语言资源视角下的语料库建设与应用研究

Corpora as Language Resources: Construction and Application

熊文新 著

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

教育部新世纪优秀人才支持计划(NCET-11-05
教育部人文社科项目"基于语料库及对应词表的英语特异组合研

语言资源视角下的语料库建设与应用研究

Corpora as Language Resources: Construction and Application

熊文新 著

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

图书在版编目 (CIP) 数据

语言资源视角下的语料库建设与应用研究 : 汉、英 / 熊文新著. — 北京 : 外语教学与研究出版社, 2015.4
ISBN 978-7-5135-5928-7

I. ①语… II. ①熊… III. ①语料库－建设－研究－汉、英 IV. ①H0

中国版本图书馆 CIP 数据核字 (2015) 第 085902 号

出 版 人 蔡剑峰
项目负责 孔方卓
责任编辑 李婉婧
封面设计 郭 子
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 北京京华虎彩印刷有限公司
开 本 650×980 1/16
印 张 19.25
版 次 2015 年 5 月第 1 版 2015 年 5 月第 1 次印刷
书 号 ISBN 978-7-5135-5928-7
定 价 57.90 元

购书咨询：（010）88819929 电子邮箱：club@fltrp.com
外研书店：http://www.fltrpstore.com
凡印刷、装订质量问题，请联系我社印制部
联系电话：（010）61207896 电子邮箱：zhijian@fltrp.com
凡侵权、盗版书籍线索，请联系我社法律事务部
举报电话：（010）88817519 电子邮箱：banquan@fltrp.com
法律顾问：立方律师事务所 刘旭东律师
中咨律师事务所 殷 斌律师
物料号：259280001

序

语言研究中的数据之争由来已久。理性主义者主张使用内省和演绎的方法，以理想的本族语者的语感和直觉为起点，步步为营，进行一系列的推导得出结论，进而解释和生成语言，并把赖以形成理论的语感和直觉称之为内省数据（introspective data）。经验主义者主张利用实证数据（empirical data），通过归纳的方法，辅以各种统计手段以获取数据中隐藏的趋势性规律，进而得出结论。语料库语言学是经验主义方法的典型代表，主张从事实出发，以真实语言使用为数据，通过一系列科学的描述进行归纳，或通过统计学手段进行推断，揭示语言使用中的规律性。

语料库语言学研究方法与传统的理性主义研究方法最大的不同点在于，语料库语言学研究中所使用的是海量真实数据，而理性主义方法所使用的则是语言学家的直觉数据。诚然，语言学家的直觉具有十分重要的价值，但若以此为理由，置普通语言使用者的语言产出于不顾，虽有严密的推导和演算，得出的结论也难免会有以偏概全、削足适履之嫌。正因为如此，近年来语料库受到越来越多的语言使用者的青睐，甚至一些生成语法研究者也时不时地借助语料库来激发和验证自己的直觉。笔者主张研究者与数据的互动，因而十分推崇已故美国语言学家Charles Fillmore（1992）的一段话：

“任何语料库，无论它有多大，也不可能包含我希望探索的所有英语词汇和语法信息，因此只靠观察数据是不够的；同样，我所见过的任何语料库，无论它的规模是何等的微不足道，我也曾从其中获取过通过其他任何途径都无法获取到的信息。因此，我的结论是，基于

直觉的语言学家和语料库语言学家彼此需要，谁也离不开谁。或者，更准确地说，我们应该把这两类语言学家的品质融于一身。”

如今，是否应该在语言研究中使用语料库已经成为一个不争的话题，真正的问题是：1）我们从哪儿可以得到这些语料库？ 2）有了语料库我们可以做什么？ 3）我们应该如何使用语料库？

熊文新博士是北京外国语大学语料库语言学团队的技术中坚，他的新作凝练了一位优秀研究者多年的心血和知识积累。该书为回答以上三个问题提供了很好的途径和答案。

有关第一个问题，熊文新博士以自己的研究实践为基础，细致描述了我们可以如何利用当今的互联网，建立满足自身需求的语料库。描述全面、系统，又不乏理论探讨。

有关第二个问题，熊文新博士从丰富的研究案例入手，以崭新的研究视角，展示了语料库在语言研究中大有用武之地。案例既涉及英语和汉语单语，又涉及双语，充分体现了作者丰富的研究经验和广泛的学术兴趣。

有关第三个问题，熊文新博士充分展示了其宽广的知识面。该书将深奥的计算机程序问题藏匿于背后，以友好、易懂的方式把问题的要害呈现给读者。在描述句法分析、语言检索、双语句级对齐等各种技术问题时，熊文新博士娓娓道来，犹如一位循循善诱的老师在手把手地教你一步一步地迈入知识殿堂。

