《社会语言学与语料库语言学》简评

浙江大学　陆娟鸿　肖忠华

提　要: 《社会语言学与语料库语言学》是英国兰开斯特大学教授 Paul Baker 的一部新作,为第一部系统介绍语料库语言学在社会语言学中应用的专著。本文将首先讨论该专著的创作背景,然后简要介绍本书各章的主要内容,最后在此基础上作出简要评价。

关键词: 社会语言学; 语料库; 研究方法

《社会语言学与语料库语言学》是英国兰开斯特大学语言学及英语系教授 Paul Baker 的一部新作,由爱丁堡大学出版社"爱丁堡社会语言学文库"于 2010 年 2 月出版。Paul Baker 博士是语料库语言学和社会语言学领域的一位重要人物,已出版和发表了大量这方面的著作和论文。本书是第一本系统介绍语料库语言学在社会语言学中应用的专著,全篇力图要回答的一个问题就是如何将语料库语言学方法运用于社会语言学研究。因此,那些"想要了解更多关于语料库技巧的社会语言学家以及想要研究社会语言学问题的语料库语言学家"自然而然成了该书的目标读者。全书共分为七章三部分,第一部分是第一章,第二部分是第二至六章,第三部分是第七章。本文将首先讨论该书的写作背景,然后简要介绍该书各章主要内容,最后在此基础上作出简要评价。

1. 创作背景

"语料库语言学"作为一个术语虽然只是在二十世纪八十年代才出现,但语料库作为一种研究方法却是由来已久,乔姆斯基时代之前的 Boas、Sapir、Newman、Bloomfield 和 Pike 等均使用过这一实证方法(参见 McEnery, Xiao & Tono, 2006: 3) 。随着信息技术、特别是计算机科学的发展,语料库语言学也得到了长足的发展。自从二十世纪六十年代第一个大型现代英语语料库 Brown Corpus 建立以来,语料库语言学作为一种系统的研究方法日渐成熟而被广泛运用于语言研究的众多领域。众多介绍语料库语言学和其他学科结合的著作也随之相继出版。正如 Baker 在书中所述,虽然社会语言学和语料库语言学的结合不是一天两天的事了,但除了个别语言学家(如 McEnery & Wilson, 1996; Hunston, 2002) 对两者结合的可能做法有所涉及之外,尚无专门的著作详细系统地探讨基于语料库的社会语言学研究方法; Beeching(2006) 也只是用了很简短的一章对这一问题做了简单介绍(参见 Baker, 2010: 1) 。作为一位在这方面做了大量研究的社会语言学家和语料库语言学家,Baker 在这个问题上是有一定的发言权的,而出版此书也正好填补了语言研究文献这方面的一个空白,为后来想从事这方面研究的研究者提供了思路和参考。

2. 内容概述

第一章为"引论"。Baker 除了在这一部分介绍其写作背景和动机外,还对社会语言学和语料库语言学分别作了简要介绍。有关社会语言学部分主要介绍了两个重要概念: 变异(variation) 和变化(change) ,及其之间的关系。Baker 强调,尽管大多数情况下语言变化之前都会有语言变异发生,但并不是所有的语言变异最终都会演变成语言变化。该章对语料库语言学的介绍略为详细,除了基本定义以外,还解释了为什么可以用语料库来研究语言、建库时要注意的平衡和代表性问题、语料库的种类以及对语料库语言学的一些误解。此外 Baker 还用了较多篇幅介绍了语料库标注及语料库分析的几个主要技巧与步骤,即频率(frequency) 、检索(concordance) 、搭配(collocation) 、关键词(keywords) 和离散度(dispersion) 。

第二章"语料库与社会语言学变异"主要介绍如何采用语料库方法对不同人群的语言使用方式进行研究。首先介绍了传统变异语言学的研究方法及其不足,并通过两个研究案例向读者展示可用语料库语言学的方法来弥补这一不足。但值得注意的是,通过语料库得到的发现反映的只是一种相对的趋势。Baker 还提醒说,影响研究结果的因素有很多,如语言的使用环境、不同统计测试的采用、语料的离散度; 对此可采用因素或方差分析,或聚类分析来检验究竟是哪些因素对变量有影响,以及在多大程度上有影响。但语料库语

言学不是万能的;它不能解释在语料库中观察到的语言现象,而需借用别的研究方法作出解释。最后 Baker 又通过几个研究个案向读者展示了如何用语料库来研究语域和语音/韵律方面的变异。

