应用语言学中的语料库

Corpora in Applied Linguistics

Susan Hunstom 著

冯志伟 导读

世界图书出版公司 剑桥大学出版社 联合出版， ISBN: 7-5062-8210-0/H-889, 2006 年 8 月

目 录

《应用语言学中的语料库》导读

冯志伟

编者按：本文首先简要回顾了语料库语言学的兴起及国外语料库的概况，然后介绍了中国语料库的发展情况，阐述了语料库在语言学各学术领域的研究中所发挥的作用，接着介绍 Corpora in Applied Linguistics 一书的作者 Susan Hunston，并对书的各章内容进行了引领导览，旨在使读者对语料库研究及本书能够得到一个鸟瞰式的认识。

语料库语言学的兴起

英国著名哲学家罗素（Bertrand Arthur William Russell）曾经用两个金字塔来比喻西方两大传统哲学流派的研究方法，他说（1976:177-178）：“方法的不同可以这样来刻画其特征……（要么）在针尖似的逻辑原则上按倒金字塔式矗立起一个演绎巨厦……假若原则完全正确而步步演绎也彻底牢靠，万事大吉；但是这个建筑不牢稳，哪里微有一点裂罅，就会使它坍倒瓦解。……（或者）金字塔基底落在观测事实的大地上，塔尖不是朝下，是朝上的；因此平衡是稳定的，什么地方出个裂口可以修缮而不至于全盘遭殃。”这里，倒立的金字塔用来比喻理性主义的研究方法，正立的金字塔则用来比喻经验主义的研究传统。

在 20 世纪 50 年代以前，现代语言学的传统，无论是规范语言学、历史语言学或是描写语言学，都注重语言事实，提倡经验主义，即“根据对大量事实的广泛观察，得出一个比较有限的结论”（罗素，1976:177）。美国语言学家乔姆斯基（Noam Chomsky）自 1956 年开始发表有关形式语言的一系列论文，在 1969 年的 Quine's Empirical Assumptions 一文中他说：“然而应当认识到，‘句子的概率’这个概念，在任何已知的对于这个术语的解释中，都是一个完全无用的概念。” 可见，乔姆斯基早期完全排斥经验主义的统计方法。他主张采用公理化、形式化的方法，严格地按照一定的规则来描述自然语言的特征，试图使用有限的规则描述无限的语言现象，发现人类普遍的语言机制，建立所谓的“普遍语法”。自此形成了转换生成语法的研究途径，60 年代末到 70 年代时期在美国兴盛一时，也大力推动了机器翻译（Machine Translation，简称 MT）和自然语言理解（Natural Language Understanding，简称 NLU）的研究和发展。

转换生成语法的研究途径在一定程度上克服了传统语言学的某些弊病，推动了语言学理论和方法论的进步，但它认为统计只能解释语言的表面现象，不能解释语言的内在规则或生成机制，渐渐远离经验主义的途径。这种转换生成语法的研究途径实际上承继了“理性主义”的哲学思源。经验主义和理性主义两者之间的争论主要体现在知识论的问题上：在英国以培根（Francis Bacon）、洛克（John Locke）等人为代表的经验主义传统（empiricist tradition）主张，知识产生的途径是根据外界世界的数据和经验来进行归纳和推理的过程，而在欧洲大陆以笛卡儿（René Descartes）等人为代表的理性主义传统（rationalist tradition）则提倡学习和推理的途径是由先验的知识和与生俱来的思想所指导的。

然而，人们逐渐发现，这种理性主义的研究所得出的语言规则似乎只能适用于一种子语言（sub-language），而不能推广到该子语言之外的于其他语言现象，具有很大的局限性。人们开始思考，乔姆斯基的“普遍语法”是否是真正的语言规则，是否能够经受大量的语言事实的检验，语言规则是否应该和语言事实结合起来考虑，而不是一头钻入理性主义的隧道？作为一位求实求真、虚怀若谷的语言学大师，乔姆斯基开始反思，表现了与时俱进的勇气。

在最近他提出的“最简方案”中，他认为，所有重要的语法原则直接运用于表层，不同语言之间的差异通过词汇来处理，把具体的规则减少到最低限度，开始注重对具体的词汇的研究。可以看出，转换生成语法也开始对词汇重视起来，逐渐地改变了原来的理性主义的立场，开始与经验主义妥协，或者悄悄地向经验主义复归。

由于语言学中经验主义方法的东山再起，注重语言事实的传统重新抬头，大多数学者们普遍认为：语言学的研究必须以语言事实作为根据，必须详尽地、大量地占有材料，才有可能在理论上得出比较可靠的结论。传统的语言材料的搜集、整理和加工完全是靠手工进行的，这是一种枯燥无味、费力费时的工作。尽管一些对于语言研究有浓厚兴趣和献身精神的语言学家对于这样的工作乐此不疲，但是一般的人对此却望而生畏。计算机出现之后，随着计算机功能的逐渐完善和强大，原先完全靠手工的工作开始交由计算机去做，大大地减轻了人们的劳动。后来，在这种工作中逐渐创造了一些独特的方法，提出了一些初步的理论，形成了一门新的学科 — 语料库语言学（corpus linguistics），由于语料库是建立在计算机上的，因此，语料库语言学是语言学和计算机科学交叉形成的一门边缘学科。

在目前的研究水平下，语料库语言学主要是利用语料库对于语言的某个方面进行研究，仅仅是一种新的研究手段。严格地说，语料库语言学还没有十分完备的理论，它还不能跟语言学中的其他成熟的学科（如计算语言学、社会语言学、心理语言学）相提并论。尽管这样，这个新兴的研究领域一出现，就引起了语言学界的普遍关注，越来越多的语言学家愿意采用语料库作为他们的工具来研究语言，并取得了令人可喜的成绩。

