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

此是一个植根于实证主义的方法，与社会科学中其他一些提取样本以推广至更广泛总体的方法有诸多共同之处。抽样和代表性这些相互关联的概念支撑着所有语料库语言学的研究——语料库需要以这种方式来抽样，才能令人确信它确实代表了我们正在研究的语言变体。例如，如果我们想要在英格兰地区的青少年女性中抽取言语样本，我们需要确保从各个地区（例如，北部、南部、中部）、社会和种族群体以及环境（家里、学校、购物、和朋友外出、访亲、看医生等）的女孩中收集言语的转写本。我们还需要保证抽样是相对平衡的，没有任何一类女孩或是言语事件较之其他情况，被过度代表。理想情况下，每一个女孩都应该贡献相同数量的言语，这样就不会有任何一个个体的言语特质使我们的数据集发生偏斜（尽管有时候，现实情况意味着这并不总是能够做得到，还需要采取其他措施，把特征的分布情况考虑在内，以避免妄自做出错误的结论）。

语料库通常由能够存储于计算机的文本文件组成。对于那些原本以书面形式出现的文本（例如，报纸文章、书籍、邮件等），这种存储是比较容易实现的，特别是如果这些文本已经为电子版格式。那些只有纸面形式的文本（例如，手写的信件）需要转换为电子版格式，可以通过键盘输入，也可以使用光学字符识别软件（如果文本是经打字录入的）。对于像是对话录音这样的口头文本，则需要把讲话转写成文字，而且语料库建设者还必须考虑如何、或者是否要使用一套编码方案来表示笑声、轻声说话和外部噪音等特征。一些语料库是多模态的，包括图像（运动的或静止的）和语言（书面的或口头的）的组合。同样，这也需要使用编码方案，以使得所有相关信息均被纳入语料库。

在语料库语言学中，即便是对于像是报纸文章这样的书面文本，为了使分析任务易于开展，添加标注（或标码）也是常见的做法。很多语料库文本含有标头，出现在文本的开头，或是链接到一个单独的文件。标头包括给出文本元信息的标注码，例如，作者是谁，写于何时何地，属于何种语类，是否已对拼写不规则的情况予以纠正等。对话的转写本可能包括每一位讲话者的年龄、性别或社会阶层等信息。倘若我们想要进一步关注某种类型的说话者，这些信息就会派上用场。标注码除了出现在标头之外，还可能出现在正文部分。例如，我们可能想要以某种方式标注标点符号，

以避免产生误解，特别是假如该文本要在各种不同的操作系统中使用。例如，“‘”这个符号可能表示引号的开始、引号的结束，亦或是一个省字号，那么，把语料库里每一个“‘”所具有的实际功能，通过标注码纳入相关信息，倘若我们想在一个语料库里提取所有引语时，就会大有助益。为了保留一个文本原始的排版格式，诸如项目符号、分段标记、斜体字这样的信息有时也需要做标注，这是因为当文本被转换成可用于语料库软件分析的纯文本时，很多诸如此类的信息都会从文本中被剥除。

语料库在消除词的某些用法或意义歧义时，也是大有助益的。例如，love 除了用以指一种情感或是网球比赛的分数（在网球比赛中，0 分被称为“love”），还可以用作动词、普通名词或是专有名词。通过对单词进行赋码，可以反映出所属的语法或语义集合。图 1.1 为取自 BNC 某文件的一小部分语料，标注码 <u> 用以标记每一句话、<w> 用以标记每个词的语法类别、<s> 用以标记句子以及 <vocal> 用以标记发出的声音，譬如笑声。标注码还包括属性和值的信息。例如，标注码 <u> 具有 who 的属性（指讲话者的身份）。附值如下例所示，像是 PS04U 这样的代码指个体讲话者。

```
<u who+PS04U>
<sn+“0012(03900)”><w PPIS1>I<w VV0>mean<w PPY>you<w VV0>know<w
RGQ>how<wDA1>much<wNN2>boxes<wVBR>are<c
YSTP>.</s>
</u>
<u who+PS04Y>
<sn+“0013(03901)”><w RR>Quite<c YSTP>. <vocal desc+laugh></s>
</u>
I mean you know how much boxes are.
Quite.
```

