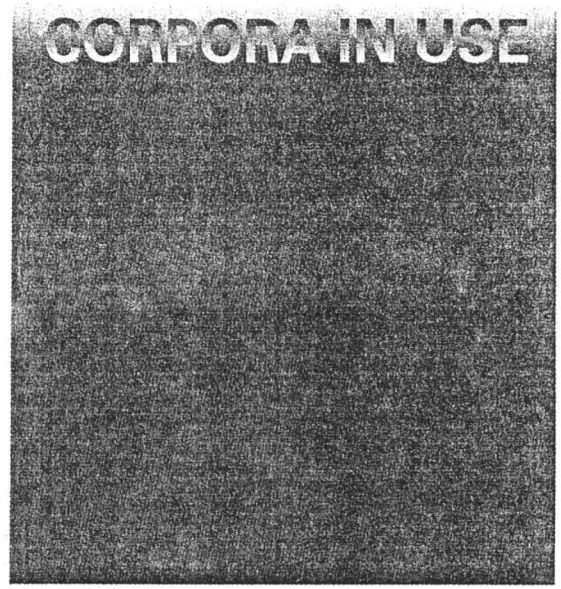
国家社会科学基金项目

卫乃兴 李文中 濮建忠 等 著

语料库应用研究



外教社

上海外语教育出版社

图书在版编目（CIP）数据

语料库应用研究 / 卫乃兴，李文中, 濮建中等著. —上海：
上海外语教育出版社，2005.8
ISBN 7-81095-781-3

Ⅰ. 语… Ⅱ. ①卫…②李… ③濮 … Ⅲ. 计算机应用研究
Ⅳ. H09-39

中国版本图书馆CIP数据核字（2005）第095020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：杨自伍

印　　刷：上海外语教育出版社印刷厂
经　　销：新华书店上海发行所
开　　本：787×1092　1/16　印张 16　字数 352千字
版　　次：2005年 10 月第 1 版　2005 年 10月第 1 次印刷
印　　数：3 100 册

书　　号：ISBN 7-81095-781-3 / H · 309
定　　价：27.00 元

In Honour of Professor Yang Huizhong!

谨以此书献给杨惠中先生！

纪念杨惠中先生从事外语教学与研究 50 年！

上海交通大学语言文字工程研究所

写在《语料库应用研究》付梓之际

在《语料库应用研究》这本小书即将付梓之际，杨惠中教授从事外语教学与研究事业也达50年之久。我们谨将此书献给中国语料库语言学的开拓者之一杨惠中先生，向先生表达由衷的敬意！

20世纪60年代初，美国布朗大学的W. Nelson Francis与Henry Kucera创建了世界上第一个电子语料库——“标准美国英语布朗语料库”，当时语言学研究中的理性主义风生水起，以实证主义为主导的语料库研究仅是狂飙中的一点微弱星火。现代语料库方法刚一出世就受到双重的压力，即在理念上因历史的包袱而不见容于当时的学术主流，在技术方法上又受制于当时的计算机软、硬件技术水平。现代电子语料库问世之前的早期语料库语言学也有不足，它过分强调实际可观察到的语言实例，彻底排斥任何其他语言学依据；另外，它还坚持句子可以穷尽列举等错误主张。所以，20世纪50年代末Chomsky语言学兴起时，早期语料库方法受到严厉批评就毫不奇怪了。60至70年代，计算机技术及其普及程度还处于较低的水平，虽然出现了现代电子语料库，语料库语言学研究却举步维艰。但是，基于语料库的语言学研究在这一时期毕竟没有被淹没，而是在理性主义坚硬的地壳中破土而出，且在80年代显示出勃勃生机。

20世纪80年代初，基于电子语料库的语言学研究在国际上开始兴起，而中国语言学界对这一概念仍感陌生。这时候，上海交通大学杨惠中先生等开始着手建立“交大科技英语语料库”(JDEST)，并于80年代中期建成。该语料库建成之初的容量为100余万词，同时还有配套的语料库索引软件、AGTS自动POS赋码系统以及技术词汇筛选技术，其技术和理念在80年代中后期处于国际领先水平。如今，其自动赋码系统仍是国内唯一自主开发的同类产品，而且语料库容量已经扩展到400余万词，收集的文本体裁更加丰富，实际上已成为一个学术英语语料库。当时，Geoffrey Leech盛赞JDEST为专业英语语料库的“先锋”之作。该语料库既是中国第一个真正意义上的语料库，也是国际上获得公认的第一代电子语料库。正因为建成了这个语料库，中国的外语教学才有了第一个基于语料库的大纲词表，并有了第一批专攻语料库语言学方向的博士生、硕士生，中国也才有了第一届语料库语言学国际研讨会。

20世纪90年代末，杨惠中先生与广州外语外贸大学桂诗春先生联合组织团队，建成了国内第一个“中国学习者英语语料库”(CLEC)。如今这个学习者语料库已成为广大外语研究者以及博士生、硕士生乐为己用的研究平台和资源，影响波及欧美和东南亚。

21世纪伊始，杨惠中先生又主持开发了国内首个“大学英语学习者口语英语语料库”(COLSEC)，目前该库已建成并投入使用。这个项目横跨国内三所高校，历时三年，吸引参与的博士生、硕士生达几十人，初期成果头角才露，学科队伍日益壮大，后继有人。