第三章"历时变化"介绍语言历时变化的语料库研究方法。Baker 首先提醒读者,语言变化的历时研究和共时研究是可以并存的。而且基于语料库或由语料库驱动的历时语言变化研究其主要对象为书面语而非口语,这主要是因为口语语料(特别是较早时期语料)的缺乏。对此 Baker 介绍了两个办法来克服这一问题。方法之一是检查一下搜集语料时某单一人口的年龄变异情况,但这一做法是建立在以下这一假设之上的,即年纪大的人说话带有过去时代的特点,而年纪轻的人则更具现代特点;方法二是以演讲稿等为说而写的(written-to-be-spoken)文本为研究对象。但是这两个办法也都存在一定的问题。然后 Baker 着重介绍了由四个"小"型(就现在标准而言)平衡英语语料库组成的"Brown family"(即 Brown、LOB、Frown、FLOB)及相关历时语言变化研究。由于这四个语料库的抽样框架基本一致,只是抽样年代不同,因此可用于研究二十世纪英国英语和美国英语的历时变化。Baker 在这一部分提到了一个被他称之为"历时抽样两难推论(diachronic sampling dilemma)"的问题,即建库时各库是采用相同的抽样模型来减少其他因素的干扰,还是不采取统一模型,但对每个时期所建库进行全面描述。因为语言不是静止的,某一特定时期的抽样框架并不一定能反映其他时期语言的使用情况。对此 Baker 也给出了参考办法。例如可在模型中增加某几个类别的文本,在需要时对其进行保留,不需要时则可将其剔除。之后通过一些文献回顾,Baker 指出语言变化和文化变迁之间的关系不是单向的,而是循环并不断相互加强的;语言不仅反映、代表了社会,同时也对社会产生了影响。接下来 Baker 对历史语料库进行了介绍,并强调了历史语料库建库者应注意的事项。最后 Baker 提醒不能轻易下结论,否则结论可能会过于简单,而应该先分析索引行或与之相关的社会、历史背景。

第四章"共时变化"主要介绍如何利用语料库方法来进行共时语言变异研究。Baker 在介绍几个其他国家(除英、美两国以外)的英语语料库和由几个子库组成的大型英语语料库——国际英语语料库 ICE(International Corpus English)的同时,还展示了几个利用这些大型平衡语料库进行共时语言变化研究的例子。不过 Baker 也指出,用小型的专门语料库来研究也有一定的价值,不应被忽视。此外,Baker 还提醒道,在进行共时语言变化研究时,除了对比语言之间的不同之处,其相似之处也值得研究,并通过几个研究案例说明了计算相似程度的两个方法:斯皮尔曼等级相关统计检验(Spearman rank correlation statistic test)和聚类分析(clustering)。不过 Baker 也指出,根据 Kilgarriff (2001)的研究,卡方检验($\chi 2$ test)才是测量语料库之间相似性最好的方法;相似性也只有在先行考虑同质性(homogeneity)的情况下才能对其进行测量。除此以外,Baker 还提到了在研究时可能出现的一些问题,如"虚假发现率(false discovery rate)"的上升和零标记语言特征的识别,并给出了相应的解决办法。最后 Baker 表示,虽然现在这方面的研究语料多为英语书面语,今后对口语和计算机中介传播的语料库研究必将大大增加(尽管这样的语料处理起来还存在一些问题,如标注等)。

第五章为"语料库与人际交流"。Baker 在本章为读者介绍了如何对口语,特别是其中的人际交流部分采用语料库方法进行分析,主要涉及以下几个方面:口语语料库的文本转录、用语料库分析韵律、如何将语料库和传统的人际交流分析方法结合起来、用语料库检索口语中的搭配和结构、话语标记的语料库研究。同时,Baker 特别提到了口语语料库研究中的一个难点,即文本的转录、标注问题。除了介绍常用的 TEI(Text Encoding Initiative)这一文本编码标准外,还介绍了其他几种编码体系。此外,针对不流利现象及其给语法和句法标注带来困难的问题,Baker 介绍了 McKelvie(1998)提出的办法:在对某一话语做句法分析时,将待修正语(reparandum)视为话语的一部分。同时 McKelvie 还制定了一系列规则将停顿、填补停顿、噪音和感叹等不流利现象归入"编辑短语(Edit Phrase)"一类,而将"right"、"yes"、"OK"、"well"等归入"话语标记(Discourse Markers)"一类。当这两类中的现象出现时,不同的句法分析规则就会被触发,从而对待修正语作出分析。