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

图 1.1 还展示了未经标注的同一个文本。由于语料经过赋码之后，读起来可能会比较费劲，尤其是对于那些不习惯带着标注码开展研究的人来说，所以大多数语料库分析软件都允许将标注码隐藏或忽略。例如，

除非是特别予以说明，WordSmith 工具会忽略掉符号 <and> 之间出现的任何内容。这些符号通常被用以在最常用的标注方案之一的 XML 中表示标注码。

有些语料库对于某个语类具有完全代表性，而另外一些则无法做到。例如，如果我们建设的语料库涵盖了某位作家于 20 世纪创作并出版的所有小说，那么我们对该作家在那个世纪所出版的小说作品就具有完全代表性。然而，倘若是青少年女性言语语料库，那么永远都无法把每一句话都囊括在内，这只能是一个有代表性的样本。Hunston（2002：14-15）在参照（reference）语料库和专用（specialized）语料库（前者通常也被称为通用语料库）之间做了重要的区分。参照语料库旨在作为某个特定语言变体的代表，诸如美国英语。参照语料库包括范围广泛的文本类型，通常规模庞大，包括数百万或数十亿词次。而专用语料库通常都不会那么规模宏大，仅包括一组文本用以（部分或者全面地）代表某个具体的语类或者文本类型。例如，一个专用语料库可能会汇集美国报纸自 2000 年至 2010 年间围绕某个话题发表的文章，或是汇集 2013 年 2 月在英国境内的推文，或是汇集某位作者的小说作品。不过，参照和专用之间的区别与其说是二元对立、非此即彼的，不如说是线性相续、亦此及彼的。可以这么说，所有语料库均旨在作为一些事物的参照，而所有语料库在一定程度上都是专用的。因此，经常用作参照语料库范例的 BNC，也还是被专门限定在了取自一个并不连续的、相对较短时期的英国英语，而不能代表所有英语，甚至不能代表所有的英国英语。

还需要根据所收集的文本类型和语料库建设者的目标来处理道德和版权的有关问题。如果语料库仅供个人使用，而且这些文本已经可以广泛获得，不受版权限制，那么研究者或可无需征得任何人的许可就能使用。不过，如果打算把研究发现公开发表在著作或是期刊文章当中，而且还想要从语料库中复制大段文本，那么通常则需征得版权持有者的许可。

对于为了商业用途而建设的语料库来说，征得使用文本的许可权就更是尤为重要了。此外，收集包括个人信息的文本，例如，邮件或是包括个人详细资料的转写本，也要求研究者从文本贡献者那里征得使用许可权。即便是讲话者或者邮件作者欣然同意他们的个人信息可以在语料库中出现，

对相关文本进行修订或修改以便使其身份难以追踪，仍然不失为明智之举。匿名性问题有必要谨慎对待，正如 McEnery 和 Hardie（2012：57-70）在讨论关于语料库建设和传播的诸多道德问题时所指出的那样，即便是对姓名做了匿名处理，可能也仍然不足以保护讲话者的身份。

我们为什么需要那么多文本呢？为什么就不能还像过去那样，仅对几个精挑细选的文本进行分析呢？在很多情况下，这些分析有效而且有趣。然而，如果我们想对鲜活的语言做出更具广泛意义的概括，那么我们所拥有的文本则越多越好。语料库能够帮助我们免受如下指责：为了证明一个先入为主的观点，我们精挑细选出了最适合的例子。例如如果我们想要证明某个广播节目具有性别歧视，我们可能会选择分析其中一段五分钟的对话，而我们事先已经知道这段对话是含有性别偏见的言辞的。然而，如果我们能够说明性别歧视充斥于很多片段，日复一日，而且并未因图一时方便而忽略掉实际上还有很多广播节目在挑战性别歧视，那么我们的论证就会更具说服力。语料库分析于是能够告诉我们哪些属于典型情况。

