外研社英语语料库研究系列

# 基于语料库的英语语言学语体分析

桂诗春 【著】

# A Corpus-based Analysis of the Register of English Linguistics

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

外 研 社 英 语 语 料 库 研 究 系 列

# 基于语料库的英语语言学语体分析

桂诗春 【著】

# A CORPUS-BASED ANALYSIS OF THE REGISTER OF ENGLISH LINGUISTICS

外语教学与研究出版社
北 京

**图书在版编目(CIP)数据**

基于语料库的英语语言学语体分析／桂诗春著．— 北京：外语教学与研究出版社，2009.12
（外研社英语语料库研究系列）
ISBN 978－7－5600－9175－4

Ⅰ．①基… Ⅱ．①桂… Ⅲ．①英语—语体—研究 Ⅳ．①H315

中国版本图书馆 CIP 数据核字（2009）第 228393 号

**出 版 人**：于春迟
**责任编辑**：段长城
**封面设计**：覃一彪
**版式设计**：付玉梅
**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**：中国农业出版社印刷厂
**开　　本**：650×980　1/16
**印　　张**：8
**版　　次**：2009 年 12 月第 1 版　2009 年 12 月第 1 次印刷
**书　　号**：ISBN 978－7－5600－9175－4
**定　　价**：23.90 元（含 CD-ROM 光盘一张）

*　　*　　*



**物料号：191750001**

# 语库分析，循此门径

# ——序

桂诗春教授这本新书，是在自建一个英语语言学语料库（English Corpus of Linguistics, ECOL）的基础上，指导我们如何用语料库去分析英语语言学的语体，进而学会如何分析其他类型的语体和文体。

在桂老师的自序里，在书中各个章节里，研制这样一个语料库并编写这么一本书的缘起、意义、使用方法等等，都已讲述得很明白。

我本无需置喙。向来自奉“无所发明则不著书，无所了解则不志状”，同时自愧学浅，所以一直谢不作序。然而，此次是桂老师惠允，出版社力邀，我恭敬不如从命，只好写点可能多余的话。

先谈作者，再谈作品。

作者桂诗春教授，是我多年来尊敬的老师，也是我所编刊物《外语教学与研究》长期的作者和支持者。对于他，我的导师许国璋先生更为了解，曾在《语言学系列教材》（湖南教育出版社，1987）“总序”中顺笔给予高度评价。桂老师是老一辈学人，在“文化大革命”后期困难的环境下，开始了他的语言学探索，因此在20世纪80年代初期改革开放后，便能率先开展语言学和应用语言学的教研工作。许老说：

“在中国，他是从独尊文学的环境中冲出来为语言学研究立课程，置图书，培养人才的第一人；又是向外语教学界的经验主义传统（即自己怎样学，于是怎样教）进行冲刺的第一人；又是在20世纪70年代即已开始使用计算机

积累教学资料和实验数据之人，又是首先提出应把外语教学看作一件系统工程之人（即入学水平和毕业水平业经规定，在四年或某段时间里，统驭各种变数去完成任务——培养目标——的一件工程）。……诗春同志是‘不论国内国外，书要靠自己读’的主张者，是怀有理想的教育家。”

许老这段话里说了好几个第一人，在我看来，是概括了桂老师的特点和可钦可敬之处：其一，是心系人才培养的教育家；其二，是勇于向自我和现实挑战的实践家；其三，是重视研究资料和数据的严谨的学问家。

年届80高龄的桂老师，特点仍在，精神仍在，于是才有我们手上的这本书。

语料库随着现代计算机技术发展起来，应用于语言研究，即语料库语言学，这是近几十年新兴的子学科。凭借语料库工具，以实证的手段对各类文本的文体或语体特征进行定量描写和定性分析，便形成语料库文体或语体的研究。语料库带来许多新的研究途径，虽然它们只是传统研究的补充，但方法论上的创新，其意义不可低估。语言组合与思想表述之间有怎样的关联、怎样的特征？大规模的语料库可以提供客观的、多样化的、经过系统标注且便于提取分析的文本，拓展了考察的视野，方便了调研的对比，令研究者兴致盎然，乐此不疲。