目前，语料库语言学主要研究机器可读自然语言文本的采集、存储、检索、统计、语法标注、句法语义分析，以及具有上述功能的语料库在语言教学、语言定量分析、词汇研究、词语搭配研究、词典编纂、语法研究、语言文化研究、法律语言研究、作品风格分析、自然语言理解和机器翻译等领域中的应用。

建立和使用语料库的意义

语料库语言学是以语料库作为研究对象的。这样的语料库必须以电子计算机为载体来存放语言材料，这些存放在电子计算机中的语言材料是在语言的实际使用中真实出现过的，因此，它们可以如实地反映语言现象，克服语言学家观察语言现象时的主观性和片面性，这样的未经加工的语料对于语言学研究已经很有用；而这些真实的语言材料需要经过分析、加工、处理之后，就可以变成更加有用的语言资源。所以，不论是未经加工的“生语料”或者经过加工的“熟语料”都是非常宝贵的。

多年来，机器翻译和自然语言理解的研究中，分析语言的主要方法是句法语义分析。因此，在很长一段时间内，许多系统都是基于规则的，而根据当前计算机的理论和技术水平，很难把语言学的各种事实和理解语言所需的广泛的背景知识用规则的形式充分地表达出来，这样，这些基于规则的机器翻译和自然语言理解系统只能在极其受限的某些子语言（sub-language）中获得一定的成功。为了摆脱困境，自然语言处理的研究者们开始对大规模的非受限的自然语言进行调查和统计，以便采用一种基于统计的模型来处理大量的非受限语言。不言而喻，语料库语言学将有可能在大量语言材料的基础上来检验传统的理论语言学基于手工搜集材料的方法所得出的各种结论，从而使我们对于自然语言的各种复杂现象获得更为深刻和更为全面的认识。

传统语言学基本上是通过语言学家归纳总结语言现象的手工方法来获取语言知识的，由于人的记忆能力有限，任何语言学家，哪怕是语言学界的权威泰斗，都不可能记忆和处理浩如烟海的全部的语言数据，因此，使用传统的手工方法来获取语言知识，犹如以管窥豹，以蠡测海，这种获取语言知识的方法不仅效率极低，而且带有很大的主观性和片面性。传统语言学中啧啧称道的所谓“例不过十不立，反例不过十不破”的朴学精神，貌似严格，实际上，

在浩如烟海的语言数据中，以十个正例或十个反例就轻而易举地来决定语言规则的取舍，难道就能够万无一失地保证这些规则是可靠的吗？这是大大地值得怀疑的。在计算机上建立了语料库之后，我们就可以使用机器学习的方法，自动地从浩如烟海的语料库中获取准确的语言知识。这是语言学获取语言知识方式的巨大变化，作为二十一世纪的语言学工作者，都应该注意到这样的变化，逐渐改变获取语言知识的手段。

语料库是语言知识的宝库，是最重要的语言资源。语料库中蕴藏着丰富的语言知识，词汇知识、句法知识、语义知识、语篇知识，都包含在语料库当中。随着语料库加工的逐渐精细和深入，我们获得的语言知识也就越加准确和深刻。

语料库同时也是语言学家有力的研究工具。语料库的使用，为语言学的研究提供了一种新的思维角度，辅助人们的语言“直觉”和“内省”判断，从而克服研究者本人的主观性和片面性，逐渐成为语言学研究的主流方法。语言学家利用语料库来研究语言学，正如天文学家利用望远镜来研究天文学，生物学家利用显微镜来研究生物学一样，能够使他们如虎添翼，其意义是非常重大的。望远镜的发明使天文学家能够观察到他们过去难以观察到的宏观世界的现象，显微镜的发明使生物学家能够观察到他们过去难以观察到的微观世界的现象，计算机可读的语料库就好比语言学研究的望远镜和显微镜，语料库的使用扩展了语言学家的眼界，使他们看得更远，看得更细，从而使他们能够发现更多的语言现象，挖掘出更多的语言事实，把语言学的研究推向一个新的阶段。从某种意义上说，语料库的使用，是语言学研究的一次革命性的进步。

需要指出的是，语料库并不是全部的研究方法和手段。它的局限性在于，语料库只能提供语言事实的例证，但是不能对之进行解释，不能进行推理，也不能为文本数据直接地提供文化和社会背景等方面的信息。它在辅助人们的语言“直觉”和“内省”判断的同时，离不开研究者本人的语言“直觉”和“内省”，因为，科学研究中的客观知识离不开主观知识，就像主观知识离不开客观知识一样。

历史上的语料库

1959 年，英国伦敦大学教授 Randolph Quirk 提出建立英语用法调查语料库，叫做 SEU（Survey of English Usage），后来他根据这个语料库领导编写了著名的《当代英语语法》。不久，Nelson Francis 和 Henry Kucera 在美国 Brown 大学召集了一些语料库的有识之士，建立了 BROWN 语料库（布朗语料库），这是世界上第一个根据系统性原则采集样本的标准语料库，规模为 100 万词次，是一个代表当代美国英语的语料库。由英国 Lancaster 大学 Geoffrey Leech 教授倡议，由挪威 Oslo 大学的 Stig Johansson 教授主持完成，最后在挪威 Bergen 大学的挪威人文科学计算中心联合建立了 LOB 语料库（LOB 是 London，Oslo 和 Bergen 的首字母简称），规模与 Brown 语料库相当，这是一个代表当代英国英语的语料库。欧美各国学者利用这两个语料库开展了大规模的研究，其中最引人注目的是对语料库进行语法标注的研究。20 世纪 70 年代。Greene 和 Rubin 设计了一个基于规则的自动标注系统 TAGGIT 来给布朗语料库的 100 万词的语料做自动词性标注，正确率为 77%。Geoffrey Leech 领导的 UCREL （University Centre for Computer Corpus Research on Language）研究小组，根据成分似然性理论，设计了 CLAWS （Constitute Likelihood Automatic Word-tagging System）系统来给 LOB 语料库的 100 万词的语料做自动词性标注，根据统计信息来建立算法，自动标注正确率达 96%，比基于规则的 TAGGIT 系统提高了将近 20%。最近他们同时考察三个相邻标记的同现频率，使自动语法标注的正确率达到 99.5%。这个指标已经超过了人工标注所能达到的最高正确率。