即便只是在研究一个文本，语料库能帮助我们验证对其中某些词或短语在用法上的直觉是否可靠。我们不妨看一篇短文，其中作者把一个女人描述为 clucking。我们可能会说这是一个负面的表征，因为把这个女人说的话比作是母鸡的咯咯声，从而说明是招人厌烦的，并且 / 或者是无关紧要的。然而，为了说明作者使用的就是大家比较熟悉的性别歧视话语，从一个大型参照语料库中找到更多的例子，不失为一个有用的做法。那么，在这样的语料库中，clucking（或是与之相关的动词，比如 quacking 和 twittering）经常在提及女性时出现吗（但却不是那么经常地和男性一起出现）？我们不妨再看一个例子，假如我们想要分析一篇报纸上的小短文，其中包括“gay activists are always calling for tolerance and understanding”这句话。我们可能想说，作者对同性恋活动家暗存偏见，尽管这句话的语境还是有些模棱两可。然而，如果我们到类似 BNC 这样的参照语料库中，检索与之类似的句子结构：[human] are always [动词的 -ing 形式]。我们发现其中有很多都包括下述情形：人们因为对某事发牢骚而受到批评，这后面经常会再跟一个追加的句型，其中某人被表征为正在说话、宣称、或是做事，作者于是批驳这是不正确的或是不明智的。因此，我们就有了证据来说明，

写同性恋活动家的这位作者使用了一个隐含表达负面立场的结构，那些比较通晓英语的读者就会受到心理触发，无需挑明就能心领神会。语料库因此能够揭示与词和短语相关联的隐藏意义，从而为我们提供一种更加有力的指出偏见的方式。

接下来这一节描述的是语料库语言学当中的一些较为常用的分析技术。

然后分析它

一旦语料收集入库，通过阅读所有语料来开展语料库分析，固然是可能的（我在第三章，某种程度上用的就是这个方法），但是，大多数语料库语言学研究都要使用能够对语言数据快速开展复杂计算的专门软件。这些工具的优势在于快速、准确，仅在几秒钟之内就能完成人类要花数小时或数日才能完成的计算，更何况人类的计算还可能会出错。

一些语料库建设者将其语料库开放的方式是绑定一个与之配套的软件包。这种模式在互联网上越来越常见。使用者通过网站（例如，第三、四、六章）和语料库实现交互。而在其他情况下，倘若我们只能获及构成语料库的文本文件，我们就需要自己找软件来开展分析。

尽管互联网上可供下载的软件包有很多，我在撰写本书时还是使用了两个非常流行的工具，即 Mike Scott 设计的 WordSmith 和 Laurence Anthony 设计的 Antconc。这两个软件都会定期更新（本书使用的是 WordSmith 第 5 版和 Antconc 第 3 版）。WordSmith 的功能更加广泛，在处理库容达数百万词次的大型语料库时，速度往往更快，不过使用者必须付费才能获得所有功能的使用权限，而 Antconc 是免费的，但是功能比较有限。另外，Antconc 的界面更加简洁，所以我可能会推荐给初学者使用，待更加熟练之后，不妨再考虑 WordSmith（在第五章，我使用的是 WordSmith，而在第七章，使用的是 Antconc）。

语料库工具能够对语料库数据执行计算操作和统计检验，不过分析者受限于软件能力，在有些情况下，还需要再使用其他工具。在使用语料库开展研究时，我经常发现有必要再去求助更加专业的统计分析软件。其中一种选择是使用开放源代码程序语言和软件环境 R（见 Gries 2009）。R

要求使用者学习一种程序语言，而其他统计软件，譬如 SPSS，只需将数据输入表格，然后选择菜单选项来开展检验。流行的微软电子制表程序 Microsoft Excel 也可以开展大量的统计运算，比如平均数（mean）、标准方差（standard deviation）、卡方检验（chi-square test）和 Pearson 相关系数（Pearson correlation）。

