外语学术普及系列

什么是语料库语言学

梁茂成 著

上海外语教育出版社
外教社 SHANGHAI FOREIGN LANGUAGE EDUCATION PRESS
www.sflep.com

外语学术普及系列

什么是语料库语言学

梁茂成 著

上海外语教育出版社
外教社 SHANGHAI FOREIGN LANGUAGE EDUCATION PRESS

图书在版编目(CIP)数据

什么是语料库语言学/梁茂成著. —上海:上海外语教育出版社,2016
(外语学术普及系列)
ISBN 978-7-5446-4433-4

Ⅰ.①什… Ⅱ.①梁… Ⅲ.①语料库—语言学 Ⅳ.①H0

中国版本图书馆 CIP 数据核字(2016)第 155620 号

出版发行: 上海外语教育出版社
(上海外国语大学内) 邮编:200083
电　　话: 021-65425300(总机)
电子邮箱: bookinfo@sflep.com.cn
网　　址: http://www.sflep.com.cn　http://www.sflep.com
责任编辑: 许进兴

印　　刷: 上海信老印刷厂
开　　本: 850×1168　1/32　印张 6.125　字数 165 千字
版　　次: 2016 年 11 月第 1 版　2016 年 11 月第 1 次印刷
印　　数: 2 100 册

书　　号: ISBN 978-7-5446-4433-4 / H · 2024
定　　价: 20.00 元

外教社外语学术普及系列

出　版　说　明

“外语学术普及系列”是上海外语教育出版社专门为外语语言学和文学方向学习者策划出版的一套入门级学术读物，主要分为语言学和文学两大部分，涵盖了这两个研究领域的众多分支，作者多是外语语言学与文学领域的知名专家和教授。

我们希望通过解惑的方式达到传道授业的目的，所以力求简明扼要、浅显易读。本系列每本书均以问答的形式讲解学术领域的专业内容，语言学部分的分册每本包含约 80 个问题；文学部分的分册每本包含约 60 个问题以及 1 篇代表性文学作品的阅读赏析，每册书后均附有中英文对照的术语汇总，以期给读者提供更便捷的阅读参考。

相信本套丛书的出版能满足对语言学、文学研究感兴趣的读者的阅读需求，引领他们进入外语研究的学术园地。

序　言

据记载，公元一世纪时，罗马人发明了玻璃，还发现玻璃有放大效应，后来直到13世纪，意大利人才发明了放大镜——大约能放大6倍，可以看清跳蚤了，于是人们就把这东西叫做flea glasses。1590年前后，一对荷兰父子试着把镜片放在圆筒内壁，结果发现圆筒远端的东西会被放大，于是就试着加长这个圆筒，把三个粗细不等的圆筒套接起来，套接后的圆筒可伸长也可缩短，缩短的时候能放大3倍，伸长的时候能放大9倍。我的问题是，这东西是放大镜还是显微镜？

其实叫它放大镜还是叫它显微镜并不重要，重要的是我们用它可以看到原来看不到的东西。一旦我们操起新式武器，仔细观察一下本来已经习以为常的东西，竟惊讶地发现，我们原本习以为常的东西根本就是错的。

这就是语料库语言学，而语言就是我们天天习以为常的东西。语料库语言学像是新工具，它促使我们反思许多语言问题，尤其是一些本来已经有了定论的问题。于是，我们开始越来越不相信定论，不时地发现那些所谓的定论其实是错误的。

从我的故事里走出来，考虑一下什么是语料库语言学。我们的第一感觉是，语料库语言学就是工具，像放大镜，或是显微镜。的确，语料库语言学具有工具性是任何人都不能否认的，然而，正是这种工具性才使得我们得以看见原本看不见的东西，这不同于我们把筷子当做工具来吃饭夹菜（我们吃馒头的时候还

不用筷子呢!)。用 Sinclair 的话说,Language looks rather different when you look at a lot of it at once。这里的 look at a lot of it at once 就是工具,是方法,是一种工作方式。有了计算机,有了语料库,我们就可以同时看到很多语言,看到语言中千万次复现的规律。Sinclair 或许是希望提醒我们,每时每刻都应该拿着那副放大镜或显微镜,要让它变成我们生活的一部分,变成我们观察语言的唯一方法。当它变成我们观察语言的唯一方法时,它就不仅仅是工具了,而成了一种根深蒂固的习惯,一种不变的态度。

