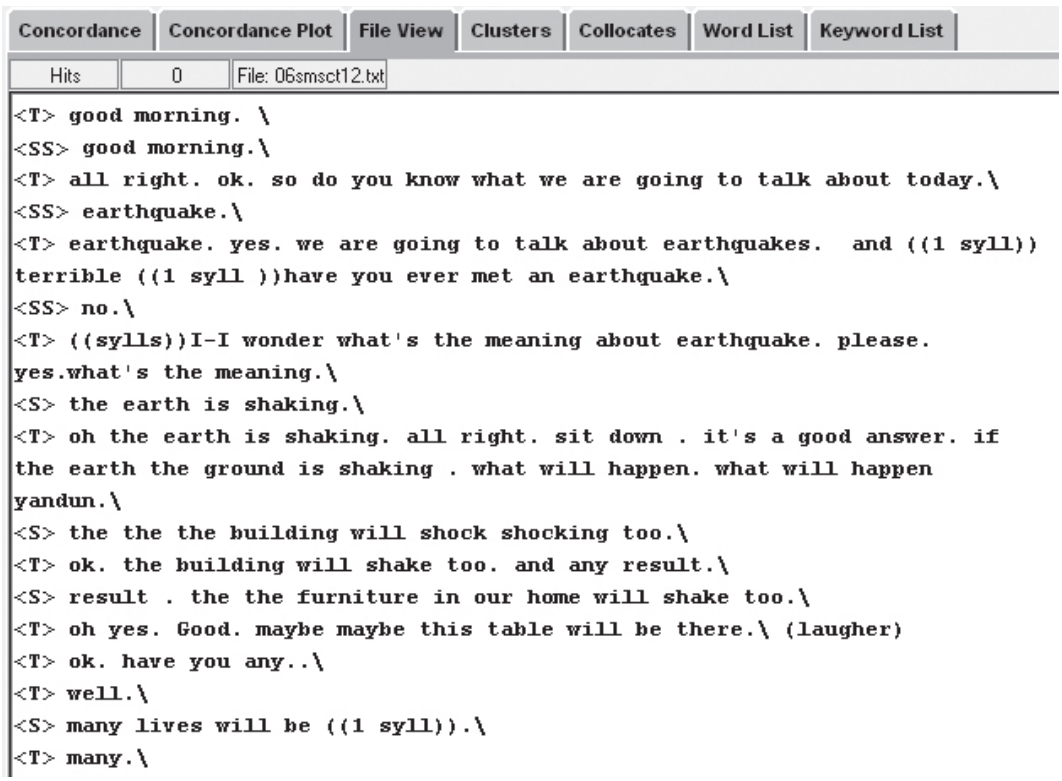


图 1-10　英语课堂教学语料库片段

观察指引

图 1-10 是华南师范大学外国语言文化学院建设的国内中学课堂教学话语语料库的一个片段。文件名 06smsct12.txt 表示是根据华南师范大学外国语言文化学院于 2006 年录入的 12 节高中英语课的视频转写。<T>、<SS> 和 <S> 分别表示“教师”、“全班学生”和“单个学生”的话语；反斜杠 \ 表示说话者的话轮结束；双括号 ((　)) 内表示录音不清楚；单括号 (　) 内表示录音中出现的非言语声音的描述；下脚点 . 表示停顿（详见何安平，2004b: 489）。

