外研社英语语料库研究系列

Using Corpora to Analyze Gender

语料库与性别分析

[英] Paul Baker 著
唐丽萍 译

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

Using Corpora to Analyze Gender

语料库与性别分析

[英] Paul Baker 著
唐丽萍 译

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

京权图字：01-2019-0248

图书在版编目（CIP）数据

语料库与性别分析：汉、英／（英）保罗·贝克（Paul Baker）著；唐丽萍译．——北京：外语教学与研究出版社，2018.11
（外研社英语语料库研究系列）
书名原文：Using Corpora to Analyze Gender
ISBN 978-7-5213-0520-3

Ⅰ．①语… Ⅱ．①保… ②唐… Ⅲ．①语料库－语言学－关系－性别差异－研究－汉、英 Ⅳ．①H0②B844

中国版本图书馆 CIP 数据核字（2018）第 280974 号

出 版 人　徐建忠
项目负责　李晓雨
责任编辑　段长城
责任校对　解碧琰
封面设计　袁　凌
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　北京九州迅驰传媒文化有限公司
开　　本　650×980　1/16
印　　张　16
版　　次　2018 年 12 月第 1 版　2018 年 12 月第 1 次印刷
书　　号　ISBN 978-7-5213-0520-3
定　　价　59.90 元

购书咨询：（010）88819926　电子邮箱：club@fltrp.com
外研书店：https://waiyants.tmall.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
物料号：305200001

河北省教育厅人文社会科学研究重大课题攻关项目成果
（ZD201526）

译者简介

唐丽萍，博士，现任河北师范大学外国语学院院长、教授。研究方向为系统功能语言学、批评话语分析、语料库语言学、英语教育。近年来，在 *Journalism*（SSCI）、《外国语》（CSSCI）等国内外权威学术期刊发表论文多篇。主要论著有《批评性跨文化阅读的主体间评价研究》、《美国大报之中国形象的语料库语言学方法辅助下的批评话语分析》。成果获得河北省社会科学优秀成果奖两项。主持的主要项目包括国家社会科学基金项目一项、教育部人文社会科学项目一项、河北省社会科学基金项目二项、教育厅人文社会科学研究重大课题攻关项目一项。入选“河北省高校百名优秀创新人才支持计划”。获国家留学基金委资助，于 2015 年赴英国兰卡斯特大学语言学系访学。

作者简介

Paul Baker，英国兰卡斯特大学语言学系教授，社会科学语料库研究中心（CASS）核心成员。国际权威学术期刊《语料库》（*Corpora*）组稿编辑，《应用语言学》（*Applied Linguistics*）、《性别与语言》（*Gender and Language*）、《新闻学与话语研究》（*Journalism and Discourse Studies*）等编委会成员。研究方向为语料库语言学、语言与身份研究、批评话语分析。在《国际语料库语言学杂志》（*International Journal of Corpus Linguistics*）、《话语与社会》（*Discourse and Society*）、《批评话语研究》（*Critical Discourse Studies*）等国际权威学术期刊发表论文多篇。出版论著多部，主要包括：《语料库与话语分析》（*Using Corpora in Discourse Analysis*）、《语料库语言学与社会语言学》（*Corpus Linguistics and Sociolinguistics*）、《语料库与性别分析》（*Using Corpora to Analyse Gender*）等。

致　谢

本书所涉及的研究已获经济与社会研究理事会（ESRC）社会科学语料库研究中心的支持（ESRC 的项目号为 ES/K002155/1）。

兰卡斯特大学性别、语言和性行为研究小组以及语料库研究讨论组的成员，对本书中部分分析的前期草稿给予了有益的反馈。

Andrew Hardie 为本书提供了英国国家语料库中男性和女性的口语语料。其输出格式令人兴奋。

John Swales、Uta Romer 和 Matt O'Donnell 耐心地为我解答了有关 MICASE 语料库的问题。

Paul Rayson、Mike Scott、Laurence Anthony 和 Adam Kilgarriff 提供了语料库分析软件。没有他们，本书不可能完成。