总之，文新兄的大作对理论问题剖析深入，对实践环节的描述细致清楚。我近水楼台先睹为快，并以此为荣！

2015年2月于北京

目 录

第一章
绪论

语言学是关于语言的科学研究（Trask & Stockwell 2007）。对当代语言学的科学研究可以从不同方面展开，如美国结构主义语言学（Structural Linguistics）借鉴物理学等自然科学方法，使用刺激反应论等行为主义研究范式来研究语言事实的使用。研究者开发了一套严谨的语言单位发现程序，为对美国印第安人语言的成功考察做出了重要贡献。生成语言学（Generative Linguistics）学派对刻画具体的语言使用现象不感兴趣，倾向研究人是如何理解和生成自然语言的深层次认知规律，更关注人的内在语言（Internal Language, I-Language），研究语言能力（competence）而非语言应用（performance）。

不同学派有各自研究的侧重点以及理念方法，对语言事实材料的重视程度不一。结构主义语言学以及语料库语言学（Corpus Linguistics）等十分注重采集真实的语言素材，并研发各类观察分析程序，试图从这些客观事实中挖掘提取出语言规则。生成语言学家则更偏重于解释同质（homogeneous）语言社区理想语言使用者的直觉（具体来说，往往就是语言学家本人），反对以异质化（heterogeneous）、不纯粹的语言使用数据作为研究现象（Clark 2006）。撇开不同学派对语言事实材料认识的不同，语言学的研究对象和检验依据都是语言事实，应该是为大多数人所接受的。因为尽管生成语言学家反对采集语料，他们也往往以研究者本人的语感（或称自造数据）来解释判断提出的语言规则。本书主要从语料库语言学的角度，探索大规模真实文本数据的采集、加工，以及语料库平台的建

设与利用，试图为基于语言事实分析的语言学研究提供基础服务。

语料库在语言研究中的重要作用越来越凸显。语言学研究历来具有实证传统。19世纪作为西方现代语言科学萌芽的历史比较语言学（Historical Comparative Linguistics）即以文本集作为研究素材，通过研究不同语言文本之间的对应关系来重构古代语言。这种借助早期语言使用遗迹来探索语言变化与重构的方法，是新语法学派的重要主张（Lüdeling & Kyto 2008）。很多理念和分析技术一直到当代计算机语料库语言学中仍得到发扬和光大，采集和整理历时（diachronic）语料文本仍是当前语料库研究中的一项重要内容。美国结构主义语言学从共时（synchronic）的角度，系统采集语言数据，研制可将口耳相传的印第安语口语转写成文本的记录手段，开发出基于分布（distribution）和替换（alternation）的文本分析方法，应用于语法编写、词典编纂及语言教学等语言应用实践。其研究取向都是选取鲜活语言事实进行概括描写。尽管摒弃语料的理性主义思潮曾经一度令基于语料的经验主义方法陷入低谷，但20世纪60年代以美国Brown语料库为代表的注重搜集语料文本、试图从中发掘语言使用概貌的实证主义思潮，预告了语言学研究经验主义的复兴。随着计算机和网络时代的到来，互联网上涌现出越来越多唾手可得的电子文本，各类统计算法工具不断被研制出来，利用语料库进行的语言学实证研究逐步走向高潮。

第一节 语料库研究现状

“沙发椅语言学（armchair linguistics）”和语料库语言学是语言学中的两种不同研究取向。前者的典型写照是理性主义语言学家身陷沙发苦思冥想，后者是经验主义语言学家盯着无数的语言事实材料，试图从中寻觅出有用的内容。早期这两支队伍是对立的，但美国加州大学Fillmore在1992年就开始寻求两者的融合。他本人是理论语言学家，对语言学几乎所有分支都有重要贡献，像格语法（Case Grammar）、框架语义学（Frame Semantics）和构式语法（Construction Grammar）等。

但在其学术生涯的后期，结合词汇语义学理论与语料库研究思想，他主持开发了著名的框架网FrameNet，既有理论观照，又有事实支撑，成为自然语言处理领域文本语义分析的重要资源。

1 语料库主题的研究发展

早期的语料库研究是一项单调枯燥的工作，但如今它已经从边缘化的活动发展成为主流，吸引了越来越多语言研究者的注意。Svartvik（2007）曾经利用谷歌检索理性主义的“生成语法”“最简方案”等术语和经验主义的“语料库语言学”，发现返回的条目数分别为12,400条、11,600条和34,500条。自从美国语言学界发生乔姆斯基革命之后，理性主义方法一度是语言研究的不二法门。随着上个世纪90年代经验主义的复兴，经过语料库学界多年的努力，语料库研究方法和理念已经牢牢地植根入语言学界。语言学者们自觉不自觉地都会以语料库作为研究对象或手段，进行语言本体及其应用研究。