对一个文本，可以做语体的分析，也可作文体的分析。两者有相似处，也有不同处。大体说来，语体的分析更侧重词、句、段、篇等语言结构，文体分析更侧重叙述视角、直接/间接引语的使用、修辞手段等。本书属于前者。

书中介绍了语料库语言学基本研究路径和研究方法，以及在语言教学上的用途，更主要的部分是三大块，即基本统计分析、语法特征分析和词汇特征分析。在基本统计分析中，作者分别解释了词频、词汇密度、平均词长、平均句子长度以及词类等常用分析项目的基本概念，并说明如何使用。在语法特征和词汇特征分析部分，则分别讲述了各个值得关注的项目以及如何去分析，尤其是语体的词

汇特征部分作了重点描述，区分定义性、分类性、分析性和修辞性四类语言的词汇特征，以及如何对少用词进行考察分析，都是非常实用的语料库语体分析。书中还附有大量的图表，或助人理解，或教人使用，一目便了然，无言胜有言。

基于语料库的分析都是量化的、统计的。语言研究量化的目的在于对各种语言现象作全面细致的描写，而描写的目的则在于对语言现象作更深层的解释，进而加深对语言和语言规律的认识。相信本书的读者，会举一反三，不仅学会对语言学语体进行语料库分析，也学会自建语料库，包括各种类型各个领域的语料库，并对其进行有意义的语体分析。

谢谢桂老师让我对大作先读为快，先学先用。相信有兴趣学习语料库语言学或应用语料库进行语体研究的读者朋友们同样能感到开卷有益：语库分析，此为门径。

王克非
北京外国语大学中国外语教育研究中心
2009-11-11

# 目 录

## 表格目录

# 插图目录

# 前　言

这本书是为语料库研究者、语篇研究者、语言学及应用语言学专业的老师和学生编写的，其主要目的是“使用语料库（在计算机上贮存大量自然发生的语言数据）和语料库过程（用各种方法处理这些数据的计算机程序）来发现那些帮助我们弄懂语言怎样建立语篇的方式的语言型式。”（Baker，2006:1）在语言学与应用语言学的领域，我们都有阅读英语语言学文献和用英语来撰写语言学论文的不同需要，可是我们对这种语体的认识往往是感性的、零碎的、粗浅的，未能形成理性的判断。从1981年我国开始建立语言学及应用语言学学科点以来，我就参与指导研究生阅读英语语言学著作和撰写硕士、博士论文的工作。但是应该承认，我始终是处在一种昏昏然而不知其所以然的状态。学生的论文有些什么我觉得不甚妥当的地方，就会提出修改的建议。但是为什么要修改，为什么改成我所建议的表达方式，却讲不出太多道理。不能说一点认识都没有，只是全凭感觉，总觉得自己的认识没有很好的梳理。语料库方法为梳理这些感觉与认识提供了很好的手段。

从历史上说，以语料库为基础的语体研究的时间并不很长，其中的一个原因是它和语境的关系很密切，而语料库往往剥离了语境，特别是早期的语料库的样本往往是片断的，不甚完整。但是“从语料库文本来推论语境并非完全不可能的”(McEnery & Wilson，1966)，以语料库为基础的语篇研究近10年内有了很大的发展，2000年在美国召开的第二届北美语料库语言学和语言教学的大会上，就有不少参加者提交用语料库方法研究语言变异的文章，后

来编辑出版了《使用语料库探索语言变异》(Reppen *et al.*, 2002)的专辑。2002 年 9 月，CamConf 2002 在意大利 Camerino 大学召开，专门讨论怎样利用语料库检索技术来研究语体，出版了专集《语料库与语篇》(Partington *et al.*, 2004)。以语料库为基础的语体研究正处在方兴未艾的阶段。以 Biber 为代表的语料库研究者在这个阶段异军独起，依托着他使用语料库资源所编辑的《朗文英语口语和笔语语法》的雄厚基础，对口语和笔语两种语体差别，展开了一系列的研究，倡导了多特征 / 多维度的研究方法，使语料库方法展开新的一页。