综观中国语料库语言学20多年的发展，杨惠中先生作为一个开拓者和播种者，其启动之力、创建之劳、引领之范，当在以后学术研究中愈见其大、惠及莘莘。然而，杨惠中先生的语料库语言学研究，其价值不仅在于实践上开创新独造之力行，还在于其学术上见微知著之卓识。建立具有独立知识产权的语料库，并以此为基础开展中国环境中外语学习和教学研究，使得中国的语料库研究从一开始就摆脱了对西方学术的依赖和亦步亦趋的追随，而是立足本土实际，在不断出新的成果和贡献中日益获得自己在国际相同领域独立的学术身份。其实，这也是近年来该领域吸引力日益增长、越来越多的学者和年轻学子积极投身该领域的心理动机。

作为一个学者，杨惠中先生的学术成就还不仅限于语料库研究方面。也是在20世纪80年代，他与同事一道，创建了中国第一个标准化考试，即大学英语四、六级考试，以他为首的研究人员设计开发的一整套标准化语言测试理论与实践已在国内外产生了深远而巨大的影响，在考生数据自动处理、分数等值处理、作文分调整、试题分析等技术方面都处于全国乃至世界领先地位。如今，这个考试以其科学性和影响力已成为中国的品牌，与多个国家和地区的考试机构建立了广泛的合作关系；他对计算语言学和机器翻译的深刻理解和造诣使他成为多项国际项目的合作者。

如今，杨惠中先生仍是一位默默的力行者。他不事浮躁张扬，不以哗众而买名，只是坚守自己的学术，一步一个脚印。

杨惠中先生也是一位忠厚的长者，其为师为人严谨认真，却不失幽默亲和。正所谓"望之俨然，即之也温"。然学高生威，德高聚望。吾侪生也有幸，能亲炙师道，得其耳提面命，叩鸣之际，多有朝夕之既。近年在语料库领域孜孜自苦，不敢稍有懈怠，欲以此书作为又一次作业，然而我辈深知，限于自身学养和能力，恐心有余而力有未逮。若有批评建议，一并欢迎。

谨以此书向杨惠中先生表达真挚的敬意！

著者

2005年2月

序言 PREFACE

历经四年的孜孜努力，国家社会科学基金项目"语料库技术与网络多媒体在外语教学中的开发应用"(02BYY016)终于结项了。读罢几位鉴定专家不约而同的赞誉勉励之词，我们生出了些许欣慰，然而又不禁诚惶诚恐①。我们真诚地将这本小书《语料库应用研究》奉献给全国的同仁，希求能对语料库语言学的发展略尽绵薄之力。

作为一门相对年轻的学科，语料库语言学自上个世纪80年代以来取得了迅猛发展。在发轫之时，语料库研究被人视为少数狂热者的轻率之举；而今，它已经成为欧洲范围内语言研究的主流，其影响触及到语言学领域的方方面面。就外语教学而言，这种影响无疑是巨大而又深刻的：传统语言教学的理念、内容和方法无一不在发生着变化。面对急速变化的语言研究与语言教学现实，如何系统地研究语料库技术在外语教学中的应用，推进我国外语教学的变革？另外，面对互联网资源的急速扩张与网络技术的日新月异，如何迎接挑战，抓住机遇，使网络资源纳入外语教学的理论与实践？再之，在过去多年里，杨惠中先生带领的语料库研究队伍已经建好了JDEST和CLEC两个语料库，并通过交换，获得了国外一些语料库；如果将这些资源转换为互联网在线索引，实现广大外语教学与研究人员对资源的共享，岂不可以最大限度地发挥语料库的作用，产生极大的学术价值和社会效益？正是基于这些背景和动机，我们启动了国家社会科学基金项目"语料库技术与网络多媒体在外语教学中的开发应用"的建设。我们期望，该项目能在一定程度上缓解我国外语研究人员长期来极度缺乏真实数据的困难，普及语料库研究的方法，促进基于数据的语言学研究，有效地推进我国的语料库应用研究！我们期望，该项目能给我国的外语教师和学生提供大批量的珍贵真实数据，使教学更加注重"真实英语"输入，在较大程度上提高输入的质量。我们还期望，通过网络在线查询、实时检索和基于语料库的数据驱动学习(DDL)，我国的外语教学能更多地注重学生个人的自我探索、自我发现、自我学习能力培养，并尽可能影响教学思想和模式的转变。这些都是我们的殷切希求。

几年过去了，项目组全体同事的辛勤劳动终于初见成效：我们基本完成了项目的计划建设内容。我们开发了CAST(Corpus Analysis and Statistics Tools)检索软件；虽然它仍有缺点，需要进一步完善，但是它使我们有了一个拥有独立版权的英语语料库分析软件，供国内语言学界同仁以及外语教师使用。凭借CAST软件的支撑，我们成功地将目

① 从国家社会科学基金项目办公室反馈的鉴定意见来看，5位鉴定专家均对该项目的创新程度、研究方法、学术价值、理论价值和应用价值给予了高度评价；同时，又对我们今后的研究提出了厚望。对专家们的肯定与赞誉，我们更多的是感受到了鼓励与压力。我等才疏学浅，绝不敢苟且懈怠；必将矢志不渝，研精覃思，不辜学界之期。