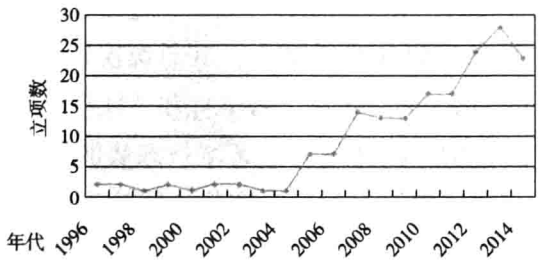

图1　国家社科基金“语料库”课题立项的历年变化

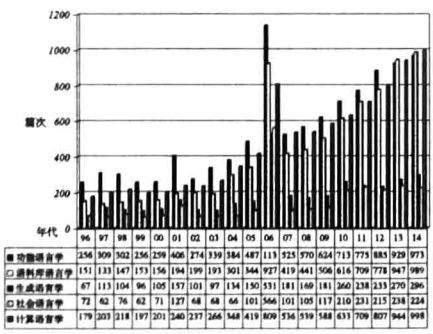

图2　语料库语言学在ScienceDirect库中的检索数据

图1显示了我国国家社科基金1996至2014年包含“语料库”一词的立项项目。国家社科基金是我国人文社科领域最高级别的政府资助项目，具有科研引领风向标的作用。从图中可以看到2006年后有一条陡峭的增长曲线。2006年前，语料库方面的立项课题还只有个位数；此后一路攀升，到2012年以后更是达到20项以上。在这里还没有统计以语言“数据库”等冠名或研究课题利用语料库却没有在项目标题出现这一字眼的立项课题。如果把这部分课题再加上，语料库对语言学研究的重要作用将更为凸显。

从国际范围考察，我们利用知名学术出版商爱思唯尔（Elsevier）推出的ScienceDirect全文数据库，分别选择语言学几个主要分支学科名，如“功能语言学（Functional Linguistics）”、“语料库语言学（Corpus Linguistics）”、“生成语言学（Generative Linguistics）”、“社会语言学（Sociolinguistics）”和“计算语言学（Computational Linguistics）”作为关键词进行查询。图2是这一查询返回的结果。结果显示从1996年开始，“语料库语言学”一词几乎始终呈增长态势。即便本书成稿检索时2014年尚未结束，其总篇次数已经达到989，比1996年增长了将近5倍，而“生成语言学”和“社会语言学”的增势平稳。与“语料库语言学”同样具有显著增长态势的是“计算语言学”。它在一定程度上与语料库语言学具有交集，因为当前基于规则和统计的“计算语言学”同样关注语料库建设和语料库分析工具的开发。

对“中国知网”2000年至2014年间主题为“语料库”的中文论文篇目检索，不难发现2000年论文数仅为93篇，到2010年以后已经超过1500篇。截至2014年9月，2014年当年语料库相关论文数已经超过2008年全年数目，是2000年的10倍以上，如图3所示。图4显示对语料库建设和利用感兴趣的不仅仅是“外国语言文字”和“中国语言文字”两大学科门类，在计算机软件及计算机应用、文学和教育以及新闻与传媒等学科领域都能看到很多语料库的应用实例，如通过采集和分析大规模真实语料文本，计算机应用系统能够从中训练和学习到语言使用规律，建立各类知识模型指导对后续文本的分析。在文

学领域，利用文本的语言风格特征可以推断出作者及年代归属。对报刊及其他各类媒体的新闻报道及评论等文本实施基于语料库的分析，能够及时了解媒体立场观点及社会舆情。

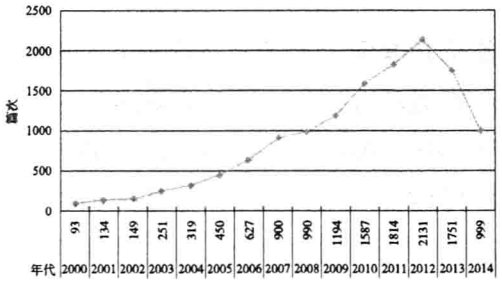

图3　知网有关“语料库”主题的篇目数

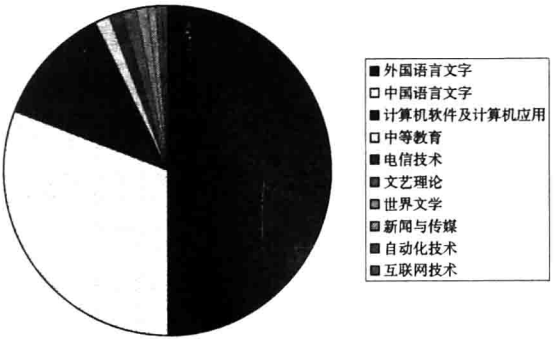

图4　知网有关“语料库”主题的学科分布

语料库语言学及其语料库方法的应用，使得各学科领域的融合态势得以加强，进一步促进了语言学作为领先学科的发展。

2 作为语言学热门研究的语料库

语料库和语料库语言学研究近些年的发展呈现出井喷态势。语料库方法已经成为语言学热门研究。这可以从以下几方面得到验证。

（1）具有确切的研究对象和方法，研究真实世界文本样本（sample）中体现的语言现象。语料库语言学关注在语境中的语言使用，从真实文本中选取有代表性的样本，将实验干扰降至最低，利用