20 世纪 60 年代初，英国伦敦大学 Randolph Quirk 教授主持的英语用法调查研究课题组曾经收集了 2000 个小时的谈话和广播等口语素材，并把这些口语素材整理成书面材料，后

来，瑞典 Lund 大学教授 J. Svartvik 主持，把这些书面材料全部录入计算机，在 1975 年建成了 London-Lund 英语口语语料库，收篇目 87 篇，每篇 5000 词，共为 43.4 万词，进行了详细的韵律标注（prosodic marking）。

以上这三个语料库都储备在挪威 Bergen 大学的国际现代英语计算机档案（International Computer Archive of Modern English，简称 ICAME）的数据库中。

20 世纪 80 年代以后，陆续建立了一些以词典编纂为应用背景的大规模语料库。在 John Sinclair 教授的领导下，英国伯明翰大学（Birmingham University）与 Harper Collins 出版社合作，建立了 COBUILD 语料库（Collins Birmingham University International Language Database，首字母缩写就是 COBUILD）。1987 年，Collins 出版社出版了建立在 COBUILD 语料库基础上的英语词典，词条选目、用法说明和释义都直接来自真实的语料，由 John Sinclair 教授担任总编辑，COBUILD 词典出版后，得到读者的广泛好评，影响很大，现在又出版了各种用途的 COBUILD 词典，并编写英语课程教科书（COBUILD English Course）。2003 年这个语料库的规模已经达到 5 亿词次，其中包含 1500 万词次的口语语料库。这个大规模的 COBUILD 语料库，又可以叫做“英语银行”（Bank of English）。

20 世纪 80 年代还建立了 Longman 语料库，也应用于词典编纂。这个语料库由 LLELC（Longman Lancaster 英语语料库）、LSC （Longman 口语语料库）和 LCLE （Longman 英语学习语料库）等三个语料库组成。这个语料库主要用于编纂英语学习词典，帮助外国人学习英语。规模为 2000 万词次。

由于这些语料库可直接用于词典编纂，在商业上获得了成功，语料库语言学的研究开始从纯学术走向实用，词典编纂是语料库语言学发展的推动力之一。

美国计算语言学学会（The Association for Computational Linguistics，ACL）发起倡议的数据采集计划（Data Collection Initiative，DCI），叫做 ACL/DCI，这是一个语料库项目，其宗旨是向非赢利的学术团体提供语料，以免除费用和版权的困扰，用标准通用置标语言 SGML（Standard General Mark-up Language，ISO 8879，1986 年公布）和文本编码规则 TEI（Text Encoding Initiative）统一地对语料库进行置标，以便于数据交换。这样的工作是很有价值的，它为语料库在不同计算机环境下进行数据交换奠定了基础。ACL/DCI 的语料范围广泛，包括华尔街日报语料库、Collins 英语词典、Brown 语料库，还有双语和多语的语料。

80 年代末 90 年代初，美国 Pennsylvania 大学开始建立“树库”（Tree bank），对百万词级的语料进行句法和语义标注，把线性的文本语料库加工成为表示句子的句法和语义结构的树库。这个项目由 Pennsylvania 大学计算机系的 M. Marcus 主持，到 1993 年已经完成了 300 万词的英语句子的深加工，进行了句法结构标注。

在美国 Pennsylvania 大学还建立了 LDC 语言数据联合会 （Linguistic data Consortium），实行会员制，有 163 个语料库 （包括文本的以及口语的）参加，共享语言资源。2000 年，LDC 发行了一个中文树库，包含 10 万词，4185 个句子，这是世界上第一个中文的树库，可惜的是规模比较小。

国外比较著名的语料库还有：

- AHI 语料库：美国 Heritage 出版社为编纂 Heritage 词典而建立，有 400 万词。
- OTA 牛津文本档案库（Oxford Text Archive）：英国牛津大学计算中心建立，有 10 亿字节。
- BNC 英国国家语料库（The British National Corpus）：1995 年正式发布，使用文本编码规则 TEI 编码和通用标准置标语言 SGML 的国际标准，有 1 亿词次，其中书面语 9000 万词次，口语 1000 万词次。
- RWC 日语语料库：日本新情报处理开发机构 RWCP 研制，包括《每日新闻》4 年的全文语料，语素标注量达 1 亿条。

- 亚洲各语种对译作文语料库：日本国立国语研究所研制，中野洋主持，北京外国语大学参加。

为了推进语料库研究的发展，欧洲成立了 TELRI 和 ELRA 等专门学会。TELRI 是跨欧洲语言资源基础建设学会（Trans-European Language Resources Infrastructure ）的首字母缩写，John Sinclair 担任主席，Wolfgang Teubert 担任协调员，由欧洲共同体提供经费，其目的在于建立欧洲诸语言的语料库，现已经建成柏拉图（Plato）的《理想国》（Politeia）多语语料库，建立了计算工具和资源的研究文档 TRACTOR （Research Archive of Computational Tools and Resources），正在语料库的基础上建立欧洲语言词库 EUROVOCA。TELRI 每年召开一次研讨会。我有幸曾多次参加 TELRI 的学术会议和部分研究工作，在 TELRI 的学术会议上发表过多篇论文。