前拥有的4个语料库的700多万词容的文本资源转换成了互联网在线KWIC(key word in context)索引；而且，我们将4个语料库的所有关键词的搭配词数据变为在线资源：针对每个关键词，我们提供100个统计值最大的搭配词；外语教学研究人员可以直接登陆http://corpus.sjtu.edu.cn网站检索并下载搭配词数据，用于自己的研究。项目的另一部分建设内容则是"数据驱动学习系统"及其相关模块：我们共建设了"在线数据驱动学习模块"、"ESP写作在线模块"、"在线词典模块"、"在线资源模块"、"交互模块"等子系统，供外语学习者对数据观察、分类、概括，从而自我探索语言、自我发现语言规则，进行自我学习。

该项目的建设得到了上海交通大学外语学院领导的关切和支持；上海交通大学、河南师范大学和解放军外语学院的部分博士生、硕士生积极参与项目建设并付出了艰辛的劳动；上海交通大学的陆元雯博士对ESP写作模块的建设，做出了有益的工作。我们对他们的支持和奉献，由衷地感激！

在项目建设的基础上，我们写成了这本小书《语料库应用研究》。全书分为三部分。在第一部分里，我们探讨了语料库与网络资源应用于外语教学的理论与实践问题。毋庸置疑，关于语料库与网络资源的应用问题，学界争执较多，可谓仁者见仁，智者见智。我们在书中讨论了这些新资源、新技术的利用所涉及到的理念问题，以及由此引起的外语教学的过程、手段和环节的变化等等。这些都是一家之言，旨在抛砖引玉，与广大同仁切磋。在本书的第二部分里，我们着重讨论了语料库资源与成果应用的相关技术问题，包括检索软件的开发、商用软件的使用与功能开发、对比中间语分析中的数据统计等等。需要说明的是，目前的语料库检索软件种类繁多，功能特点各有千秋。我们只能以自己开发的CAST和功能较为齐全的商用软件POWERGREP为例，阐述软件开发与使用的技术问题；希望能收到一隅三反之效，有助于读者克服技术恐惧症，较好地使用现有软件，并逐步自行开发一些软件。语料库资源与技术在外语教学中的应用，不可避免地要涉及学习者中间语研究。基于学习者语料库数据，参照本族语语料库，对比、概括和描述学习者语言的特征和规律，发现中间语存在的问题，从而对症下药地采取相应的教学措施，是语料库语言学与外语教学研究结合的有效途径。所以，在本书第三部分里，我们集中报道了一些个案研究。这些个案研究涉及学习者语言的多个层面：词语搭配的、句法的、语篇的、学习者策略的、学习者心理词语网络的，等等；这一部分研究的学习者语言既有书面语交际数据，又有口语语域的语用材料。我们并不试图勾画学习者语言的全貌，那几乎是不可能的；我们只是想通过各种不同层面和语域的个案研究，向读者揭示一些可能的研究方法，以期能有借鉴之用。

除了第十一章由解放军外国语学院陈立平教授撰写外，本书的所有章节均由我们三人，卫乃兴、李文中和濮建忠，执笔撰写，或者根据我们的指导，由我们的研究生执笔撰写；全书内容由卫乃兴最终审定。国际著名的语料库语言学家、ICAME(International Computer Archive of Modern English)协会的现任主席Antoinette Renouf教授(University of Central England, UK)专门为本书写了"导语"，阐述了有关语料库语言学的种种问题；希望她的阐述能解决国内读者可能的关切和疑虑。王晓婷、陈凌明、吕艳华

等几位研究生同学帮助校对了文稿清样。上海外语教育出版社的有关领导为本书的出版提供了宝贵支持，责任编辑杨自伍先生花费了大量时间和精力，提出了许多真知灼见。

我们衷心感谢国家社会科学基金项目鉴定专家的厚爱与鼓励！

我们真诚感谢所有同仁、朋友和学生的参与、建议和支持！

我们诚挚希望读者对书中尚存的错误不吝指正！

著　者

2005年8月

于上海交通大学

CONTENTS 目录

表格一览表

图形一览表

第一部分 理论与方法讨论

第一章 语料库资源及技术与外语教学

卫乃兴

语料库语言学的持续、深度发展产生了强大的数据资源。语料库建设、语料库数据检索与处理又形成了一套独具特色的技术手段和研究方法。在过去近 30 年里，该学科的诸多研究成果促人深省和反思，触动了语言研究的方方面面。几乎所有相关的语言研究领域都感受到了语料库语言学的影响；在不同程度上，研究人员都在借鉴语料库的数据和方法。在外语教学与研究领域，无论是上个世纪 90 年代初兴起的数据驱动学习(Data-driven learning)方法，还是世纪之交建立的大批学习者语料库，及至后来异军突起的“中间语对比分析”(Contrastive Interlanguage Analysis)，都反映出人们试图借鉴语料库方法，应用其资源和技术于外语教学的努力。在我国，从杨惠中教授等人建立 JDEST 语料库至今，一批研究者一直坚持不懈地将语料库技术与外语教学结合。然而，二者结合的进展并不尽如人意；总的来说，进展较为缓慢。相关资源与技术在外语教学中的应用，仍存在不少理论与实践的问题，涉及学术立场、理念与认识、技术障碍等等。在本书的开篇之章，我们从多个层面和视角讨论语料库资源、技术和方法应用于外语教学的问题。我们将讨论语料库资源应用于外语教学的合理性与价值，并澄清学术界某些不尽正确的观点。我们还会探讨解决相关问题的途径、资源利用的可能性与可行性等等。