我还要感谢 Jane Sunderland 和 Judith Baxter 多年来给予我的支持和宝贵意见。

目　录

致谢 I
第一章　引言 1
第二章　再看性别差异：女性真的比男性更爱说“lovely”吗？ 21
第三章　不是，但是个不错的猜测：女性学者如何表示异议？ 49
第四章　重男倾向和历时变化：所有的女发言人在哪儿？ 77
第五章　话语韵和合法化策略：再看《每日邮报》对男同性恋者的表征 112
第六章　男孩和女孩是什么做的？使用 Sketch Engine 分析搭配模式 143
第七章　三角互证：克雷格列表网的交友广告对性别差异有何揭示？ 170
第八章　结论 214
参考文献 226
索引 240

表格目录

表 1.1　feminist(s) 的样本，取自英国国家语料库中的报纸部分 16
表 2.1　曼哈顿距离的样例 29
表 2.2　英国国家语料库中不同组别讲话者相比较的曼哈顿距离 31
表 2.3　男性和女性言语的总量 33

表 2.4 不同语料库相较产生的曼哈顿距离 35
表 2.5 排名前 10 的男性和女性主题词的分布情况
(源自 Rayson *et al.* 1997) 39
表 2.6 语料库中排前 50 的男性和女性讲话者的主题词使用情况 41
表 2.7 BNC 人口特征部分的男性和女性主题词
(FLOB 作为参照语料库) 43
表 2.8 she 在 BNC 中的每百万词出现率 44
表 2.9 数字（标注为 CRD）在 BNC 中每百万词的出现率 45
表 3.1 检索项和所发现的异议 61
表 3.2 异议的元信息 63
表 3.3 用于异议表达的语言特征 65
表 3.4 不同异议策略的组合 69
表 4.1 COHA 中 -man 和 -woman 功能化词的词频 86
表 4.2 chairman 在 COHA 不同部分出现频数的比较 90
表 4.3 COHA 中具有特殊性别标记的角色 92
表 4.4 COHA 中性别化家庭关系对应词 97
表 4.5 COHA 中男性优先的证据 99
表 5.1 2001—2002 年间《每日邮报》中比较普遍的话语韵 123
表 5.2 在 2001—2002 和 2008—2009 两个时间段之间，
《每日邮报》中差别不大的话语韵 124
表 5.3 在 2008—2009 中更为常见的《每日邮报》话语韵 127
表 5.4 《每日邮报》中同性恋话语韵的小节 140
表 6.1 对 BNC 中 man（作为普通名词）的搭配进行测量的不同方式，
自高值往低值排序 144
表 6.2 man 及其相关词在 BNC 中的搭配词数量 147
表 6.3 man 在布朗家族中的搭配词数量 148
表 6.4 关于搭配词识别的决定 150
表 6.5 搭配词之间的语法关系 155
表 6.6 使 BOY/GIRL 作主语或作宾语的动词（词目）搭配词 156
表 6.7 对女孩的表征 159
表 6.8 对男孩的表征 161
表 6.9 顽劣的男孩 164
表 6.10 赢得荣誉和骄傲的男孩 164

表 7.1 布朗家族英国部分四个时间段中的频数 177
表 7.2 每个语料库的一致性主题语义标注码 182
表 7.3 每个语料库中排序前 20 的自我描述形容词 194
表 7.4 自我描述语特征的频数 195