此外，还有快捷的在线计算器，把数字输入方框，就能完成某些检验。例如，我有时会使用由 Paul Rayson6 开发的在线对数似然比（log-likelihood, LL）计算器。这可以用来在两个语料库中比较同一个词（或是在同一个语料库中比较两个词）的出现频率。我若想要比较两个以上的语料库，就会使用一个用于卡方检验的网络应用程序（Preacher 2001）。对于更加基础性的数据计算（例如，对一栏中的数字求和或者是对一栏中的每一个值与另外一个数字求积），我往往靠 Excel 来完成。我在第二章计算曼哈顿距离（Manhattan Distance）时，用的就是 Excel。我和很多语料库语言学家都做过交流，对于这些工具，各有所好，至于什么是最好的，我所听到的看法不一（有时矛盾冲突）。我的建议是，新手应该努力让自己意识到拥有各种选择，但是，宁可善用一个低级别的工具，也不要误用一个高度复杂的工具。

语料库语言学的诸多分析方式多多少少都会用到语料库频率（某个词、词串或是标注码单独出现或是与类似特征毗邻出现的次数）。所给出的频率可以是原始数据，例如，yes 这个词在 BNC 中出现了 58,521 次，或者也可以是百分比，例如，yes 在语料库中的百分率是 0.059525。如果我们在两个语料库中比较某个词的出现频率，特别是倘若语料库规模不等，那么明智的做法是，既给出原始频数，也给出百分率，以便更加易于比较。然而，因为很多词相对生僻，百分率可能小到难以想象，所以还会经常用到另外一种形式的比例数据，即给出每百万词中某词的出现次数 x。例如，我们可以写道，在 BNC 中，yes 每百万词出现 595.25 次。该比率较之 0.059525%，要更好理解一些。本书在第四章对频率有非常细致的探讨。

因为语料库常常包括来自诸多不同渠道的语言样本，所以重要的是要考虑数据中的语言型式在多大程度上有所体现，是在整个语料库中，还是仅在部分语料库中。例如，词语 lady doctor 在 BNC 中出现了 13 次。尽管

这是比较生僻的词汇，但是如果所有这 13 例出现在了不同的文本（并且出现在了不同的语类或语域），那么，与该词仅限于一两个文本相比，我们就拥有了更多的证据来说明，这是一个使用范围更加广泛、更加为人所熟知的词语（我在第二章，对有关分布的问题有更为详细的讨论）。

词频表（frequency list）是一个由语料库中最频繁出现的单词（或单词序列或标注码）所生成的列表，排序按照词频或字母顺序。词频表有助于揭示语料库中的常见型式或主题。不过，诸多语言都包含很高比例的闭合类语法词，譬如冠词、介词、连词和代词，于是这些词经常位于词频表前列。这会告诉我们有关语言的某些总体情况，但是却不能告诉我们某具体语料库所独有的特征。为此，就会经常用到另外一种被称之为主题性（keyness）的测量方式。主题性建立在对语料库中的每一个词都进行统计检验，将其频率与更大规模的参照语料库做比较（或者在有些情况下，把两个规模相当的语料库放在一起比较，二者互为参照）。参照语料库的作用是为语言的正常频率提供一个标准量度，那么在另外一个语料库中，相对更为频繁出现的任何词语都将被视为具有主题性，这时若对数据进行对数似然比或方差检验，p 值为低值。那么，主题词（keywords）未必非得是高频词，只是相较而言，频率高于预期。我在第二章，对主题词有深入探讨；在第七章，研究了一个与之相关的现象，即主题性语义码（key semantic tags）。

搭配（collocation）与频率相关，或者说某些词吸引另外一些词的趋势，搭配词出现时要么紧挨着，要么靠得比较近。搭配词有助于揭示意义是如何通过语言的重复使用而获得的，这是由于某些概念随着时间的推移而紧密联系在了一起。“强”搭配理论指出，如果一些词足够频繁地共现，而且彼此都与其他词存在互斥关系，那么，这些词就会对某些概念产生心理触发作用。在第五章和第六章，我分别对搭配词如何能够揭示性和性别的相关话语进行了深入探讨。此外，在第七章，我还通过观察搭配网络（collocational networks），研究如何把更为复杂的单词关系也考虑在内。