1 语料库资源应用的相关理论问题

1.1 语料库语言学与应用语言学

语料库语言学的学科属性揭示了它与应用语言学的紧密关系。就其方法特征而言，语料库语言学不妨称之为“基于语料库的语言学”(Corpus-based linguistics)或“语料库驱动的语言学”(Corpus-driven linguistics)①。它业已形成的一套理念、方法及其研究成果确立了自己作为一门独立语言学科的地位。就学科属性而言，它大体上属于哲学上的经验主义一派，认同于功能语言学；它的思想与方法直接渊源于英国的 Firth 语言学，具有极强的应用性、实践性，与语言教学的理论与实践密切相关。Michael Stubbs(1993：2)在《文本与技术》一书中曾全面论述了自 Firth 以来，英国语言学的“社会科学属性”、“应用性”、“实证性”、“文本整体性”、“意义形式统一性”等 9 条重要原则。Stubbs 阐述的这些原则实质上也是语料库语言学的原则，揭示了其主要特征和属性。首先，语料库语言学是一门社会科学。它视人类语言使用为一种社会行为方式，关注的是语言的社会作用和功能。其次，它又是一门应用科学。Stubbs 认为，它主要应用于语言教育，与教育语言学密

切相关。这两个属性又决定了语料库语言学的研究对象为外化语言，而不是内化语言。Noam Chomsky一派研究的是内化语言，使用的是内省数据；内省数据是不可观察和验证的，靠本族语者的直觉产出。语料库语言学则使用可观察的和可记录的真实语言使用数据。内化语言与客观世界的真实语言使用相去甚远，研究者呈现的是一套高度理想化的抽象系统；系统越理想、越抽象，与客观世界的真实语言使用可能相距越远。而语料库语言学研究的就是客观世界的语言使用本身，对其形式、意义和功能概括与归纳；其研究成果与语言教学具有极强的相关性。语料库语言学采用的一套量化研究方法和计算机技术手段又可快速、准确、高效地处理大批量数据，其研究深度和广度非别的方法可以比拟，从而人类语言使用的多维度、深层次问题得以更为有效的探讨，其对语言教学的启示也越来越深刻。

然而，这并不是说语料库语言学可以代替应用语言学。应用语言学是关于语言教学的一门综合学科，涉及语言学、社会语言学、心理学、教育学等多学科理论和原则；语料库语言学的研究内容则在某些方面与应用语言学重合。在这个问题上，以 Henry Widdowson 为代表的一些学者(Widdowson，2000)认为，语料库语言学属于描述语言学，而外语教学属于应用语言学；将语料库研究成果直接应用于外语教学是“语言学应用”，甚至是“语言学误用”，而不是“应用语言学”②。Widdowson 长于对概念做严格的两分区分，“语言学应用”与“应用语言学”便是其多年坚持的一个重要区分(Widdowson，1980)。这样的区分颇有益处。任何语言学科，包括结构主义语言学、形式主义语言学、功能主义语言学等等，无论成果再显著，都只是见长于一部分语言事实的研究，而疏于另一部分事实；因此都有其局限性，不宜直接使用于教学。从这个意义上说，应用语言学家的重要任务之一就是综合各种流派的研究成果，折中或调和尖锐对立的观点，探讨并规划语言教学的原则与方法。

然而，应用语言学家论述的原则与方法可以直接使用于语言教学吗？答案是否定的。应用语言学虽是一门直接关于语言教学的科学，但仍有其相当程度的抽象性和概括性；否则便不成为科学。那些抽象的原则和理想化了的教学方法，如不加以改造、折中或调和就直接应用于外语教学，只能是事倍功半，甚至事与愿违。以教学法为例，多年来，教学理论家们先后倡导了“功能—意念法”、“情境教学法”、“交际语言教学法”、“认知学习法”、“建构主义教学法”，林林总总。但是，这些教学法只是代表了一种教学理念或一套方法论；每套方法论都侧重于语言教学的一个方面，而忽视其他方面。而语言教学是一个人所共知的复杂系统和过程，具体教学法的成功实施有赖于教师根据具体的教学对象、教学目的、教学内容和客观条件进行量体裁衣的调整、折中和取舍。部分教师削足适履地生搬理论家的教学法，结果只能适得其反。不少外语教师抱怨：这些教学法“好听”、“好看”，但“不好用”。它们“好听”、“好看”，因为它们是理论的和理想化的；它们“不好用”，因为它们是抽象的，与教学实践仍有距离，需要教师的调整。事实上，就连应用语言学家自己，也在不断折中、调和甚至修正自己的理论。Widdowson(1978)极力主张“交际语言教学”(Teaching Language as Communication)，而 Widdowson(1979)却又为此懊悔，提出“为交际而教语言”(Teaching Language for Communication)；Widdowson(1978)完全背书“交际能力”这一概念，而 Widdowson(1983)又认为仅此远远不足，应加上“交际能量”(communicative capacity)这一概念；不一而足。试想，如果这些主张和原则都直接应用于教学实践，又怎能不引起认知和过程上的混乱？道理很明显，任何语言学科的研究成果都不可能直接应用于教学，应用语言学也概莫能外。其原则、方法和主张也需要紧密结合教学实际加以折衷、调和或改造；这些折衷或调和工作只能落在语言教师肩上。

