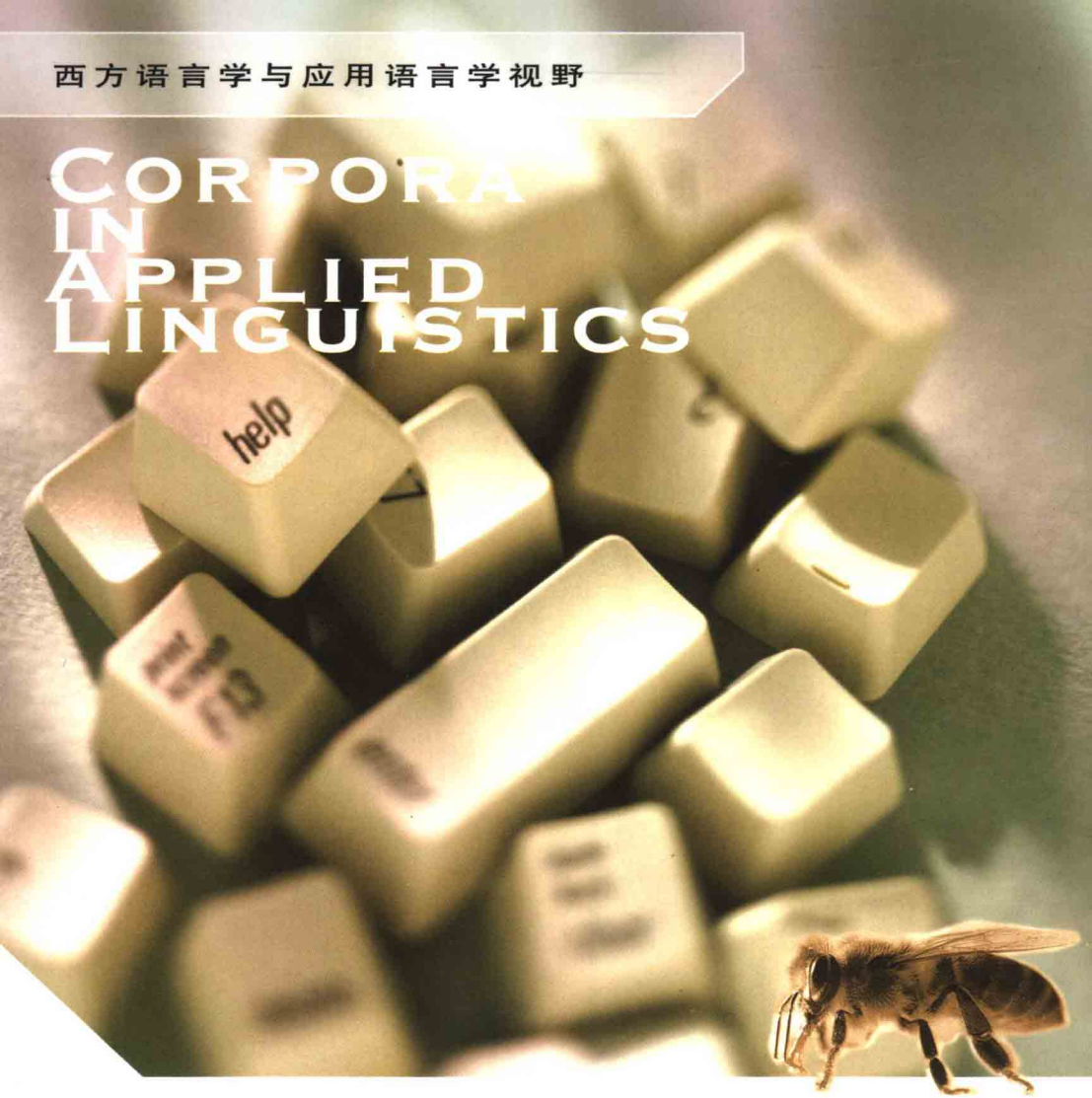
西方语言学与应用语言学视野

CORPORA IN APPLIED LINGUISTICS



应用语言学中的语料库

Corpora in Applied Linguistics

Susan Hunston / 著

世界图书出版公司
剑桥大学出版社

西方语言学与应用语言学视野

应用语言学视野·应用语言学专题

Corpora in Applied Linguistics

应用语言学中的语料库

Susan Hunston 著

冯志伟 导读

世界图书出版公司

剑桥大学出版社

图书在版编目（CIP）数据

应用语言学中的语料库 = Corpora in Applied Linguistics /（英）霍斯顿（Hunston, S.）著. —北京：世界图书出版公司北京公司，2006.8
（西方语言学与应用语言学视野. 应用语言学专题）
ISBN 7-5062-8210-0