插图目录

图 1.1 英国国家语料库中文件 KBF 的节选，包括赋码及未赋码文本....10
图 4.1 MAN 和 WOMAN 在 COHA 中每百万词的出现率....84
图 4.2 MAN/WOMAN 和 PERSON 在 COHA 中每百万词的出现率 85
图 4.3 husband 和 wife 在 COHA 中每百万词频率的历时变化....96
图 4.4 男性和女性家庭关系词在 COHA 中每百万词合计频率的历时变化 97
图 4.5 男性优先的历时性变化（每百万词频率）....101
图 4.6 泛指性 man 在 COHA 中的历时出现情况（基于每个年代所抽取的 100 条索引行）....103
图 4.7 mankind 在 COHA 中每百万词出现频率的历时情况....105
图 4.8 与女性性行为放荡有关的贬义词每百万词频率的历时变化 107
图 4.9 性别中立词语在 COHA 中每百万词频率的历时变化....109
图 5.1 搭配词表截图（2001—2002 年语料库）....119
图 5.2 allegations 与 gay 或 homosexual 相搭配的索引....120
图 7.1 比较三个语料库，以识别相似性的程度 180
图 7.2 比较三个语料库以识别差异 181
图 7.3 搭配网络：澳大利亚人的自我描述语 199
图 7.4 搭配网络：印度人的自我描述语 200
图 7.5 搭配网络：新加坡人的自我描述语 201

第一章 引 言

今天你说了多少次“I love you”？

2012 年的一天，距离情人节还有四天，在某国际报社工作的一位记者联系到我。当时他正在写一篇题为《论男人和女人之语言差异》的文章，意在概括出男性和女性在语言上的主要差异，同时摘其中几句话或几段话来说明这样的差异。他希望我给他列举一些性别化差异。他特别提到了 Harrison 和 Shortall（2011）的研究。该项研究通过对 171 位大学生开展调查，发现男性先于女性宣告恋爱了，更早说“I love you”（我爱你）。

我是这样答复那位记者的：要列举出这些差异，是一件挺困难的事儿，因为所需要的语料量是极其庞大的。我们需要获得数百万词次的口语语料，而这些语料需要从来自各种不同背景和地点的大量人群中提取。从不同的时间点抽取语料，以确保我们发现的任何差异都是稳定的，而不是由于某一社会某一阶段的某个特定因素所致，这样的做法应当是不错的。至于那位记者所援引的文章，我的建议是，或许我们不能从那项研究中做出太具有广泛意义的概括。这项研究所使用的参与者人数比较少（99 位女性和 72 位男性），他们年龄相仿且所处环境类似（学生），而且他们被要求记住并报告自己的语言行为（Harrison 和 Shortall 在文章的讨论部分，也对这些问

题有所提及[1]）。为了说明情况，我给这位记者发送了一些英国国家语料库（British National Corpus, BNC）中有关短语“I love you”的信息。BNC是一个大型参照语料库，包括1亿词次的英国英语，其中1,000万词次是录音对话的转写本。在该对话语料中，有约71%的部分，我们可以知道说话人是男性还是女性。尽管BNC只能直观地显示在语料采集的时间点（20世纪90年代早期），英国社会的语言使用情况，但是，作为最大规模的自然发生的口语语料资源之一，它在本书撰写之时仍然是语料库语言学者可以获得的最好资源之一。

我发现在BNC的口语语料中，I love you仅出现了64次。尽管在该语料库中，女性说这句话的次数是男性的三倍，但是在绝大多数情况下，大多数人并没有说“我爱你”（至少在被录音采样时如此）。我半开玩笑地提议，应该鼓励人们多讲这句话。

不出所料，那位记者没再回复。我既没有列举出一些词和短语，去证实关于性别化语言使用的刻板印象，也没有予以驳斥。即便二者选其一，哪一个发现都有可能构成“新闻”。然而，我的回答可以总结为一句话：“没有足够多的证据可以得出像样的结论”。2月14日就快到了，我的回答可能与那位记者想要创作的任何一种故事版本都不合拍。

从性别差异到性别话语

这件趣事说明，在过去20年左右的时间里，性别和语言领域所开展的学术研究与公众/媒体对性别和语言的感知是不一致的。但是，也并非总是存在这种出入。“性别差异范式”（gender difference paradigm）[2]实际上是