主张将语料库研究成果应用于外语教学，并不排斥对其他学科成果的兼容并蓄和调和折衷。语料库语言学的学科属性，语料库研究者对教学问题的强烈关注，不可能在运用研究成果时不紧密结合教学实际，根据诸多因素而抉择。语料库研究者从来没有提出过“直接应用”之类的主张，也没有试图用本学科的成果去否定应用语言学存在的价值；相反，不少语料库研究者乐于视自己的研究为应用语言学的一部分。在很大程度上，Widdowson(2000)是在针对自己虚拟出的一些立场进行指责与批评。但是，无论“语言学应用”还是“应用语言学”，首先是乐于“应用”的态度；乐于“应用”，语料库研究成果就可能极大地丰富应用语言学的理论与实践。过于僵硬地区分两个概念，则会排斥或拒用语料库研究成果；这并不是 Widdowson 的初衷与立场。

1.2 语料库的自然数据与外语教学的真实输入

语料库是由自然文本构成的资料库。“自然数据”或“验证数据”是普通语言学领域的常用概念。而“真实性”则是应用语言学领域的惯用说法。外语教学的中心任务是培养学生的交际能力，这是不争的共识。而提供真实的语言输入应当是培养能力的有效途径。在这方面，语料库的自然文本构成了可供外语教学使用的重要资源。John Sinclair(1984，再版于 1996：93--101)提出了语言数据“自然性”的概念，认为“自然性”等同于“真实性”；他主张，外语教学“只能提供真实例子”。

然而，针对“真实性”，应用语言学界历来观点不同。上面谈到的观点可称之为“文本真实性”(text authenticity)。另一种观点是所谓的“学习者真实性”(learner authenticity)，以 Widdowson 为代表。这一派认为(Widdowson，1978；1979)，真实性不应当指文本的绝对质量，而应当指学生的反应；只有当学生对输入作出了合适的反应，才算产生了真实性。照此，真实语言使用中的自然数据可能不具真实性，因为它们可能激不起学生的合适反应；而经过特别加工或处理的输入则可能具有真实性，因为它们有可能激起学生的合适反应。显然，Widdowson 意在使用非真实输入。或许，该主张有其合理成分：外语教学须根据学生的具体水平和其他因素，渐进地使用真实输入，使学生语言逐步接近本族语者的语言使用。但是，非真实输入不可能发展学生处理真实语言数据的能力。究竟如何处理由非真实到真实的循序渐进问题？Widdowson 一派并未作出回答或解释。事实上，Widdowson 的观点在应用语言学界并未成为主流。多数应用语言学者认同的是“文本真实性”的概念(见 Wilkins，1976；Johns，1983 等)。

语料库问世前的语言描述和现时的语言教学材料不乏非真实的、生造例句的使用。比如，Randolf Quirk 等人在《综合英语语法》中阐释形容词和副词比较级用法时所用的例句：

Walter played the piano more often in Chicago than his brother conducted.

concerts in the rest of the States. (引自 Francis, 1993: 138)

I've never seen a dog more obviously friendly than your cat. (引自 Francis, 1993: 138)

基于 COBUILD 语料库的研究表明,这是两个生造的句子,与真实的英语使用相差甚远:英语本族语者不会说出那样的句子。生造句子的价值在于证明或阐释语言学家设计的理论模型;在教学过程中也有助于规则知识的学习。但掌握规则知识是一回事,语用能力是另一回事;后者只能通过处理大量真实数据而发展。正确的选择只能是真实输入,包括真实的教学材料、真实的课堂语言等等。

语料库研究成果对外语学习者的技能发展也提出了重要启示。Sinclair(2004)认为,除了一般的读、写、听、说技能外,学习者需要掌握另一套技能:1) 将话语切分为有意义成分的能力;2) 区分向心式结构与离心式结构的能力;3) 元语言能力,即使用语言对语言认识、讨论、重组的能力;4) 释义的技能。这一套复杂技能不可能靠真实语言使用之外的生造数据获得,唯有通过接触大量的真实数据,处理大量的真实文本信息,也就是 Sinclair 说的"在语言内部"学习语言,方可获得。

在我国,外语教学的环境条件对真实语言输入,乃至学生的外语语用能力发展都有极大制约;国内缺乏英语本族语者真实的交际材料,外语教师的语用能力和语言直觉又相对有限。对此,语料库数据无疑是可资利用的有价值资源。语料库资源的共享可在一定程度上克服由于教师语用能力和语言直觉欠缺而产生的障碍,解决真实教学材料缺乏的问题。

1.3 语料库研究者:一身二任

从学术背景与研究目的来看,语料库研究者必然要将成果应用于外语教学。大部分语料库研究者都承担着双重角色,既要探讨、摸索和研究语料库建设方方面面的问题,又要高度关切资源与成果应用于外语教学的实践问题。我们不妨将全球各地从事语料库工作的人员分为三类。第一类是计算语言学家。计算语言学家大都是计算机科学家和数学科学研究者;他们的主要学术关切是语言形式和语言信息的计算机处理、自动文本的生成、机器翻译模型及其技术难题的攻克等等。他们面向自然语言处理建立的语料库及其相关研究,与语言教学并无太大关系。第二类是语言学家,如 John Sinclair, Geoffrey Leech, Jan Svartvik, Stig Johansson 等等。他们的主要关切是语言学研究,尤其是语言描述;由于英国语言学长期来的"社会科学"与"应用科学"属性,他们对语言教学问题具有浓厚兴趣。这部分人建立的语料库,如 LOB、COBUILD、BNC、LSWE 等及其相关研究开创并确立了语料库语言学这一学科,催生了语言研究体系和方法的一系列变革,并对语言教学产生了重要启示和应用价值。第三部分人可称之为语料库研究者,为数众多。他们本来就是应用语言学者或语言教师。多年的外语教学与研究实践促使他们转向语料库领域,寻求证据、方法和新的理论,解决传统方法难以处理的问题。他们建立的语料库具有明确的教学研究与应用目的、极强的针对性、相应的合适规模,如全球各地建立的专门用途英语语料库(JDEST 等)与学习者英语语料库 ICLE(International Corpus of Learner English)、USE(Uppsala Student English)、CLEC