尽管上述各分析方式都是以量化为基础的，但还有另外一种分析方法会经常用到（尤其是当语料库语言学用于研究社会问题时），这就是索引（concordancing）。索引是一个一览表，包括某一语言特征（单词、词—码组合、标注码或者序列或者单词或标注码的组合）与位于其左右两侧直接

语境中的若干词共同出现的所有实例。索引能够使语料库语言学者基于对某个词所出现语境和型式的理解，开展功能性的定性分析。为了使之更容易操作，索引可以对位于检索词（短语、标注码等）左右两侧的词，按照字母顺序排列。

以表 1.1 为例，这是从 BNC 报纸文章中提取的 feminist 及其复数形式的索引行样本。尽管所提供的每一例语境均十分有限，但是我却很有把握地发现了大量看似涉及类似概念的词语——axe-grinding、blame、irk、vociferous、battle-cry、vehemently、enraged、militant、outraged、outrage。总的来看，大部分词似乎都指向“女性主义者好战寻衅”的话语韵（discourse prosody）（Stubbs 2001：65）。女性主义者的负面刻板印象在态度研究（attitude research）（见 Goldberg *et al.* 1975）中已有发现。Williams 和 Wittig（1997）提出，这些负面刻板印象可能会妨碍女性主义者的自我标签（self-labelling）。那么，BNC 中的这些例子说明了记者们如何通过把两个概念（女性主义和愤怒）反复捆绑在一起，使得这种联系在读者心里得以固化，从而集体作用于女性的负面刻板印象。

表 1.1　feminist(s) 的样本，取自英国国家语料库中的报纸部分

the important thing is the play, not any	feminist	**axe-grinding**. I've really no ambitions
has caused this terror of female flesh?	feminists	can't **blame** it all on men — since time
[with-a] brain.' Most likely to **irk** the	feminists	is 'F O long blonde hair,' referring to
defined by a tiny, **vociferous** group of	feminist	leaders, is in direct contradiction to
was a justifiable **battle-cry** of the	feminist	movement and that ambition was
'native races'. Most seriously, the only	feminists	of any note who were not **vehemently**
restore their 'caring' image among the	feminists	they **enraged** by opting out of Europe's
this type of dotty left-wing and **militant**	feminist	thinking,' she said. 'Now they want to
of sexual harassment years earlier, and	feminists	were **outraged** when the court rebuffed
remarks which caused **outrage** among	feminists.	He admitted his language was

当某个词在一个语料库中特别频繁地出现，所产生的索引行数以百计或千计时，我们可能就会想到要观察索引行的随机样本，或者只考虑那些还含有最高频或最显著搭配词的索引行。我在第五章研究对男同性恋者的表征时，采用的就是这种方法。索引相当于一种降采样（down-sampling）的方式，因为我们聚焦在了语料库中更为少量的数据，这时涉及的只是含有某个语言项的数据。不过，还可以纳入另外一种降采样的方式，这就是挑选出那些可被视为对语料库特别具有典型性的文本。例如，我们可能会从一个由 1,000 个报纸文章组成的语料库中提取一个主题词列表，然后决定以含有最大数目主题词为标准，对缩减数量之后的文章开展更加精细的定性分析。再有，降采样的方式还可能包括使用词频，辨别语料库中的常见主题或概念，然后聚焦于提及那个话题或概念的文本，开展更为详细的分析。这是我在之前项目中所采用的一个方法，当时我研究的是英国报纸中有关穆斯林的文章。对词频的初步分析说明，某些话题是很流行的，譬如“戴着面纱的女人”，这为进一步开展更为细致的研究指明了方向（见 Baker *et al.* 2013）。那么可以认为，语料库语言学有望为文本研究提供一系列方法：从频率的统计检验，到指引分析者对那些尤具代表性特征的文本（或部分文本）开展定性分析。