[1] 我无意批评Harrison和Shortall的研究，而是想要质疑该记者以一项对处于类似环境下的少量人群的报告行为所开展的研究为例，来支持性别之间存在普遍性差异，这种方式是存在问题的。

[2] 性”（sex）和“性别”（gender）的概念通常被学术界分别用以指代身份的生物特性（例如，X染色体的数量以及/或者一个人是否拥有阴茎或是阴道）和身份的行为/社会特性（例如，人们作为男性或是女性，如何做事/思考/讲话）。有时候，这两个词语是互换使用的，“性别”于是可以被用作表达“性”的一个更加文雅、委婉的词语。尽管存在阴阳人和变性人，性经常被描述为（对于大多数人而言）具有稳定性的男性/女性的二元对立。而性别却被理论化为更加复杂、易变，而且可能包括多重变化梯度（例如，有的人可能在一些方面表现阳刚，而在其他方面又表现阴柔，而且这还可能随着年龄的增长而发生改变）。

一种早期的学术方法，与 Lakoff（1975）所提出的语言使用中的“男性主导”论（male dominance）（男性用语言支配女人的观点）相关。Fishman（1977）提出了女性从事“互动性家务”（interactional shitwork）的概念，其中包括使用问题和模糊语促使男性做出回应，使对话顺利进行，这一概念对 Lakoff 的理论有所拓展，作出了新的贡献。

尽管与性别差异本身相比，Lakoff 和 Fishman 更加关注男性主导的观念，但是这里存在着一个潜在的假设：由于男性主导，女性被主导，那么，不同的性别势必也就会以不同的方式使用语言。到了 20 世纪 80 年代末期，另一种方法经 Tannen（1990）得到推广，该方法强调性别差异，而非男性主导。这个角度受到了互动社会语言学的影响，基本依据是：男性和女性拥有互不相同的“性别方言”（genderlects），这会导致产生“跨文化沟通误解”（cross-cultural miscommunications）。Tannen 认为，男人把对话视为一场竞赛，而女人交谈则是为了相互确认和得到支持。乍一看，该“差异”范式对于性别和语言的思考，是一种政治上更为中立，不至引发争议[1]的方式。为了避开第二波[2]女性主义者关于父权统治的主张，“性别差异”不把男人描述为是压迫者，女人是受害者，也不把任何人的语言运用放在比其他任何人“更优越”的位置。而差异范式认为男性和女性是在分离的言语社区中成长，学习不同的社交方式和语言运用方式。语言的性别差异因此被用以“解释”（异性恋）夫妻内部发生的人际冲突。冲突据说是由误解所致，因为男性和女性除了有不同的需求之外，对相同言语的意义有不同的理解。该范式的一些倡导者建议，不同性别之间需要学着理解彼此的语言。

[1] 尽管性别差异范式在媒体中一直广泛流行，但是在性别和语言研究领域，却一直受到强烈地批评。例如，Troemel-Plotz（1991：490）认为，正是一种“置身事外的、与政治无关的立场”，“使我们亲身经历的不公正和对话支配变得平凡琐碎、不足为奇；它掩盖了是谁应该做出改变；它一次又一次地给差异蒙上面纱，用执迷于拉平差异的狂热，均衡掉我们感受身为男女的任何差异”（同上：501）。

[2] 第一波女性主义与 19 世纪和 20 世纪早期的争取选举权的运动有关，而第二波与 20 世纪 60 年代的平权运动有关（Bucholtz、Liang 和 Sutton 1999）。第三波被确定为始于 20 世纪 90 年代，与后女性主义相连，主张在男人内部和女人内部存在多样化差异，把所有男人都视为运用权力凌驾于所有女人之上的观点是过于简化的。第三波关注对导致不平等的社会结构予以解构（见 Brooks 1997；Tandon 2008）。