ELRA 是欧洲语言资源学会 （European Language Resources Association）的首字母缩写，由意大利比萨大学 Zampolli 教授担任主席，ELRA 负责搜集、传播语言资源并使之商品化，对于语言资源的使用提供法律支持。ELRA 建立了欧洲语言资源分布服务处 ELDA（European Language resources Distribution Agency），负责研制并推行 ELRA 的战略和计划。ELRA 还组织语言资源和评价国际会议 LREC（Language Resources & Evaluation Congress），每两年一次。第一次会议于 1998 年在西班牙的 Grenade 举行；第二次会议在 Athens（Greece）召开（31.May – 02.June.2000），第三次会议于 2002 年在西班牙的 Las Palmas de Gran Canaria 召开（27.May – 02.June 2002），第四次会议在 2004 年 6 月在葡萄牙的里斯本举行。我有幸曾担任 LREC 国际顾问委员会的成员，积极参与了国际语料库研究的学术交流活动。

中国的语料库状况

从 1979 年以来，中国就开始进行机器可读语料库（machine-readable corpus）的建设，早期在中国建立的主要的机器可读语料库有：

- 汉语现代文学作品语料库（1979 年），527 万字，武汉大学。
- 现代汉语语料库（1983 年），2000 万字，北京航天航空大学。
- 中学语文教材语料库（1983 年），106 万 8 千字，北京师范大学。
- 现代汉语词频统计语料库（1983 年），182 万字，北京语言学院。

早期的这些语料库多数是采用手工键入的方式建立的，耗时耗力，缺乏规范，规模较小，重用性差。为了建设这样的语料库，需要付出艰辛的劳动，北京航空航天大学计算机系刘源教授在该校 2000 万字的语料库建设中积劳成疾，健康受到严重的损害，不幸早逝。我国语料库的早期建设者的敬业精神是值得我们尊敬的。

北京航空航天大学的语料库还进行了词频统计和汉语书面文本自动分词研究，发现了两种不同的分词歧义字段：交集型歧义字段和多义组合型歧义字段：

- 交集型歧义切分字段：例如：“地面积”可能切为“地面”或“面积”，“面”成为交段，从而产生歧义。
- 多义组合型歧义切分字段：例如：“马上”本身是一个词，但也可以切为“马”+“上”两个单词，而“马上”与“马”+“上”的含义不同。

他们曾对一个 48092 字的自然科学、社会科学样本进行了统计：交集型切分歧义 518 个，多义组合型切分歧义 42 个。据此推断，中文文本中切分歧义的出现频度约为 1.2 次/100 字，交集型切分歧义与多义组合型切分歧义的出现比例约为 12:1。

为了推动汉语语料库的深入研究，我国还建立了初步的分词规范：1990 年 10 月，在计算机界和语言学界的共同努力下，我国制定了国家标准 GB-13715《信息处理用现代汉语分词规范》，这个国家标准提出了确定汉语单词切分的原则，是汉语书面语自动切词的重要依据。

1991 年，国家语言文字工作委员会（现已并入国家教育部）开始建立国家级的大型汉语语料库，以推进汉语的词法、句法、语义和语用的研究，同时也为中文信息处理的研究提供语言资源，计划其规模将达 7000 万汉字，当时宣称，这将成为世界上最大的汉语语料库。这个语料库是均衡语料库，其语料要经过精心的选材，语料的选材应受到如下限制：

① 时间的限制：语料描述具有历时特征，着重描述共时特征。选取从 1919 年到当代的语料（分为 5 个时期），以 1977 年以后的语料为主。

② 文化的限制：主要选取受过中等文化教育的普通人能理解的语料。

③ 使用领域的限制： 语料由人文与社会科学类、自然科学类和综合类 3 大部分，人文和社会科学再分为 8 大类 29 小类，自然科学再分为 6 大类，综合类再分为 2 大类。主要选取通用的语料，优先选取社会科学和人文科学的语料。

为了加工这个国家级语料库，国家社科基金设立了社科重大项目“信息处理用现代汉语词汇研究”，希望利用该项目的成果来加工这个语料库。该课题分 10 个子课题：

① 信息处理用现代汉语分词词表

② 歧义切分与专有名词识别软件

③ 词的构造研究

④ 现代汉语词类及标记集规范

⑤ 汉语词类兼类研究

⑥ 现代汉语的语法属性描述研究

⑦ 现代汉语述语动词机器词典和槽关系研究

⑧ 汉语知识词典建立及词汇内部语义网络描述研究

⑨ 汉语文本短语结构的人工标注

⑩ 常用动词语义特征及词义搭配研究

现在，该课题已经结项，国家教育部语言文字应用研究所成立了“汉语语料库深加工”的课题组，已经完成了 7000 万字语料的深加工，正在逐步地把这个生语料库变为熟语料库。

1992 年以来，大量的语料库在中国研究中文信息处理的单位建立起来，语料库成为了研究中文信息处理的基本语言资源。没有语料库的支持，中文信息处理的研究将会寸步难行。建设大规模真实文本语料库的单位有：《人民日报》光盘数据库、北京大学计算语言学研究所、北京语言大学、清华大学、山西大学、上海师范大学、北京邮电大学、香港城市大学、东北大学、哈尔滨工业大学、中国传媒大学、中国科学院软件研究所、中国科学院自动化所、北京外国语大学日本学研究中心、台湾中央研究院语言研究所（筹备处）。

例如，中国传媒大学的语料库包括文本语料库（7000 多万字）、音视频语料库（900 小时的音频和视频语料）和精品语料库（如著名主持人的节目、获奖节目的音频视频语料），这是世界上规模最大的、多模态的汉语传媒有声语言的语料库，语料库加工体系从语音开始，到文字、词语、句子、篇章都进行了标注和处理。