I. 应... II. 霍... III. 词语—研究—英文
IV. H03

中国版本图书馆 CIP 数据核字（2006）第 043285 号

应用语言学中的语料库

Corpora in Applied Linguistics

作　　者：Susan Hunston
导　　读：冯志伟
责任编辑：江　枝
装帧设计：然则设计公司

出版发行：世界图书出版公司北京公司　　http://www.wpcbj.com.cn
地　　址：北京市朝内大街 137 号（邮编 100010，电话 010-64077922）
销　　售：各地新华书店及外文书店
印　　刷：北京世图印刷厂
开　　本：711 × 1245　1/24
印　　张：12
字　　数：200 千
版　　次：2006 年 8 月第 1 版　2006 年 8 月第 1 次印刷
书　　号：ISBN 7-5062-8210-0/H · 889
版权登记：京权图字　01-2006-0748
定　　价：23.00 元

西方应用语言学视野

总　序

世界图书出版公司北京公司正在策划引进一系列的国外语言学学术专著，其中包括“西方应用语言学视野”丛书。他们延请了各路造诣很深的专家教授组成一个阵容强大的“智囊团”，从遴选书目到撰写导读，都为这套书献计献策。他们想让我为这个系列写几句话，于是我就认真地翻阅了即将付梓的首批专著，印象不错，就写了这个所谓的序。

在此之前，世图北京公司已经引进了国外学术刊物，其中包括语言学期刊。翻阅这些专著和刊物时，我想到了两点。第一，在信息时代的今天，信息和知识的可获性，对国民经济和民族素质几乎起着决定性的作用。我国学者到欧美留学时，往往感受很深的一点就是：国外大学不仅图书资料几十倍或几百倍于国内大学图书馆的馆藏，学术刊物的种类之多和旧刊之全也是国内大学不能相比的。我看，大量引进国外学术专著和学术刊物有利于为我国的学者提供完备和最新的信息资源，有利于为国内培养出一流的研究者，同时，也有助于为我国大学图书馆逐渐接近世界一流大学图书馆创造条件，有利于营建品味高雅的文化氛围。第二，投身学术之路，一般都是先读教科书，再系统地读专著，然后定期地读学术期刊。专著和学术期刊是做科研和写论文须臾不可离的。专著使我们了解学科的发展历程，系统理解学科的理论和方法；期刊使我们及时知晓学科的前沿，跟上学科发展的步伐。这套开放性丛书的推出，实则为我国语言学研究人士“拿来”国外的最新理

论和方法，进一步开阔视野，拓宽思路，从而能够建立中国广袤土地上自己的语言学理论。

被经典著作、精品图书和权威期刊所包围，犹如与学术大师亲密接触，除了倍受激励，决心奋起直追以外，似乎还可以净化灵魂，提升人生的内涵与境界。中国图书市场上，许多产品过于商业，过于功利，过于垃圾。铺天盖地的粗制滥造的英语试题就是一例。这不利于人才的培养，不利于人文的教化与性情的陶冶，更不利于可持续性教育。这样的文化风气迟早要改变。

引进国外名著的重要性，我在二十年前就有所察觉，我曾为我们办的研究生班胶印过部分语言学专著。那时还没有版权问题。虽然那只是“小打小闹”，但得到书的那二百多位学生都受益匪浅。接着，我又主编了《西方语言学名著选读》，让学生有机会领略语言学大师们的风采，也收到良好的效果。今天，世界图书出版公司北京公司大批引进语言学专著和学术刊物，必将为我国语言学研究做出不可估量的贡献。

刘润清

北京外国语大学语言研究所所长

《应用语言学中的语料库》导读

冯志伟

编者按：本导读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以及本书的全貌有一个概要的了解和认识。

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，并始与经验主义妥协，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？这是很值得怀疑的。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。其规模为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－2 June 2000），第三次会议于 2002 年在西班牙的 Las Palmas de Gran Canaria 召开（27 May－2 June 2002），第四次会议于 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：语料有人文与社会科学类、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。该课题分为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）。

其中，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