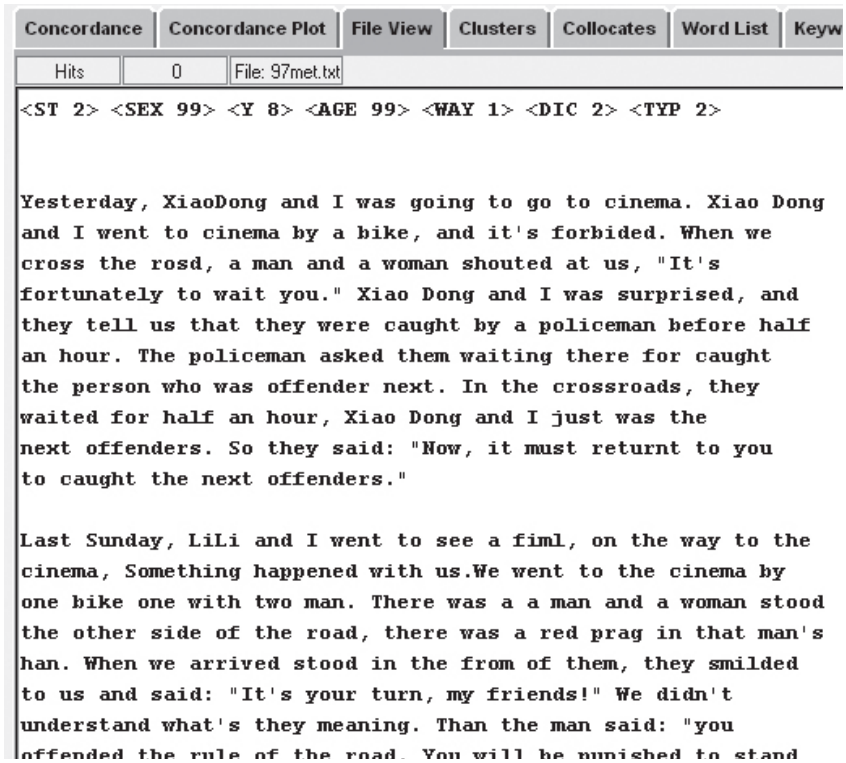


图 1-11　学生英语作文语料库片段

观察指引

图 1-11 显示的是华南师范大学外国语言文化学院收集的广东英语高考作文语料库的一个片段。文件名 97met.txt 表示其为 1997 年广东高考英语语料库。首行 < > 内的内容与图 1-8 的 CLEC 语料标注一致，不同的是 99 表示信息不详。注意作文与作文之间有若干空行作间隔。

Chinese	English
他早就兩眼發黑，耳朵里嗡的一聲，覺得全身仿佛微塵似的迸散了。	All had turned black before his eyes, there was a buzzing in his ears, and he felt as if his whole body were being scattered like so much light dust.
羿吃著炸醬面，自己覺得確也不好吃；偷眼去看嫦娥，她炸醬是看也不看，祇用湯泡了面，吃了半碗，又放下了。	While eating, Yi had to admit that this was not an appetizing meal. He stole a glance at Chang-ngo. Without so much as looking at the crow sauce, she had steeped her noodles in soup, and she set down her bowl half finished.
他覺得她臉上仿佛比往常黃瘦些，生怕她生了病。	Her face struck him as paler and thinner than before—suppose she were to fall ill?

图 1-12　中英平行语料库片段

观察指引

图 1-12 是以中文“觉得”为检索项提取的中英平行语料库片段，检索自香港教育大学在线英中平行语料库（English-Chinese Parallel Concordancer），[1] 展示了两种语言在句子层面上的平行对译。第一个中文例句中的“觉得”在其右边英译句中被翻译为“felt”，而第二句和第三句中的“觉得”并没被直译出来。

【活动 1.3 述评】

以上展示的八种不同类型的语料库仅仅是语料库的冰山一角。各种类型的语料库是为了满足不同的研究和应用的需要，了解每种类型的特征是为了学会使用这些特征来进行有效的检索，同时也为建设符合自身教学研究需要的语料库提供参照或借鉴。

【本章小结】

本章展示了语料库语言学界多位著名学者对语料库与语料库语言学的定义和发展史的陈述，对比了三种不同的文献阅读方式。通过观察八种不同类型的语料库样品，我们对语料库的定义、不同时期的发展特点、发展历程及其概貌有了初步的认识。以下几点可作为拓展学习的议题：

(1) 以下截图中学习者对语料库的定义表述，你认同哪些？是否有补充意见？

我才知道语料库是一个汇集语言各方面数据的综合体
语料库通俗一点讲就是存放语言的仓库
语料库是一本活字典
语料库是一个非常强大的知识宝藏
语料库是一个全新的词
语料库是一个数据库
语料库是一种电子阅读
语料库对于我是一个全新的概念

(2) 思考各种不同类型的语料库（例如笔语语料库和口语语料库、原始语料库和附码语料库等）分别可以为英语教学提供哪些帮助。

(3) 尝试在互联网上搜索至少两种英语语料库并与同伴交流不同语料库在英语教学中的作用。

1　详见 English-Chinese Parallel Concordancer. Project leader: Dr. Wang Lixun. The Education University of Hong Kong. https://corpus.eduhk.hk/paraconc/ or https://corpling.eduhk.hk/paraconc/。