我国语料库的建设与语言学研究有着密切的关系。例如，在中国传媒大学语料库的基础上，进行了汉语同类词短语的研究、汉语插入语的研究、网络语言研究、汉语熟语标记研究、汉语“有”字句研究、汉语“吧”字研究、汉语“然后”研究、主持人韵律特点研究等。语料库成为了语言学研究的语言资源，又成为了语言学研究的工具，有力地推动了语言学研究的发展。

我国在 20 世纪 80 年代中期就建立了第一个英语语料库，即上海交大科技英语语料库，简称 JDEST（Jiao Da English for Science and Technology），这个语料库是由上海交通大学建成的。JDEST 的建成，为我国大学英语教学大纲的制定和词表统计做出了积极的贡献。这个语料库当时在欧洲受到语料库语言学界广泛关注，JDEST 成为国际第一代语料库。后来在中国建成的英语语料库还有：ICLE 中国子语料库、中国英语学习语料库、大学学习者英语口语语料库、中国专业英语学习者口语语料库、CEC 中国英语语料库、中学英语口语语

料库等，这些英语语料库都与中国的外语教学和外语学习紧密相联。外语教学和外语学习是我国应用语言学的重要内容，是语料库推动我国应用语言学发展的又一个重要内容。

目前，语料库的深加工受到各国学者的普遍重视，很多国家都对语料库文本进行句法标注（syntactic annotation）和语义标注（semantic annotation），把语料库进一步加工成树库。例如，英语有英国 Lancaster-Leeds 树库、美国的宾州大学的 Penn 树库，德语有 TIGER 树库和 NEGRA 树库，捷克语有布拉格大学的 PDT 树库。汉语树库的建设也取得可喜的成绩，例如，清华大学的 TCT 树库、台湾中央研究院的 Sinica 中文树库、哈尔滨工业大学的汉语依存树库、中国传媒大学的依存树库、中国科学院计算技术研究所的汉语树库、美国的 Penn 中文树库等，都为成为语言资源的自动获取的重要工具。我们可以确有把握地说，树库建设将成为今后语料库研究的一个发展趋势。

总而言之，语料库给语言学研究提供了无比丰富的语言资源。很多几乎已经成为定论的语言规则需要我们根据语料库去重新认识和评价，许多新的语言学思想将从语料库的研究中产生出来。语言本身确实是无比复杂的，观察语言现象时，我们决不能掉以轻心，我们应当借助于语料库，更加努力地工作，从而推动语言学的发展。

本书及其作者

本书作者 Susan Hunston（苏珊•霍斯顿）是英国 Birmingham 大学英语系教授，2003 年我到 Birmingham 大学语料库中心访问时，曾经与她见过面，并且与她讨论过语料库语言学的一些问题。从交谈中，我知道她亲身参与过 John Sinclair 教授主持的 COBUILD 语料库的研制，专门负责语法方面的工作，是著名的语料库分析专家，是 COBUILD 语料库的核心成员；John Sinclair 教授退休后移居意大利托斯坎（Toscan）地区建立了“托斯坎词中心”（Toscan Word Center，简称 TWC），Susan Hunston 教授担任英语系主任，她继承了 John Sinclair 的传统，积极进行语料库的研究，并且把语料库与应用语言学的研究结合起来。

Susan Hunston 教授具有丰富的语言教学和应用语言学研究的经验，又掌握了 COBUILD 语料库研究的第一手资料，在这本《应用语言学中的语料库》中，她广泛地使用了 Bank of English 语料库中丰富的语言材料，把应用语言学与语料库密切地结合起来，对于如何在应用语言学中发挥语料库的作用，提出了许多独到的见解。

Susan 的研究十分细致。她指出，研究语料库中单词的频度（frequency）有助于发现出不同文体和语体的特点。例如，材料科学的语料中 surface，energy，electron，particles 等单词的频度高，政治语料中 International，policy，socialization 等单词的频度高，反映了频度的领域特性。在口语中，表示“必须”时多用 have to，而在比较严肃的文本中多使用 must，反映了频度的语体特点。

Susan 还指出，在语料库中，可以使用“词语索引”（concordance）来发现词语的不同用法（phraseology），从而建立词语的使用模式。例如，interested 通常使用于短语 interested in 中，其模式是 someone is interested in something；而 interesting 通常使用于名词前面，其模式是 an interesting thing。使用语料库的“词语索引行”（concordance lines）可以帮助我们发现词语的搭配（collocation）。例如，在“词语索引行”中，我们可以观察到 shed 和它后面的不同名词的各种搭配关系，从而具有不同的含义：shed light（阐明），shed tears（哭），shed blood（受害），shed stuff（开除），shed pound（减肥）。

Susan 认为，语料库还可以用于语言教学，教师在课堂中可以鼓励学生自己开发语料库来观察词语用法的细微差异，比较各种语言之间的不同。双语并行语料库可以帮助翻译工作者比较语言之间的差异。例如，通过语料库发现，英语的 still 可以翻译为法语的 toujours 或 encore，但是，有时包含 still 的英语句子在并行语料库中却找不到法语的等价翻译，而法语的 toujours 或 encore 却总是可以在英语句子中找到与它们等价的词语 still。语料库还可以用

于文体学、病理语言学、法律辩论语言学（forensic linguistics）的研究，用于调查语言所表达的文化态度以及作为批评性话语分析的语言资源。

本书还介绍了语料库的类型。Susan 认为，语料库的类型有：专用语料库（Specialized Corpus）、通用语料库（General Corpus）、比较语料库（Comparable Corpus）、并行语料库（Parallel Corpus）、语言学习者语料库（Leaner Corpus）、语言教学语料库（Pedagogical Corpus）、历史语料库或历时语料库（Historical or diachronic Corpus）、监控语料库（Monitor Corpus）。