以上我们说的是学界对语料库语言学的两种认识。一些学者认为语料库语言学是工具,是方法;而另外一些学者认为语料库语言学是新理论、新学科、新视角,预示着新发现。语料库语言学中充满了这样的争议。

本书试图解释语料库语言学中的一些基本问题,有的可能是我们曾经思考过的,有的则是我们从未想过的,甚至笔者在写作之前自己也没太想过的。

在写作本书之前和过程中,曾不时地与李文中教授、许家金教授、熊文新教授研讨过若干问题,讨论令我思考,让我感到似乎开窍了,有时讨论激发的思考不知不觉就落在了纸上。在本书写作过程中还承蒙王克非教授、王文斌教授的大力支持,在此一并表示感谢。同时,十分感谢上海外语教育出版社编辑同志,特别是许进兴先生对我一再拖延交稿时间的谅解以及为本书付出的辛勤劳动。

本研究为教育部人文社会科学重点研究基地北京外国语大学中国外语教育研究中心研究成果,得到教育部人文社科重点研究基地重大项目"英语学习者语误自动检测系统的研制"(11JJD740011)和国家社科基金重大研究项目"大规模英汉平行语料库的建设与加工"(10ZD&127)的支持,特此鸣谢。

梁茂成

2016 年 2 月于北京外国语大学小白楼

目 录

（一）什么是语料库

1. 什么是语料库？

语料库，英文作 corpus（复数常用 corpora）。corpus 一词源自拉丁语，本义为 body，即“身体”、“躯干”之义。据 Murray 主编的《牛津英语词典》（*Oxford English Dictionary*）记载，早在 15 世纪，该词在英语中就开始使用，但到 18 世纪才获得了与我们当今所说的语料库相近的意义，指有关某一专题的“文本汇集”。时至今日，在英语中还会有 the corpus of Shakespeare、a corpus of Greek poets 等说法。

在当今的语言学研究领域，特别是语料库语言学研究领域，人们已赋予 corpus 一词以更为具体的意义。在许多人看来，corpus 已经成为专门的、需要给出严格学术定义的术语。如著名语言学家 David Crystal 在 *A Dictionary of Linguistics and Phonetics*（1991）中给出的定义是：

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.

在以上定义中，David Crystal 不仅指出语料库是笔语或者口语的转写，同时还指出了语料库对语言学研究的意义所在，甚至对如何用语料库来从事语言研究也给出了指导性思想。

著名语料库语言学家 John Sinclair 在 *Corpus, Concordance,*

Collocation（1991）中对 corpus 一词的定义则是：

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

在这一定义中，John Sinclair 虽未对如何使用语料库来进行语言研究做出具体说明，但却对语料库中应该包含什么、为什么要使用语料库等问题给出了明确的答案。

众多学者对什么是语料库这一问题的认识既有共识，也有差异。学界的共识是，语料库中的语言可以是书面语，也可以是由口语转写而来的文本。但无论它是口语，抑或是书面语，必须是自然发生的语言（而不是通过问卷或访谈得来的语言，更不是语言学家的语言直觉）。对于如何使用语料库这一问题，学界的意见不尽相同。在 Crystal 看来，语料库是可以用于语言研究的数据，其作用在于可以验证已有的假设，也可以作为语言描写的起点。当然，仅有起点是不够的，我们还需要对语言事实进行概括和理论化，语料库在其中的最大功用是作为原始素材。在 Sinclair 看来，语料库之所以存在，其意义在于可以代表一种语言或语言变体。换言之，只要我们充分观察语料库中的语言使用实例，基于此所得出的结论注定就是对该语言最客观的描写。因此，基于语料库对语言所进行的描写是最客观、最可靠的语言描写，十分不同于一些研究者基于内省和直觉对语言所进行的描写和解释。

McEnery, Xiao & Tono（2006）对语料库必须具备的特性进行了概括，认为语料库必须：1）是机器可读的电子文本；2）是真实发生的语言（口语或书面语）；3）是经过严格的取样而得来的（而不是随意收集而来的）语言样本；4）目的是为了代表一种语言或语言变体。