那么，在对语料库语言学的一些主要特点做过讨论之后，本章最后要简要描述一下本书接下来的内容。

本书概述

以下两章研究的是身份和语言用法的特点，不过所采用的方法各不相同。就此，有必要先对 Tognini-Bonelli（2001）在语料库驱动（corpus-driven）和基于语料库（corpus-based）的调查之间做出的区分进行简要概括。前者把语料库用作数据，其中的模式被视为表达语言规律（和例外）的一种方式。研究者对语料库数据“一无所知”，不施加任何一种预先存在的范畴体系，而是使用基于词频的计算程序，使语料库自己来推动分析的进展。例如，分析的路径可能会基于一个无法提前预知的主题词列表。与之不同的是，基于语料库的调查更多是把语料库作为一种检验研究直觉的方式，或者是作为一种在较小规模的数据集里研究语言的频率和 / 或其合

理性的方式。研究者心中可能有一套假设，并提前决定好要研究哪些词或短语，以对那些假设加以检验。

如同专用语料库和参照语料库之间的差别一样，把基于语料库的研究和语料库驱动的研究视为一个渐变体的两级而不是一个二分对立的选择或许更容易。例如，McEnery（2006：8）等指出，以一个完全一无所知的方式来处理语料库是很困难的，也很难做到不对语言数据施以某种范畴体系（即便像是关键词这样一也需要把“词”归为一类语言范畴）。本书第二章主要采用语料库驱动的方法，研究语言的性别差异。该章在对语料库语言学者之前所开展的研究予以综述之后，对 BNC 中男性和女性语言进行了比较。笔者提出的研究问题是：如果我仅把男性或者仅把女性的语言随机划分为两组进行比较，那么其呈现的差异是否小于这两组数据之间的差异？对于其他研究者通过观察 BNC 而假定性别差异存在的一些发现以及在多大程度上，这些差异实际具有可推广性，笔者也尝试予以解释。第三章讨论的问题依然是语言的用法和性别分析。但是为了从将男女性别对立起来的研究中走出来，我仅把重点放在了女性学术导师的语言中，研究她们是如何表达异议这种语言形式的。有些语言和性别研究的理论认为刻板印象中，女性更有可能采用合作的、不具威胁性的言语形式，那么，执掌权位的女性如何讲话就成为一项非常有趣的研究课题，因为为了在这样的语境中获得成功，女性的言行就得违背其刻板印像（可能会因此而面临非议）。正如其他学者所注意到的，当女性经理人和领导以对抗性的身份进行协商时，很有可能会导致相当复杂的语言形式。因此本章采用了基于语料库的方法。也就是说，在我开展语料库调查之前，就决定了要研究一种特定的语言形式（表示异议）。笔者建立了一套分类体系，从她们是否考虑听众的面子而采取礼貌态度的角度，来对异议进行归类。此外，本章并未停留在检索语料库中的简单词语上，而是想要发现在表达异议时所出现的所有情况 - 这是一个更加充满变数的现象。

第四章不再关注女人和男人如何使用语言，而是聚焦于他们如何被语言表征。该章调查的是一个可回溯至 19 世纪早期的美国英语大型历时参照语料库，在研究过程中对词频信息的价值予以了充分考虑。笔者讨论了如何比较诸如 man/woman, wife/husband 这样一些性别化词对的频率，以证明

语言中存在男性偏向（male bias）的现象。追踪词频如何随时间的推移而发生变化，有助于证实抵制语言性别歧视或隐性偏见的努力是否已见成效。笔者引用了了大量实例以告诫人们不要过度解释。例如，chairman 这个词未必总是指男人，father 也并非总是指有子女的男人。一个词高频出现，也未必说明人们认为这个身份是有权势的（wife 的词频就高于 husband）。除了观察单个词的频数之外，笔者还仔细研究了词序。那么，比起 women and men，人们更加频繁书写 men and women 吗？这条规则的例外情况能告诉我们什么呢？本章最后研究的是一些性别中性的词语，比如说 police officer 和 Ms。那么，在最近的几年里，这些词语是否已经对美国人语言的使用产生了影响？