Susan 讨论了语料库对于应用语言学的重要性和它的局限性。她认为，语料库对于应用语言学的重要性在于：语料库可以帮助人们克服语言“直觉”（intuition）的不足，语料库是收集和存储语言数据的一种手段，借助于“词语索引”和“频度”，语料库可以大大地改善人们的语言“直觉”，弥补了“内省”方法的不足，从而克服主观性和片面性。

Susan 认为，语料库的局限性是：语料库只能给我们关于某种语言现象频度高低的信息，但是不能给我们某种语言现象是否可能的信息；语料库只能告诉我们语言事实本身，但是不能进行推理；语料库只能给我们提供例证，但是不能对于例证提供解释，对于例证的解释要依靠语言的直觉和内省；文本语料库不能提供关于声调、手势以及社会背景方面的信息。

下面，我们介绍各章的主要内容，以便引导读者更好地理解此书。

第一章　语料库使用概述

本章是全书的导论。首先简短地介绍了全书的内容安排。然后介绍在应用语言学中语料库的各种用途，从而说明：使用语料库可以帮助人们发现语言学规律。

本章介绍了在应用语言学中使用语料库的一些基本概念，举出了语料库的一些可能的应用实例，同时也指出了语料库的某些局限性。

此外，本章还解释了后面各章将使用的一些关键性术语：token（词例），type（词型），hapax（一次出现词），lemma（词目），word-form（词形式），tag（标记），parse（剖析），annotate（标注）。

第二章　作为对象的语料库：它的设计和目标

本章讲述了语料库的设计和目标。讨论了语料库的规模大小，语料库的内容，语料库的平衡性和代表性，语料库的维护等问题。

从使用的角度，可以把语料库看成一篇篇文本的集合，或者看成带有上下文的单词的集合，或者看成是范畴的集合，从而从不同的角度对语料库进行抽象的解释和研究。

第三章　语料库语言学中的方法：对于词语索引行的说明

本章着重介绍语料库研究中广泛使用的“词语索引行”（concordance line）。

为语料库构造“词语索引行”是语料库信息处理的最重要的途径。大多数的语料库用户都是使用词语索引行来分析和解释语料的。

使用词语索引行进行“词语索引”（concordance）时，“词语索引程序”（concordancer）把我们感兴趣的单词或短语放在计算机屏幕的中心，而把与这些单词或短语有关的上下文放在它们的左边和右边，以便于我们观察和研究。放在屏幕中心的单词或短语叫做“结点词”（node word）。

要搜索的对象可以是具体的单词（如 point），也可以是词目（词目用大写字母表示，如

词目 CONDEMN，包括它的各种变化形式 condemn，condemned，condemning），也可以是单词的序列（如“BE + 形容词” is clear，was clear，is important，was important 等）。有时，搜索的对象可以不是具体的词语，而是在某种情况下为表示某个概念而使用的词语（如，当表示假定的时候，经常使用 what would）。如果一次搜索达不到目标，可以进行一次以上的搜索，以便找到搜索的目标。

词语索引行的排列可以是随机的，不一定要排序，当然也可以按字母表来排序。词语索引行还可以进行选择或组合，以说明特定的行为。

在本章中，词语索引行主要是用来说明词语或词语组合的含义和用法，并说明给定的短语在上下文中的语用意义。

词语索引行可以作为语言研究的信息资源，可以帮助我们发现模式与词汇意义和语用意义之间的联系，由于使用了词语索引行，我们的研究就会比仅仅依靠主观的“内省”（intuition）更加可靠，而且，从具体的词语索引行提供的语言数据中，我们还可以抽象出一般性的范畴，对语言现象进行综合的解释。

本章还说明，从词语索引行中我们可以观察到什么？怎样使用词语索引来处理更多的语料，以便进一步研究词语的用法？怎样使用更广泛的语境来观察隐含的意义？怎样使用试探的方法？

最后，本章还说明，为什么在语料库研究中我们更倾向于使用“基于单词”（word-based）的方法。在第四章中，我们将把这种“基于单词”的方法与“基于范畴”（category-based）的方法进行对比，阐明它们各自的优点和缺点。

第四章　语料库语言学研究方法：超出词语索引行的研究

本章讲述语料库语言学中除了词语索引行之外的其他研究方法。

词语索引行是研究语料库的有用的工具，但是，由于观察语言现象的人处理信息的能力是有局限性的，所以，词语索引行这样的工具也免不了带有局限性。在本章中，作者还介绍超出词语索引行之外的研究方法，例如，语料库中词语搭配的统计方法和语料库的标注方法。

本章还讨论了在语料库分析中基于单词的方法和基于范畴的方法。这两种方法的区分不但有方法论上的意义，而且还有理论上的意义。

如果我们统计语料库中的某个“词型”（type）的所有“词例”（token）出现的次数，就可以计算出这个词型的频度。频度表有助于揭示不同文章的差异。例如，经济学文章中的高频度单词 price，cost，demand，curve 等在一般性文章中的频度都不高。在英语中，the 和 of 的出现频度很高，但是，它们都难以反映文章的特点，因此，我们还要统计出那些能反映文章特点的高频度单词，这样的单词叫做“关键词”（keywords）。

单词之间的搭配（collocation）也就是单词之间共同出现的倾向，这样的倾向也可以使用语料库来发现。词的搭配可使用“互信息分值”（Mutual Information score，简称 MI-score）和“t 分值”（t-score）来计算。彼此有搭配关系的两个单词之间的跨度（span）有时会很长。例如，在口语中，I wonder 往往与它后面的单词 because 搭配以说明原因，这时候，在 I wonder 与 because 之间可能出现若干个单词，形成很大的跨度。通过词的搭配统计得到的信息经常比观察词语索引行得到的信息更加可靠，但是，我们在解释词的搭配信息时要慎重，免得误入循环论证的迷津。