我们必须看到，语料库语言学家所定义的语料库和普通语言研究者所理解的语料库可能有着重要的差别。在不少普通语言研究者看来，语料库就是语言研究中可以赖以观察并提供线索的语言素材，而在语料库语言学家看来，语料库的真实性和代表性是同等重要的首要问题。失去了真实性或代表性，所观察到的任何语言现象注定是片面和不可靠的，而对这一点，至少从语言研究现状看，普通语言研究者似乎

并不十分重视，在研究中对自建的语料库的代表性避而不谈，甚至毫无意识。我们时常会看到学术期刊上发表一些语言研究成果，其结论所依赖的只是几位作家（甚至某一位作家）一两部作品中的语言素材，或者是源自某一部词典的一些例句，而在语料库语言学家看来，这些作品中的语言素材只是一些破碎的语言事实，而并非系统性的语言数据，因而远远不能代表语言的整体。语料库语言学家特别强调研究中所使用的语言数据必须具有广泛的代表性，能够充分表现该语言或语言变体的各种特征。

我们十分希望读者在阅读本书之后，对什么才是真正的语料库这一问题能有更为准确的认识，同时对如何利用语料库进行语言研究也有大致的了解。

2. 为什么要使用语料库？

长期以来，语言研究主要依靠两种方法，即依赖直觉的内省方法和依赖真实数据的实证方法。这两种方法具有不同的哲学基础，前者是理性主义（rationalism）的，后者是经验主义（empiricism）的。不可否认，理性主义的思辨方法在语言学研究中常常给人一种直击事物本质的感觉，可以让我们少走弯路，因而获得更高的效率；而经验主义依赖客观数据的方法耗时费力，而且不时还会爆料出有些研究者操纵数据、歪曲事实的事情。我们认为，操纵数据、歪曲事实终究是极少数人的个体行为，语言研究需要科学化，决不能因为耗时费力而放弃，况且，语言学家的直觉也未必总是正确的，我们每个人自身的经验注定任何研究者对语言的认识总会或多或少地带有偏见，不顾语言事实的语言研究必然是缺乏科学性、没有解释力的。

Francis & Sinclair（1994：191）指出，语言学家的直觉一经提出，常常给人一种似乎合理的感觉，然而正是这种语言直觉使得我们无法观察到一些重要的语言事实。究其原因，我们认为，受限于过去的技术，我们很难处理大量的语言数据，况且当时也无法采集到大量的语言数据，只好退而求其次去依赖直觉。如今不同了，计算机技术已经变得十

分发达和普及，处理文本数据十分容易。有了丰富的语言素材和发达的计算机技术，语言研究方法可以也应该发生一些变化了。也正因为以上两方面的原因，语料库语言学近年来被越来越多的语言研究者所接受。对此，Sinclair（1991：1）给出这样的描述："三十年前，我们刚刚开始做语料库研究之时，几乎人人都认为我们不可能处理数百万词的语料；二十年前，人们认为此事可能性极小，几乎是狂想；十年前，人们认为这事很有可能，但仍然是狂人之举；而到如今，这种做法已经变得十分普遍。"Sinclair 的以上描述是在二十多年前作出的，期间计算机技术获得了突飞猛进的发展，与此同时，语言数据的量级也呈几何倍数增加。在当今的大数据时代，语言研究没有理由置语言事实于不顾，依赖语料库已经成为语言研究者的共识。

语言研究中使用大型语料库有无可比拟的优势：

1）出于对研究的科学性的考虑。科学研究的目的是从现实中总结规律，最终解决现实问题。语言理论与语言事实之间存在一种张力，不顾语言事实的理论是片面的（徐烈炯 1997；许余龙 2000）。语言学与其他科学一样，应该注重语言事实，完善语言理论，而语料库中包含着我们靠直觉无法想象的事实，等待着我们去探索和挖掘。

2）出于对数据分析效率的考虑。近几十年里，计算机硬件技术和自然语言处理软件技术取得了一系列突破性进展，当年百万词级的所谓大型语料库今天成了微型语料库，处理起来几乎不费吹灰之力，而相同的工作由人工几乎是不可能完成的。高效率的语言学研究，特别是实证性研究，需要语料库分析技术。