在第五章，笔者继续关注表征，研究在由英国大众报纸《每日邮报》的文章所构建的语料库中，同性恋人群是如何被书写的。与其他研究者的观点一样，笔者认为不考虑性倾向，就难以理解有关性别的话语。在本章前半部分，笔者述及 Butler 的异性恋矩阵。其中有一个例子谈到，诸如 man 和 woman 这样一些词语时常含而不露地表示异性恋，正如 Connell 所言，霸权男性气质（hegemonic masculinity）的关键要素是基于“异性恋男人的主导地位和同性恋男人的从属地位”（1995：78）。在早前发表的一项研究中（Baker 2005：60-92），笔者调查了英国报纸《每日邮报》在 2001 年到 2002 年间，如何通过大量的关于同性恋人群的文章，将他们塑造成淫乱的、无耻的、丢脸的、暴力的和政治上激进的负面形象。在第五章，笔者重新再看这家报纸，观察正在发生改变的社会态度是否已经影响了 2008 年到 2009 年间同性恋人群表现的方式并且重新回顾了之前所采用的研究方法，即对 gay 和 homosexual 的搭配词加以识别，然后通过阅读含有那些搭配词的索引行，来推定韵律存在的证据。笔者还反思了这种方法是否依然存在某种形式的“挑樱桃”之嫌（例如，我当初发现的模式之所以主要是负面的，是否正是因为我心中有那样的期待）。最后，笔者谈到了分析扩展索引行的优势，认为这能够使那些跨越多个句子的特征得以识别，譬如合法化策略（legitimation strategies）。

对搭配词予以更加深入的审视，包括探讨不同的搭配方法，考虑一些相关的问题，比如说：搭配跨距应该为多大，应该采用基于信心

（confidence-based）的技术还是假设检验（hypothesis-testing）的技术（或是二者兼用）。笔者首先对使用搭配词研究性别表征问题的一些重要的且基于语料库的研究进行了综述，然后研究了搭配词如何通过使用在线工具“Sketch Engine”，根据语法模式进行自动归类。通过使用了一个由英国网页文本构建的大型语料库，笔者对 boy 和 girl（及其复数形式）的搭配关系进行比较，目的是发现这些身份在得以稳定建构的方式上，存在哪些异同之处。例如，男孩更有可能被那些对行为好坏予以评价的搭配词修饰，而 girl（boy 却并非如此）经常被用以指成年人。本章最后围绕对这些研究结果的解释、阐释以及批评性评价，对相关问题进行了讨论。

作为分析的最后一章，第七章研究了由取自克雷格列表网（Craigslist）的男性交友广告所构建的三个小型语料库，由此把之前所探讨的性别化用法和表征的各个主题有机联系在一起。笔者首先通过这一章来例示说明语料库建设者在决定自建语料库之后所可能遇到的一些问题。例如，从互联网下载资料，去除样板文件（boilerplate），识别和去除无用文件、重复文本以及对于文件的标注。此外，笔者还对几个尚存争议的问题进行了探讨。当大多数语料库分析软件目前只允许对两个语料库开展比较时，如何比较三个或三个以上的语料库？当使用不同规模的语料库开展研究时，如何保证临界点的一致性？本章还旨在探讨在多大程度上，各种不同的语料库分析技术会得出各种不同的结果。为此，这项分析对相同数据集采用了三角互证（triangulation）的方法，包括对主题性语义码、搭配网络和索引行的分析。

最后在结论这一章，笔者对本书的主要发现进行了总结，同时对各章的研究结果进行了反思，对所使用的不同方法做出了批判性评价，并试图对采用语料库方法分析性别所可能具有的局限性进行了探讨。本书最后对语料库和性别研究的未来发展方向做出了思考。