因此，“差异”是一个“宏大”的理论，容易把握，又免受诟病，并且还为夫妻之间的冲突提供了一个广泛适用的解释和解决办法。这就不难理解性别差异研究为什么会变得这么受欢迎，在媒体中尤为如此。同时衍生出数不清的关于两性关系的“自助”书籍和报纸文章，谈及风趣的语言性别差异，证实了我们对于男人和女人既有的了解和猜想。

但是，尽管各种性别差异范式在媒体中广为流行，但在学术界，关于男人和女人是否存在语言使用上的显著差异，尚存诸多分歧。有些研究者主张差异确实存在（例如，Locke 2011），而有些学者却表示语言性别差异实为谬谈（例如，Cameron 2008）。在支持差异论的学者中，关于差异来自哪里的观点也是林林总总——或许来自于不同的脑部化学物质、不同生殖系统或是人体肌肉组织和身材体型有关的生物差异，这些因素可能都对人们如何看待自身以及如何被他人看待产生影响。或许，语言性别差异与社会以不同的方式对男性和女性有关。何为得体的语言行为，人们对男孩和女孩的期待也是不同的。20 世纪 90 年代，Judith Butler 采用后结构主义视角，提出性别是述行的（performative）——是一种行为方式，而非一种存在方式。因此，他们之所以使用某种方式讲话，不是因为他们身为男性或女性，而是根据当前社会对不同性别行为举止的规范，利用语言（还有其他方面的行为）去扮演一个男性或女性的身份。Butler 指出，女性名人模仿秀说明性别述行可以被颠覆，因此并不是与一个单一的性别固定不变地联系在一起。人们通过观察和效仿周围的人，获知就其性别而言，什么是正确的性别表现。因此，Butler（1990：31）强调指出，“对于‘原型’的模仿性的重复。……表明原型不过是对自然性和原始性的模仿”。Butler 还把性别述行和性意识联系起来，提出“异性恋矩阵”（heterosexual matrix）（同上：5）。她主张，“……为了让身体保持连贯而具有意义，必须存在一个通过稳定的社会性别（阳刚表示男性，阴柔表示女性）来表达稳定的生理性别，而社会性别通过强制性的异性恋实践被对立性地、等级性地加以界定”（同上：151）。

自从 20 世纪 90 年代以来，在性别和语言研究领域，学者努力改变把所有男性和女性都硬塞进各自不同的范畴做比较的做法，并转向对女人或男人内部差异的研究，例如，关注性别与其他身份范畴相互作用的方式

(Eckert & McConnnell-Ginet 1992)。这样的方法形成了另外一组别样的研究问题，这些问题围绕语言使用有助于创造、反思和挑战社会规范的方式展开，正如 Butler 指出，这些社会规范影响着男性和女性讲话的方式。诸如社会规范（social convention）和期待（expectations）这样一些术语是与话语（discourse）的概念相关联的。Foucault（1972：49）将话语界定为“系统性地构成他们所谈及物体的实践”，而 Burr（1995：48）提议，话语是“以某种方式，对某个事件版本的诸如意义、隐喻、表征、形象、故事和陈述”的协同生产。Gill（1993：166）强调，语言在社会科学的各个领域已然变得越来越重要，这是由于受到了“强调社会生活具有彻底的话语性和语篇性的后结构主义思想的影响”。Cameron（1998：947）指出，这个“语言学的”转向实际上主要是向话语分析的转向。Livia 和 Hall（1997：12）主张，“是话语生产讲话者，而不是讲话者生产话语。这是因为行为只有‘出现在具有约束性规范的语境中’，才是有意义的”。