3）出于对研究结果可靠性的考虑。在已有的研究中，对相同的研究问题，因研究视角不同、研究方法差异等原因，常常会出现研究结果不一致甚至相互矛盾的现象。相反，几乎所有的语料库研究都是可以重复的。也就是说，基于相同的语料库，采用同样的方法，我们不可能得出不同的结论，因而基于语料库的研究便于重新验证已有研究。

在现时的语言研究中，广泛使用语料库已经成为一种普遍趋势。语言研究者需要更多考虑的已经不再是是否应该使用语料库，而是如何利

用语料库。语料库已经成为语言研究中的默认资源(Teubert 2005)。

总之,出于科学性、可靠性和工作效率等方面的考虑,对于任何可以通过语料库解决的问题,我们没有理由不使用语料库。

3. 语料库与其他语言研究数据有什么不同?

语料库是一种数据源,所记录的是自然发生的语言,其最大特点在于其真实性(authenticity)。此外,语料库规模之大,是任何其他语言研究数据所不及的。

早在20世纪50年代初,Voegelin & Harris (1951: 323) 就对语言研究中各种数据进行过讨论和分类。在他们看来,语言研究中的数据可以分为:

1) 诱发(elicitation)数据。研究者对受试发问,对语言使用的正确与否或可接受度(acceptability)进行评判;

2) 语料采集(corpus collection)。研究者从语言发生情境中采集大量自然发生的语言;

3) 控制性实验(controlled experimentation)。通过控制性实验,直接测量受试的各种语言能力。

在Voegelin & Harris (1951)的分类中,并没有内省数据之说,这或许是因为20世纪50年代初期,在美国甚至国际语言学界,理性主义的研究方法不是主流的研究方法,语言研究更多地依赖田野调查、实验等经验主义的方法。

Widdowson (2000)按照语言产出者的人称,对语言研究中的数据进行了更为简单的分类:

1) 第三人称观察数据(比如,人们在何种场合下才会使用X这个词呢?);

2) 第一人称内省数据(比如,我在何种场合下才会使用X这个词呢?);

3) 第二人称诱发数据(比如,您在何种场合下才会使用X这个词呢?)。

在 Widdowson 的分类中，第一种数据是观察数据，观察对象正是普普通通的语言使用者，因而语料库数据理所当然应该归入此类；第二种数据是主观数据，是典型的纯理性主义的做法，语言学家既无实验又无观察，仅依靠自己的直觉对语言使用作出判断；第三种数据是实验数据，实验者诱发受试产生数据。也就是说，实验者主观地设计任务，由受试尽可能客观地在完成任务的同时产出数据。我们认为，以上三种数据各有千秋：第一种客观可靠，反映的是普通语言使用者的日常言语行为，但数据稀疏是一个无法回避的问题，真理可能埋藏在大量的语言事实之中，需要我们细心观察后才能发现。正因为如此，语料库的规模才显得特别重要，数据挖掘技术也才需要大力发展；第二种数据直击问题本质，研究者所想即研究所需，省去所有的数据采集之苦，但有时可能受限于研究者自身的语言观、生活经历和生活环境，所得到的数据未必能够客观地反映语言事实；第三种数据不像第一种数据那么稀疏，也不像第二种数据那么主观，但实验的可靠性受到多种内部变量、外部变量和中间变量的影响。

我们无意否定内省方法在语言学研究中的价值，毕竟内省是研究得以开展的重要基础。没有了内省，我们可能连研究问题都没有了。但因为内省的价值而一味地排斥客观数据，或者因为客观数据的价值而一味地排斥内省，都会使语言学研究失去科学性。笔者更愿意接受 Widdowson（2000）的看法，把内省也看作为一种数据源。毕竟，研究者自身也可能是语言的产出者，其对语言用法的判断应该成为语料库的重要补充。我们支持徐烈炯（1997）和许余龙（2000）的观点，坚决反对不顾语言事实的内省，主张“让数据说话”，带着内省去观察数据。

总而言之，语料库作为语言研究的数据源，其丰富性和多样性对语言研究的价值是不容置疑的，基于大规模语料库所得到的结论不能轻易地被否定，这正是语料库与其他语言研究数据之间的差异所在。从某种意义上说，语料库是语言理论的试金石。如果我们提出某种理论，或者利用某种研究方法得出某种结论，但这种结论与语料库中的语言事实不符，我们则有理由认为，这种理论尚需进一步完善，需要我们细