(Chinese Learner English Corpus)、HKUST (Hong Kong University of Science & Technology)语料库等等。

就中国的具体情况而言,语料库语言学与外语教学可以说具有天然的联系。从杨惠中等上个世纪 80 年代中期建立第一个语料库 JDEST 至今,语料库语言学的发展就与外语教学结下了不解之缘。建立 JDEST 的直接目的就是为当时的大学英语教学改革,尤其是为制定新的"大学英语教学大纲"提供决策依据。该库包含的文类、题材、通用词汇、次技术词汇、技术词汇等数据信息,为当时的教学大纲以及后来的"大学英语考试(CET)大纲"提供了坚实的参照数据。由桂诗春、杨惠中两人主持,于 21 世纪初建立的 CLEC,及至后来何安平建立的 CEEC (The Corpora of English Education in China)、文秋芳建立的中国英语专业学生口语语料库等等,都直接服务于外语教学研究和实践。可以看出,中国语料库语言学的开拓者都是应用语言学者,大部分语料库研究者本身就是英语教师。他们的研究关切必然是外语教学、学习者英语等有关问题。这样的双重身份与研究内容已使 Widdowson 坚持的"应用语言学"与"语言学应用"之区分不再有效。

然而,语料库研究者的双重角色和任务更多的还在于调和,既要在某种程度上调和各种学派的不同观点,更要调和研究成果与教学实践的需求。学生语用能力的发展并非简单地将资源与成果用作真实语言输入。从输入到产出涉及复杂的过程活动。Widdowson 外语教学理论的重要立场之一是他对三种知识和能力的阐述(Widdowson, 1983)。他认为,外语教学涉及的三种知识与能力是:1) 语言体系知识(systemic knowledge),即常说的语言能力;2) 图式知识或能力(schemata),即客观世界经验在认知体系中形成的参照框架,包括各种专业背景知识、语篇惯例等等,大致上等同于常说的交际能力;3) 过程能力(procedural capacity),即在体系知识之间、图式知识之间以及体系知识与图式知识之间调和,将知识实现为交际行为的能力。而过程能力的发展则需要一系列的过程活动(procedural activities)。显然,过程活动至关重要,既涉及到教学法层面的教学活动,又涉及到教学内容的层面。应当看出,任何一种新的教学理念、新的教学资源的采取都意味着新的教学环节与过程的产生。语料库研究者的调和工作不妨包括研究成果的利用与呈现、合适资源的选择、语料库软件技术的利用、各种语料档案的建立与利用、相关教学环节和活动的设计等等。这些都将在下面讨论。

2 语料库研究成果与外语教学

语料库驱动的语言学已深入到词语搭配、句法结构、语义和语用、话语分析、社会文化因素等等研究领域。许多研究成果涉及和描述了传统研究方法不可能涉及的语言事实、用法模式和规律,揭示了语言运作的深层次机制。将研究成果运用于外语教学,极为必要和重要。

2.1 频数、概率信息的利用

"概率是语言系统的内在属性"(Halliday, 1991: 31)。语料库统计数据显示的有关形式的频数信息与概率信息是教学设计的科学参照依据。上个世纪 30 年代,Thorndike 和 Lorge 基于非机读语料库,用人工计算的办法编撰了《英语教师词汇手册》(The

Teacher's Word Book of 30,000 Words)。到了 50 年代,Michael West 采用同样方法编制了《英语教学通用词表》(A General Service List of English Words)。这些早期的英语教学研究者深知数据的重要,为教学设计付出了今天的人们难以想象的艰辛劳动。80 年代末,COBUILD 语料库研究表明(Willis, 1990: 46):

1) 频数最高的 700 个英语词覆盖了英语文本 70%的内容;

2) 频数最高的 1 500 个英语词覆盖了英语文本 76%的内容;

3) 频数最高的 2 500 个英语词覆盖了英语文本 80%的内容。

Sinclair 和 Renouf (1988: 148)明确提出"词汇大纲"的概念,认为:任何英语学习者的学习重点都应当是:

1) 语言中最常用的词形;

2) 这些词形的核心用法型式;

3) 由这些词形构成的典型词语组合。

杨惠中教授认为,语料库的三类数据是教学大纲设计的重要基础:1) 语料库总的词频信息,包括词形的频数信息和词目(Lemma)的频数信息;2) 词汇的覆盖率信息,即上面讨论的 COBUILD 语料库的研究发现;3) 词汇的分布信息,即高频出现的词汇分布于不同领域文本的数据信息。三类数据相互参照,方能严谨制定出大纲的词汇学习内容(这些观点系由本章作者与杨教授讨论所得)。但这并不是说,语料库数据是教学大纲制定的唯一依据。语料库数据是量化了的实证研究结果,是基础和出发点;大纲设计者当然还要参照一套定性的标准,二者结合,取得某种折中和平衡。但是,教学大纲设计者必须有一套明确表述的参照数据,明确界定的学习目标等等;这些似乎都仍被现时国内的大纲制定者忽视。