于是在性别和语言研究领域中，很多重要的方法都利用了话语转向，其中的话语心理学（discursive psychology）研究就结合了来自会话分析（conversation analysis）、民俗方法学（ethnomethodology）和修辞社会心理学（rhetorical social psychology）的不同元素。一些研究者把后结构主义理论或是批评话语分析（critical discourse analysis）的元素引入话语心理学，例如，Edley 和 Wetherell（1999）对男性青年如何谈论父亲身份的研究。还有学者展示如何把会话分析的方法用于女性主义研究，例如，Kitzinger（2008：136）展示了“性别——或性行为、或权力、或压迫——如何在互动交流中得以生产和再生产”。还有一个不同的方法是把批评话语分析和女性主义语言学（feminist linguistics）相结合，形成女性主义批评话语分析（Feminist Critical Discourse Analysis, FCDA）。FCDA 批判“维持男权社会秩序的话语：即各种权力关系系统性地赋予男人作为社会群体的特权，而女人作为社会群体，则被置于劣势、被排除在外、被剥夺权力的位置”（Lazar 2005：5）。FCDA 于是聚焦于概述语言是怎样维持不平等的性别关系的，目的在于解放和改革这一不平等关系。尽管 FCDA 除了研究那些有关性别的深以为然的假设如何被（再）生产之外，也研究其如何被协商、被争辩，不过，Baxter 提出的第三种方法却更加坚定地聚焦在协商

上。Baxter的女性主义的后结构主义话语分析（Feminist Post-Structuralist Discourse Analysis, FPDA）“提出，女性总是采取多个主体立场，对女性泛泛而谈或是把任何个体的女性仅仅看作是受男性压迫的牺牲品，都太过简化了”（Baxter 2003：10）。而FPDA对语篇（通常是会话的详细转写本）开展细致的定性分析，展现参与者（尤其是那些可能被视作相对弱势的人）如何有可能体会“权力时刻”（moments of power），而有权势的人如何有可能被置于暂时无权的位置。

在性别和语言研究领域，性别话语（gendered discourses）这个概念非常有用。Sunderland（2004）提出，性别话语可以通过分析语言使用中的踪迹加以识别：

> 人们不能……以任何一种直接的方式……识别出话语……它不仅不能被识别或是被命名，而且作为某特定文本中的一段话，也不是那么不言自明或显而易见，它永远都无法完整地存在于“那里”。那里存在的是某些语言的特征：“纸页上的标记”、说出来的词语、亦或是人们对从前对话的记忆……这些——如果是充分并连贯的——可能表明，它们就是某个话语的“踪迹”。（同上：28）

Sunderland承认，对于一个性别话语的识别和命名是一个高度主观的过程。她的方法包括从功能（例如，保守的、抵抗的、颠覆的或是破坏的）和关系（例如，两个话语可能是相互竞争的或是相互支持的，或者一个可能是主导性的，而另外一个是从属性的）的角度对话语进行归类。话语之间的相互关系有助于解释为什么人们在立场上似乎是不一致的，这是因为他们可能在使用相互冲突的话语。

上述各种女性主义的话语分析方法还都强调文本间性（intertextuality）（文本之间的关系）、话语间性（interdiscursivity）（话语之间的关系）和自反性（self-reflexivity），提倡研究者应该承认自身的理论立场，并对研究实践进行反思，“以免这些在不知不觉间使女性在性别等级秩序中的区别性待遇得以固化而不是颠覆（Lazar 2005：15）。”

在很多基于话语的各类语言和性别研究中，还有一点共同之处：常常

是对少量短小文本开展“细致的”或是定性的分析（此外，还把那些与文本创作、传播和接受有关的实践也考虑在内）。这么做是有充分理由的。理由之一是，对于话语的识别和批判是一个复杂且耗时的过程，要求关注细节以及考虑诸多类型的语境（见 Flowerdew，即将出版）。正如 Mills（1998：247-248）所指出的，尽管女性主义的成功已经对性别歧视语言中的一些显而易见的有害形式起到了遏制作用，但是应该说，性别歧视话语并未被彻底根除，而是已然变得越来越错综复杂、老练世故、含混不清，因此更加难以辨别。Mills（同上）主张，“现在亟需一种能够分析性别歧视复杂性的女性主义分析形式，……因为女性主义已经使性别歧视变得更加问题重重。”