心审视是否我们的问卷、访谈或受试取样出现了问题。说到底，否定语料库就是否定语言事实，而任何不顾语言事实的理论注定是不完善、不全面、有缺陷的。我们支持使用交叉验证（triangulation）的方法，由不同的研究者使用不同的数据、采用不同的方法去回答相同的问题，但在交叉验证中，基于语料库的研究结果始终不能轻易地被推翻。毕竟，来自语料库的答案是成千上万人未经诱导而自然给出的答案。

诚然，语言理论不会从语料库中自动浮现出来，需要语言学家依靠直觉和语言经验对其中的数据进行细心的解读，因此我们不能简单地说语料库中的语言事实高于一切。正如 Fillmore（1992：35）所说的那样：

> I have two main observations to make. The first is that I don't think there can be any corpora, however large, that contain information about all of the areas of English lexicon and grammar that I want to explore; all that I have seen are inadequate. The second observation is that every corpus that I've had a chance to examine, however small, has taught me facts that I couldn't imagine finding out about in any other way. My conclusion is that the two kinds of linguists need each other. Or better, that the two kinds of linguists, wherever possible, should exist in the same body.

4. 什么是语料库的代表性？

语料库的代表性（representativeness）指的是一个语料库在多大程度上能够代表一种语言或语言变体中各种不同的语言现象（Biber 1993）。只有在建设语料库的过程中合理地收集各类样本，才能保证基于该语料库的研究所得出的结论能够适用于语料库所代表的语言或语言变体。语料库的代表性是语料库建设者需要关注的首要问题。

具有代表性的语料库，其中的语言具有极大的丰富性，能够充分代表一种语言或一种语言变体。词语等各种语言单位的频率是语言的重要信息，因此，具有代表性的语料库，其中各种语言单位频率的高低应

能够准确反映语言使用的真实情况。换言之，只要该语言中存在的东西，在语料库中都应该可以找到，而且各种语言现象在语料库中出现频率的高低与该语言实际使用中的频率高低情况基本相符，即人们在日常生活中经常使用的语言现象在语料库中的频率也相对更高，而人们在日常生活中相对少用的语言现象在语料库中的频率相对较低；如果一种语言内部同时存在着两种特定现象，但使用频率存在差异，则这种频率差异也应该在语料库中得到准确的体现。比如说，英语正式书面语中会更多地使用 whom，而在非正式书面语和口语中，人们更趋向于使用 who。再如，操各种语言的人都有男女性别之分，如果语料库中过多收集男性语言样本或女性语言样本，则势必会导致语料库带有过多的男性语言色彩或女性语言色彩。如果我们准备建设一个通用英语语料库，需要广泛且按合理比例收集各类正式文本、非正式文本、男性语言样本、女性语言样本，保证语料库能够较好地反映语言中的各种变异。如果语料库缺乏代表性，那么基于语料库所编撰的词典、编写的语法乃至基于语料库所进行的一切语言研究，其结果都是不可靠的。

为了确保语料库的代表性，在建设语料库时至少需要考虑以下几个因素：

1）收集样本前的理论研究。在着手建设语料库之前，首先需要考虑的理论问题是语料库需要代表的整体（population）是什么（Biber 1993）。语料库建设者需要深入分析整体的边界（比如：何为汉语？广东话算不算汉语？港澳地区人们讲的是不是汉语？），以及整体内部有哪些变异类型。常见的变异类型包括地理区域变异（是南方还是北方）、活动场所（是公司还是家庭）、语言使用场合（是会议还是聊天）、传播介质（是印刷文本还是电子邮件或是网络文本等）、说话者性别、文本的互动性（是自言自语还是两人间的书信往来；是情侣间的窃窃私语还是公共场合的大会发言）等等。同时，整体内部的层级关系也是我们需要考虑的重要问题之一。比如，商务活动应该至少包括公司业务会议、商务谈判、商务邮件往来等等，而所有这些变量都归入“商务”这一个上位概念。作为语料库设计者，应该尽可能考虑到语言使用的各