用语料库方法研究语体不但在理论上有深远的意义，而且在教学上有重大的应用价值。从理论上说，它追求定量和定性方法的统一，例如 Halliday 就指出，语料库语言学把数据收集和理论化的活动重新组合，导致我们在了解语言时的定性的变化。[1] Tognini-Bonelli(2001:3) 还根据 Saussure 关于言语（Parole）和语言（Langue）的区别来进一步讨论语料库方法对认识语言（系统）的作用。她指出文本和语料库的不同在于：

| 文　本 | 语料库 |
|---|---|
| 完整阅读 | 零碎阅读 |
| 横向阅读 | 纵向阅读 |
| 为内容而阅读 | 为形式上的型式而阅读 |
| 作为唯一的事件而阅读 | 作为重复的事件而阅读 |
| 作为意志的个人活动而阅读 | 作为社会实践的样本而阅读 |
| Parole 的实例 | 对 langue 的洞察 |
| 连贯的交际事件 | 不连贯的交际事件 |

从以上对比可以看到，语料库研究方法的特点在于它通过大量语料从纵向找寻重复出现的语言型式，以提高对语言系统的洞察力，从而为社会实践服务。尽管人们对用

[1] 转引自Tognini-Bonelli (2001) 第一章。

语料库方法来研究语篇不无疑虑，但这种方法的优点还是逐步得到很多研究者的认可，Baker（2006:10-17）将其归结为下面几点：

1．减少研究者的偏颇。Chomsky 提出依赖本族语使用者的直觉来区别语言形式的语法性，但牵涉到语言的使用，问题却并不那么简单。因为语言的使用（也就是语言的变异）和语境紧密相关，而本族语使用者对语言使用的直觉往往是不大可靠的，因为他们把注意力集中在不寻常的事件，而忽略典型的事件。对变异和使用的合适的描述不但要以对自然文本的实证性分析为基础，而且还要以从很多说话人身上收集到的复式文本为基础。这样的分析必须同时考虑各种语境因素的影响。以语料库为基础的分析提供了一种同时处理大量数据和追踪语境因素的手段，开辟了考察语言变异和使用的坦途。

2．语篇的叠加效应（Incremental Effect）。语篇有一种叠加效应，使语料库的方法得以充分发挥。语篇通过使用语言而在社会传播；一个单词、短语或语法结构本身都可以说明一个语篇的存在。但是要证明这个语篇是否典型，则必须有许多支撑的例证。从这些例证的收集中可见语篇的叠加效应，因为这些反复出现的型式说明我们所评估的意义是一个语篇群体共享的。

3．对抗性和变化中的语篇。在展示反复出现的型式的同时，语料库数据还展示相反的例证——反证，这在小规模的研究中是不容易发现的，而且还会误以为是主流。语篇不是静态的，了解语言变化是一种表明语篇在社会中流动性位置的方法。一个在 10 年前被认为是主流的语篇特征在今天看来，也许成为对抗性的、不能接收的。例如 blind 这个词在 20 世纪 60 年代的语料里都是用于字面上的意义，如 blind man，到了 90 年代，则常用于隐喻性（甚至否定性）的意义，如 blind ambition, sheer blind anger, blind panic, blind patriotism, the blind lead the blind, blind to change。

4．三角印证（Triangulation）。也就是使用多种分析方法来互相印证。它可以促进假设的效度检验，使研究成果得到更有力的解释，让研究者更有弹性地应对一些未能预见的问题。语篇研究者不一定非要自己从头开始去建立语料库，只要把语料库作为一个参照点，也能获得数据支持他们的小规模文本分析的结果。

按照Leech(1997:5-6)的归纳，语料库在教学上的应用，可体现在两个方面：其一是核心，它是中心和焦点，在教学中作为教学资源而直接使用语料库；另一个是延伸的边缘，它包括对语料库的一系列间接的应用。还有一些更为边缘的活动则牵涉面向教学而开发的语料库。这包括：