因此，尽管话语分析在性别和语言研究领域已经非常流行，但是，这往往都是基于少量文本片段的详细定性研究，而不是使用擅长处理大量语料（数百万或是数十亿词次）的语料库语言学技术的方法（详见下述）。为了说明语料库语言学对性别和语言研究领域的影响程度，我在《性别和语言》（*Gender and Language*）期刊 2007 年至 2012 年间第 1 至 6 期发表的 63 篇文章中，对 corpus 及其复数形式 corpora 的词频进行了检索。其中有 25 篇文章至少一次提到了 corpus 或 corpora，但这未必表明这些文章使用了语料库语言学的研究方法。的确如此，这些文章的作者主要是使用该词来表示他们的数据集是语料库，而在分析时使用的却是纯粹定性的方法。我认为只有四篇文章（占总数的 6.3%）可以被界定为使用了语料库语言学研究方法（Baker 2010；Charteris-Black；Johnson & Ensslin 2007；King 2011；Seale 2009）。此外，Holmgreen（2009）使用了一个语料库去验证她的一些发现，但其主要方法还是定性的。有证据说明，在性别和语言研究领域，有研究者在使用语料库研究方法，尽管他们似乎是少数。

写作本书的主要初衷是提出并证明语料库语言学的一些方法对于性别和语言领域的研究者来说，是有价值的。我并非鼓励研究者们放弃已有的方法，而是把语料库语言学作为一个方法上的补充。因此，本书主要有两类读者：第一类读者对性别和语言研究感兴趣，并且愿意更多地了解语料库语言学如何能够帮助他们开展研究；第二类读者正在开展语料库语言学研究，对性别和语言领域不熟悉但又想在语料库研究中纳入对性别的调查。

我权且认为，读者具有基础水平的计算机能力（譬如，懂得如何在电脑上创建、更改以及找到文件和文件夹，能够使用 IE 等浏览器，在互联网上发送邮件和获取信息），但无需成为电脑程序设计师或是统计师。的确，本书的一个目的就是要证明，在无需成为计算机或数学行家的前提下，在语料库语言学的范式里，能够取得哪些成果。话虽如此，这些行家一直都是，并且将继续作为支撑该领域发展的中流砥柱。我希望本书能够帮助“非技术专家型”的性别和语言研究者，使他们在建设和使用语料库时感到自信，同时也鼓励语料库语言学者能够把性别和语言研究中的一些最新思考融入他们自己的研究。因此，本书的每一个分析章节（第二章至第七章）都综合了对各种不同语料库的分析，带着不同的目标，使用不同的技术，面对和解决随之出现的各种话题和问题。我尽可能做到内容全面。内容介绍如本章最后一节所示。

不过，为了更好地了解这种方法，接下来，我要首先讨论与语料库语言学有关的一些主要术语和概念，目的是要更好地解释为什么在性别和语言研究中，这是一种值得考虑的方法。

建语料库

因为语料库语言学主要就是基于它的研究方法，也就是开展分析的各种方式，故其用途足够广泛，可以应用于很多话题的研究。不过，就像我在别的地方已经谈到的（Baker 2005：7-14），性别和语言领域的研究者对于开发语料库语言学的潜力，相对缓慢。原因可能是各种各样的：不熟悉，没有掌握数据和分析工具，错误地认为这是一个纯粹定量的方法，亦或是不喜欢计算机。在本节，我首先要谈谈语料库语言学背后的依据，然后就能够用以分析的一些主要方法展开讨论。

“corpus”（语料库）这个词在拉丁语中是“body”（身体）的意思，所以语料库语言学指的就是语言的身体。这个身体通常包括文本的集合，要么是完整的文本，要么由其中的较小片段组成。关键的一点是，这些文本都是在“真实世界”中发生的语言使用的真实情况，而不是语言学家为了证明某个观点而编造的句子。像是“The cat sat on the mat”这样一些虚构的句子，并不总是准确地反映人们实际使用语言的方式。语料库语言学因