除了应用于大纲制定,语料库的频数、概率信息等数据对于教学词典编撰、教材编写、教学辅助材料开发等工作也有不可忽视的作用。在欧洲各国,没有语料库而去编撰词典已是不可想象之事;依据个人语言经验和知识编撰的词典,其科学性、学术价值与实用价值都有极大的局限;照搬或引借他人词典的词条、义项、例证来编撰,则为知识产权法规不容。国内的教学词典开发,急需转至语料库方法上来。然而,什么是语料库,开发什么样的语料库,需要多大规模的语料库等认识问题以及相关的技术标准问题,似乎仍远未解决③。语料库数据在词典编撰、教材开发等领域的运用,仍有艰巨的工作要做,但前景十分广阔。

2.2 可能性、典型性与教学内容

语料库研究提供的形式频数与概率信息,区分了语言使用中的可能性与典型性。可能的事件不计其数,但只有高概率事件是典型事件。语料库研究常用的概率信息包括 Z 值或 T 值信息、MI 信息(互信息)、卡方值等。Z 值依据任意两个词形在给定语料库中的出现与共现频数,测量它们组合的几率和典型性;MI 测量任意两个词形在给定语料库中的搭配强度。这两种概率信息均能揭示词汇行为的核心模式,如表 1.1 和表 1.2 分别显示了两个词形在 COBUILD 中的搭配数据:

表 1.1 "issue"的部分搭配词信息

搭 配 词	MI 值	搭 配 词	MI 值
thorniest	8.454911	clouded	5.092006
prejudged	8.092305	debated	5.067172
contentious	6.893772	resolved	5.059870
thorny	6.849325	skirted	5.013996
vexed	6.784052	abortion	4.923036
emotive	6.770245	unilateral	4.794296
fudged	6.738533	substantive	4.746197
ticklish	6.722935	resolve	4.720078
divisive	6.355166	sensitive	4.692036
complicating	6.217649	euthanasia	4.683383
weaning	5.835741	pervasive	4.261848
touchy	5.217550	hostage	4.252859
unresolved	5.159700	raise	4.205205

表 1.2 "argued"的部分搭配词信息

搭 配 词	T 值	搭 配 词	T 值
against	5.861002	attorneys	2.435164
case	5.587856	passionately	2.434353
strongly	3.754228	party	2.329915
lawyers	3.353118	unions	2.286318
government	3.258613	opposition	2.187544
court	3.047679	consistently	2.157700
evidence	2.854282	opponents	2.105291
critics	2.852426	judge	2.041156
officials	2.683231	heatedly	1.997793
lawyer	2.680562	forcefully	1.988193
secretary	2.515034	convincingly	1.984551
defense	2.509449	prosecutors	1.969985
closely	2.495101	speakers	1.910727
persuasively	2.446967		

这些数据提供了两个关键词较大范围内的搭配词实例及其概率信息。这些搭配词的概率信息不同,典型程度也不同,但大都达到了显著水平,都可视为本族语者使用的典型

搭配。可能性与典型性孰轻孰重？信奉转换生成语法的学者会告诉我们，可能性远重于典型性。但可能性几乎是无法确定的：大量事实表明，本族语者的语言直觉有限；一个"理想本族语使用者"认为"可能的"、"可接受的"形式，会被另一个"理想本族语使用者"判定为"不可能"和"不可接受"。作为非本族语者的外语教师相信哪一个呢？再则，让学生习得本族语者常用的、典型的语言形式，理应成为教学的首要任务之一。语料库研究的重点是频数超过 1 的形式，尤其是高频发生的形式。形式的概率信息使得典型性成为有形的、可观察的实体。翔实的典型形式实例是选择、安排和组织教学内容的有效依据。当然，外语教学不应当排斥可能性。但是，可能性与典型性之间有何关系？确立可能性的研究方法是什么？在这些问题解决之前，我们似乎只能依赖典型性；语料库提供的典型形式及其使用价值，不容忽视。

2.3 词汇与语法的共选

语料库研究的重要成果之一是词汇与语法的共选机制。所谓共选，指二者的不可分割关系和统一性：一定的语法结构受词汇选择的制约；一定的词汇形式又受一定的结构选择制约。共选的关系由结构与意义的关系使然：语言交际受意义驱动；为表达给定意义，一旦结构选定，相应的词汇也随之选定，反之亦然。下面的词语索引（JDEST 数据）表明，在 find＋形式宾语 it＋形容词＋to 不定式结构中，可选的形容词范围不外乎 *difficult*, *easy*, *important*, *hard*, *helpful*, *impossible*, *necessary* 等有限的几个。

1. of the humans who use them. People find it difficult to remember passwords and
2. or the computer to type them, we find it easy to give multiple versions of an
3. ch students have more freedom, may find it even more important to have a library
4. Richardson's Pamela the reader may find it hard to believe that the heroine, a
5. rihan I saw a dog (You may find it helpful to jot down the Beja words w
6. department. "At the same time, I find it impossible to hire a good clerk-typi
7. in one of these areas will usually find it necessary to cite generative researc

一个词项往往具有不同义项，词义不同往往意味着选用不同的结构。语言教学的正确决策是将结构及其相应的词汇制约一起教给学生，或者说将词汇意义及其相应的结构制约一起教给学生。传统语言描述对语法和词汇的"分而治之"是方法和体系的谬误，应用在教学上则是严重的误导，导致外语学习的低效和无效。语料库研究提供了丰富充足的证据，使词汇与语法融为一体成为可能。