- 语料库在教学中的直接应用
  - 关于语料库的教学，向学生开设语料库语言学课程。
  - 关于怎样开发语料库资源的教学，教会学生亲手利用语料库所提供的信息。
  - 关于怎样在语言和语言学课程的教学中，学会有选择性地使用语料库方法来研究问题。
- 语料库在教学中的间接应用
  - 参考资料的出版
  - 编写教材
  - 语言测试
- 面向教学开发语料库
  - 专门用途语言（Language for Specific Purposes，简称LSP）的语料库
  - 一语和二语发展的语料库
  - 双语/多语语料库

我们所建立的英语语言学语料库（English Corpus of Linguistics，简称ECOL）就是面向教学，特别是语言学和应用语言学的教学的语料库。其根本目的就是通过100万词的语言学样本去归纳出其语言特征（包括语法和词汇特征），供我国英语教师和研究生阅读语言学著作和撰写语言学论文参考。

让我们先看一段关于语言流失（Language Attrition）（Schmid & De Bot, 2004）的文字：

"Language attrition (for the purpose of this article, the discussion *will be confined to* [分析性语言：说明] the attrition of an L1) is often *considered to be* [定义性语言] a reversal of language acquisition. On the *most general* [修饰性语言] level, this definition is *fairly* [修饰性语言] uncontroversial: where language acquisition is a process during which the proficiency in a first or second language increases, in the process of language attrition, lack of contact *leads to* [分析性语言：功能] a *reduced* [修饰性语言] level of proficiency in the *attriting* [修饰性语言] language. (We find definitions which base language acquisition not only on actual [修饰性语言] loss of knowledge that can be shown or assumed to have been there at a previous time, but on "incomplete acquisition" as well (Polinsky, 1994, p. 257) *to be unhelpful* [分析性语言：功能] for the description of language attrition.)

The *task* [分类性语言：种类] of the study of language attrition is to provide a *more detailed analysis and explanation of* [分类性语言：支持中心词] this rather idealized picture, to describe the observed process of loss from linguistic *as well* as [词汇包：语篇组织器] sociolinguistic perspectives, and to try and model the (contact) *variety* [分类性语言：种类] of the *attriting* [修饰性语言] language within given theoretical *frameworks*. [分析性语言：结构] Such an analysis has to take into account *observed*[修饰性语言] *differences* [分析性语言：比较] in the application of rules of grammar and lexical selection between *attrited* and *non-attrited* [修饰性语言] language use (i.e., what are *commonly* [修饰性语

言 ] perceived as "mistakes"), but *ideally* [ 修饰性语言 ] it should also attempt to describe the *linguistic* [ 修饰性语言 ] behavior of *attriters and non-attriters* [ 分析性语言：比较 ]from a more *holistic*[ 修饰性语言 ] perspective. The analysis should *therefore include* [ 分析性语言：结构 ] aspects of the *attriting*[ 修饰性语言 ] language even where it is not "deviant" in an *immediately obvious* [ 修饰性语言 ] way, *e.g.*, [ 分析性语言：说明 ] by establishing factors such as type-token frequency, lexical richness, or grammatical *complexity*. [ 分析性语言：结构 ] *Any study* that focuses *merely* [ 修饰性语言 ] on "what is lost," i.e., on "mistakes" in the speech of an attriter, fails to *take into account* [ 分析性语言：说明 ] avoidance strategies that she might have developed in order to deal with her reduced capabilities. If these strategies are perfected in a simplification of the linguistic system, her speech might very well show up little or no "interferences" at all, and the *emerging* [ 修饰性语言 ] picture might be skewed if "*deviant*" [ 修饰性语言 ]utterances are all that is *considered*. [ 分析性语言：功能 ]

*The picture*[ 分类性语言：支持中心词 ] of the *attrited*[ 修饰性语言 ] language which thus emerges should help us understand *how* different linguistic levels are affected by the attritional process, *how* different sociolinguistic variables affect the attritional process, and *whether* any of the theoretical models available can *account for* these observations. [ 分析性语言：功能 ]"

从这段文字里，我们可见：