第六章是"揭示话语"。在本章,Baker 不再向读者介绍如何运用语料库的手段来揭示语言的

社会变异和变化，使人们意识到各种具有不同身份的群体之间语言使用的模式和趋势；而是展示了语言是如何被运用于构建、保持或挑战被不同传统的语言研究者称之为态度、思想意识、解释性代码(interpretative repertoires) 或话语(discourse) 的东西。首先，Baker 对在本章所要讨论的“话语(discourse) ”作出了定义，然后在介绍传统的话语分析方法(即批判话语分析) 以后，又通过几个具体的研究案例展示了将语料库语言学的一些技术，如频率、搭配、关键字和检索等，与(批判) 话语分析结合，对话语进行综合研究，利用语料库的实证方法来减少(批判) 话语分析的主观性。Baker 指出，语料库语言学的这些技术能揭示的并不是话语本身，而只是一些语言特征(linguistic features) ；它们只是话语的表现。接着 Baker 指出了这一研究方法的不足之处，例如它并不能解释为什么能或不能发现某些语言模式；基于语料库的分析关注的往往是已经使用过的而非能用但未用的语言等。最后，Baker 给出了三点提醒：第一，某一语言模式并不总是意味着某一意义；第二，分析要深入，得出的结论不能流于表面；第三，不要误以为所有的说话人都是想通过话语来施加力量，对他人进行操控，或误导他人等。

第七章为“结论”。除了对前几章的内容进行总结，Baker 还对未来社会语言学和语料库语言学结合的新发展、新方向做出了展望。

3. 简要评价

本书的特点在于：

首先，正如前文所说，这是第一本全面介绍如何将语料库语言学的研究方法运用于社会语言学研究的著作，填补了该领域的一个空白，为想用语料库方法研究来社会语言学的社会语言学家以及想对社会语言学进行研究的语料库语言学家提供了参考，打开了思路。

其次，正因为这是一本介绍性质的著作，所以大多数章节的内容和语言都比较浅显易懂，理论与实践也紧密结合在一起。实践部分的研究案例较多，而且作者在介绍每个案例时都会注意提醒可能出现以及需要注意的各种问题，为一些常见问题的解决提供了参考模式；有的甚至是不厌其烦地反复提醒，因为这些往往就是非常重要而又很容易被人忽视的问题。稍显不足的是其理论部分的介绍。但这也是可以理解的，毕竟这不是一本专门介绍语料库语言学和社会语言学这两个领域基本知识的书；而对于一些相关的知识，作者也都给出了参考文献、链接等，以供进一步研究学习。

最后，本书的框架结构清晰。每章开头作者都会简要介绍本章主要内容，让读者对本章所要涉及的内容有个整体印象，免得中途有种迷失方向的感觉。而在结束部分也会作出总结，点明其中的重点和要点，以免得读者的注意力被分散而抓不住重点。

本书的不足之处在于，书中的介绍及其案例多针对英语语料库，较为单一，略有语言殖民主义之嫌。而不同的语种各有其特点，在进行具体研究时所碰到的问题肯定也有差别，从这点看来本书的参考价值需要打一定的折扣。不过现在建设得比较多也比较好的就是英语语料库，大量前人研究运用的也都是英语语料库，而作者本人的母语也为英语，从这几点看来，本书中大量关于英语语料库运用的介绍似乎也变得情有可原了。这也再一次提醒了我们需要建立更多更好的其他语种语料库，并用其开展研究的必要性和迫切性。

参考文献

McEnery, A., Xiao, R. & Tono, Y. 2006. *Corpus-Based Language Studies: An Advanced Resource Book*. London: Routledge.

Baker, P. 2010. *Sociolinguistics and Corpus Linguistics*. Edinburgh: Edinburgh University Press.

(通讯地址: 310058 浙江省杭州市浙江大学外国语言与国际文化交流学院)

(文字编校: 陈家刚)

A Review of *Sociolinguistics and Corpus Linguistics*

By LU Juanhong & XIAO Zhonghua

Abstract: *Sociolinguistics and Corpus Linguistics* is a new book by Dr Paul Baker at Lancaster University. This is the first book of its kind that systematically explores the application of corpus linguistics to sociolinguistic studies. The article first discusses the background of the book, and then briefly introduces the main idea of each chapter, which is followed by brief comments on the book.

Key words: Sociolinguistics; corpora; research method