“标注”（tagging）的目的就是给语料库中的每一个单词附上一个适当的词类标记。这样的标注一般由计算机自动地进行，由于语言中存在大量的兼类词，词类的自动标注并不容易。在自动进行词类标注时，可以使用基于规则的方法，也可以使用基于统计的方法。目前语料库自动词类标注的正确率大约为 90%左右，所以，在自动词类标注之后，还需要进行手工校对。

“剖析”（parsing）就是对于语料库中的句子进行句法分析，这样的工作也可以自动地进行。目前计算机自动剖析的正确率还不高。此外，还可以给语料库进行“语义标注”（semantic annotation），说明单词或词组的语义性质。

第五章　语料库在应用语言学中的应用

本章介绍语料库在除了语言教学这个领域之外的应用语言学其他领域中的应用。这些领域是：词典和语法参考书的编写，意识形态和文化研究，翻译研究，文体论研究，法律语言研究和写作。

语料库使词典和语法参考书的编写发生了革命性的变化。由于使用了语料库，词典和语法参考书的编写更加重视单词频度的作用，更加强调单词的搭配关系和词语用法，更加注意语言变异，更加关注词汇在语法中的作用，更加重视语料的真实性。

近年来，应用语言学日益关注语言与意识形态之间的关系的研究，出现了“批评语言学”（critical linguistics）和“批判性话语分析”（critical discourse analysis）这样的新学科。这些新学科试图把文本放在产生它们的社会背景下进行研究，试图揭示在文本的表面命题后面隐藏着的意识形态，试图采用不同的途径来解释某些社会事实，使其具有新意，从而对人所共知的常识提出挑战。语料库技术可以为这样的研究提供有效的帮助。

翻译是语料库应用的一个重要领域。语料库可以为“机器翻译”（machine translation）提供真实的语言事实，从而帮助改进机器翻译系统，双语语料库的对应文本还有助于帮助我们进一步认识翻译过程的实质。此外，语料库还可以作为翻译人员的实用的翻译参考工具，帮助他们提高翻译的质量。

语料库的词频分析技术、词语索引技术以及搭配分析技术有助于分析作家的文体风格。近来出现了“文体计量学”（stylometrics）这样的学科，使用统计方法来研究文学作品的文体风格，从而发现文学作品的问题风格在若干个世纪的历史过程中的发展线索。

“法律语言学”（forensic linguistics）是应用语言学的一个分支学科。法律语言学可以鉴定磁带录音的语音，确定犯罪嫌疑人在被捕时是否理解了向他们提出的问题，判断两个不同的文件是否为同一个人所写，判断一个文件是一个人写的还是两个不同的人写的。语料库技术可以用来比较不同的法律文件，或者用来比较文件中的不同部分以便证实文件是什么人写的，或者文件的内容是说什么的，或者用来分析文件中语言的性质以便区分其中哪些是真实的，哪些是不真实的。

语料库还可以为写作者提供帮助。联机的电子词典可以帮助写作者校正拼写错误，查找不熟悉的科技术语和陌生的单词。写作者也可以直接使用语料库来替代联机电子词典，从语料库中发现更加合适的表达方式，区分词语的典型用法和非典型用法，从而提高写作的质量。

第六章　语料库与语言教学：语言描写问题

语料库的发展对于语言教师的职业生涯发生了两方面的影响。首先，语言教师所教的内容发生了根本性的变化，由于语料库中包含了词语用法的丰富的信息，语言教师所教的语言实际上就是词语的用法，因此，所谓教语言就是教词语的用法。其次，语料库本身就可以作为语言教学的材料，语料库成为了语言教学大纲研制和语言教学方法论研究的基础。

从语料库的观点看来，语言就是词语的用法，词语的用法是语言描写的核心内容；语言中的意义可以使用模式来描述，因此，意义和模式之间并没有原则性的区别；语言的语法存在于词汇当中，因此，词汇和语法也没有原则性的区别。语料库语言学认为，语言有两条组织原则，一条是“惯用原则”（idiom principle），一条是“开放选择原则”（open-choice principle）。在语言教学中使用语料库，通过模式来分析意义，通过词汇来解释语法，一定可以提高语言

教学的效果。

“语言变异”（language variation）是语言在不同的环境下产生的差异。语料库为研究语言变异提供了有力的手段。在语言的地域差异、性别差异、社会群体差异、时间差异、语域（register）差异的研究方面，由于使用语料库来观察单词的频度、特征的频度、单词的意义和用法等变异参数，都取得了可喜的成绩。

第七章　语料库与语言教学：一般应用

语料库是真实的语言数据，因此，可以使用语料库进行“数据驱动的学习”（dada-driven-learning，DDL）。学生从语料库所提供的数据中，往往会发现被教师所忽视的一些语言现象，学生们也可能发现一些教科书中没有提到的语言现象，从而发挥他们的学习主动性，使他们通过语料库学习到更多的语言知识，思考更多的语言学问题。在 DDL 中，学生可以使用“生语料库”（raw corpus）进行学习，从生语料库中发现语言规律，根据生语料库中的数据来编写学习材料。

所谓“交互式学习”（reciprocal learning）就是指操不同语言的学习者结成伴，彼此学习对方的语言。例如，操法语的学生与操英语的学生结成伴，操法语的学生学习英语，操英语的学生学习法语。在交互式学习中，双语言“对应词语索引”（parallel concordance）可以帮助语言学习者观察不同语言之间的对应规律。

DDL 并不直接地向语言学习者讲述语言的特征，而是让语言学习者自己去观察语言现象，自己做出假设，最后自己得出结论。这样的学习方法显然有助于激发语言学习者的主动性。

语料库还有助于“词汇教学大纲”（lexical syllabus）的编制。所谓“词汇教学大纲”并不是仅仅教词汇的教学大纲，而是以词汇教学为主线的教学大纲，通过词汇来学习词语的用法，从而掌握语言。

有人提出，不应当过分地估计语料库在语言教学中的作用。语料库只是在非常有限意义上的“真实语言”，语料库只是文本的线索和踪迹，但并不是“话语”（discourse）本身。有一种观点认为，语言教学中也需要通过语法规则和词汇表来学习，这样学习起来更加方便，因此，我们不能只依靠语料库来进行学习。这样的观点是对于语言教学中使用语料库的一种挑战，也是我们应当面对的一些急待解决的问题。在本章中，作者陈述了她对于这些问题的看法。

第八章　语料库与语言教学：专门应用

本章讨论语料库在学术英语写作以及在语言测试等专门领域中的作用，进一步说明语料库在语言教学中的重要性。

“学术英语”（English for academic purposes，EAP）写作是指使用英语来写学术论文或科技文章。语料库可以用来帮助 EAP。使用语料库可以发现特定的学术英语中常用的各种语言框架（frames），这种框架随着学科领域的不同而各有特点。

语料库还可以作为语言测试的资料来源，帮助语言测试者进行语言调查，帮助语言测试者设计测试的题目。

使用语言学习者语料库可以比较母语学习者与非母语学习者的差异，从而改进语言教学，提高语言教学的效果。

第九章　一名应用语言学家看语料库

作为一名应用语言学家，Susan 最后试图简短地回答这样的问题：语料库究竟使应用语言学发生了什么样的变化？

Susan 认为，这样的变化表现在以下 4 个方面：

1. 语料库使许多过去不可能进行的语言调查变得可能了。
2. 语料库改变了我们观察语言的方式，至少对于语言教师来说，语料库改变了我们观察自己的作用的方式。例如，DDL 和交互式学习使得我们必须重新评价教师在语言教学中的作用，除了语言教师在语言描写中所表述的内容之外，在语料库中还蕴藏着更加丰富的内容等待着我们去发现。例如，在电子词典中，除了给出通常的词条定义和例子之外，还给出了词语索引行供我们进一步去观察和思考。
3. 语料库使我们的生活变得更加简单。通过语料库，我们可以很容易发现语言事实，翻译者可以从语料库中很快地找到得体的翻译等价物，语言教师可以从语料库中找出更加充分的例证来帮助学生纠正他们在语言学习中所犯的各种错误。
4. 语料库也使我们的生活变得更加复杂。语料库更加细致地揭示了语言的本来面貌，使我们认识到，很多一般性的语言规则都是要在一定的上下文中才可以适用，很多我们原来认为天经地义的语言规则实际上都是有漏洞的。

小　结

Susan Hunston 是一位非常务实的语言学家，她的这本关于语料库实践的论著条理清楚，例证丰富，语言流畅，深入浅出，是当前语料库语言学研究中的不可多得的优秀著作。细心地阅读这本著作，认真地思考其中的各种问题，不仅研究国外语言或汉语的人士能够从中受益，学习外国语言或汉语的读者也可以从中得到启发。

［参考文献］

Sinclair, J., S. Jones and R. Daley. 2004. English Collocation Studies. The OSTI Report. London: Continuum

Halliday, M.A.K., Wolfgang T., Colin Y., and Anna C. 2004. Lexicology and Corpus Linguistics. London: Continuum

Chomsky, N. 1969. Quine's Empirical Assumptions, In Davidson, D. and J. Hintikka, eds., Words and Objections, Dordrecht: Reidel.

罗素著，马元德译，1976，西方哲学史，下卷，商务印书馆

注：本文中未提供的文献出处均来自Corpora in Applied Linguistics一书中的参考文献，读者可从原书中找到。

国内相关重要参考文献

（以姓氏拼音为序）

陈建生，1997，关于语料库语言学，国外语言学，1：1~11

陈建生，1998，语篇的自动词性附码，当代语言学，1：25~27

丁树德，2001，浅谈西方翻译语料库研究，外国语，5：61~66

冯志伟，1996，自然语言的计算机处理，上海外语教育出版社

冯志伟，1998，标准通用置标语言 SGML 及其在自然语言处理中的应用，当代语言学，1998，4：1~11

冯志伟，1999，语料库语言学与机器翻译，信息网络时代与日本研究，山东大学出版社。

冯志伟，中国语料库研究的历史与现状，载 Journal of Chinese Language and Computing 11：2，127~136 ，2002，Singapore，*http：//cslp.comp.nus.edu.sg/cgi-win/journal/paper.exe*

冯志伟，2004，机器翻译研究，中国对外翻译出版公司

何安平，1999，语料库研究的层面和方法评述，外国语，2：10~17

何安平，2004，语料库在外语教育中的应用：理论与实践，广州，广东高等教育出版社

胡明扬，1992，英语用法调查语料库及其他语料库，国外语言学，4：37~42

黄昌宁，李涓子，2002，语料库语言学，北京，商务印书馆

李文中，1999，语料库，学习者语料库与外语教学，外语界，1：51~55

孙茂松，1999，谈谈汉语分词语料库的一致性问题，语言文字应用，2：89

王建新，1996，介绍当代三个英语语料库，外语教学与研究，3：37

王建新，1998，索引软件：语料库语言学的有力工具，当代语言学，1：37~42

王建新，1998，语料库语言学发展史上的几个重要阶段，外语教学与研究，4：52~58

王建新，2005，计算机语料库的建设与应用，清华大学出版社

王克非等，2004，双语对应语料库的研制与应用，外语教学与研究出版社

谢应光，1996，语料库语言学与外语教学，外语教学与研究，3：29

徐一平，2000，关于《中日对译语料库》的研制与应用研究，见：刘利民等主编，语言，首都师范大学出版社，258~262

杨惠中，2002，语料库语言学导论，上海外语教育出版社

周强，张伟，俞士汶，1997，汉语树库的构建，中文信息学报，4：42~51