2.4 成语原则、扩展意义单位与词块教学

语言使用中的词组特征久已引起研究者的注意。Dwight Bolinger(1976)称这种现象为语言预制件(Linguistic prefabs)和块结构(chunks)；Andrew Pawley & Frances Syder (1983)研究了词语化句干(lexicalized sentence stems)对形式选择、即时编码乃至流利言语产出的至关重要作用(见卫乃兴，2002：32)；James Nattinger(1988)，从功能的角度探讨了各种不同的词语片语，等等。但是，John Sinclair(1987，1991)依据语料库研究发现提出的成语原则(idiom principle)无疑标志着此类研究理论上的突破。Sinclair 认为，可用

两种理论模型解释语言的组织机制，一个是开放选择原则(open choice principle)，另一个是成语原则。前者视文本为一连串的空位，要用符合语法限制的词汇来填充，故又可称为空位填充原则；在每个空位上，几乎任何词都可能出现。后者则认为，"语言使用者有大量的半构筑词组可供使用，这些词组构成单一选择单位，尽管它们可被分割开来分析"(Sinclair，1987：320；1991：110)。根据成语原则，词组是主要的语言选择单位；词组包括各种不同的词语组合、搭配、固定词组、半固定词组、成语等等。继之，Sinclair(1996)又提出了扩展意义单位(extended units of meaning)的概念，即词汇的意义实现体现在更大的交际单位内。扩展意义单位包括：1) 词项的搭配；2) 词项的类联接型式(colligational patterns)；3) 词项的语义选择趋向(semantic preference)；4) 词项的语义韵(semantic prosody)等要素。比如，复合词"naked eye"的扩展意义单位一般为"visibility＋preposition＋the＋naked eye"(其中，"visibility"是搭配词的语义选择趋向，"preposition"和"the"是类联接选择，它们连同"naked eye"一起实现具体的意义和功能)：

too faint to be seen with the naked eye
it is not really visible to the naked eye ...

外语教学当然不可能、也不必要将这些抽象的语言学概念教给学生。但是，强有力的解释机制却给我们提供了重要启示：教学内容的焦点不能是单个的词项或语法结构；词语搭配、比搭配更大的半固定词组、融词汇与结构一体的词块等等必须予以更多的重视，如：

词语搭配：
commit suicide；commit crime；commit murder

半固定词组及其结构选择变体：
be sorry to keep you waiting
be sorry to have kept you waiting
be sorry to keep you waiting all this time

词块及其词汇选择变体：
I do not have the slightest/least doubt
I haven't the faintest/slightest/foggiest/remotest idea
I don't have the faintest/slightest/foggiest notion

2.5 潜藏规则

在一定程度上，任何语法书、词典、教科书对语言规则所作的描述和概括，都只能是不完备的、以偏概全的和限于表象的。这是由描述者或编撰者受个人学术立场限制，侧重一方面因素而忽视另一方面因素所致，也更由大量真实证据的缺乏，不得不依赖有限的经验和直觉所致。语料库语言学的数据、方法和手段使得真实语言使用呈现出的庞大体系、复

杂要素和深层次的潜藏规则有可能被较为全面地探索。传统方法难以发掘的潜藏规则正在逐步显现和被人们认知。以"perfectly"和"utterly"两个强化语的行为特征为例。多数词典都会告诉我们,两个词作强化语用时都表示"完全地"、"彻底地"之义,大体上同义。然而语料库证据会显示,"utterly"强化的形容词多为"消极涵义": *confounded*, *contemptuous*, *destroyed*, *devoid*, *discredited*, *doomed*, *evil*, *irrational*, *ridiculous* 等等(见下列检索自 JDEST 的词语索引),常用来表达说话者的批评、贬斥、否定态度。

1. n a spiritual world, they have been utterly confounded. Of their inspired humani
2. le merits: simple, iconoclastic and utterly contemptuous of the opinions of the
3. d faltered momentarily. The web was utterly destroyed. Birds aren't the only
4. the (male) director's thesis, being utterly devoid of any autonomous life or fem
5. e accuser and the charges now stand utterly discredited, lacking even the courag
6. immediate means. Ultimately, he is utterly doomed. Whilst sanctions remain, pre
7. terly evil. And unless the enemy is utterly evil, war is not justified at all...
8. diction? It is to us all part of an utterly irrational concept of "progress' tha
9. only ethically unaccept able, but utterly fallacious even in its own terms. Th
10. sure the troops would have found it utterly ridiculous. Dean had worked with a n
11. self-appointed task, which seemed "utterly formidable, completely ludicrous". S
12. cally according to her performance. Utterly hopeless when the season began, she
13. inseminated. There then followed an utterly disproportionate uproar... TX. — The
14. tay outdoors on a day so bitter, so utterly grim. My tooth, aggravated by the co
15. impair our humanity as to make life utterly meaningless. The active partici pat

与"utterly"截然不同,在 BNC 语料库中,"perfectly"强化的对象主要分属下列几个语义群:

1) 话者的态度: acceptable, defensible, tolerable, understandable, valid
2) 物的特质: intelligible, proportioned, symmetrical, balanced, elastic, matched
3) 合理性等: feasible, legitimate, adequate, harmless
4) 人物的品质: respectable, honest, capable, sane

可见,"perfectly"的搭配伙伴主要是些"积极涵义"类的词项,常用来表达说话者的褒扬、肯定、赞赏态度。语料库研究者趋于称之为语义韵,但并不尽然④。无论如何,这些都是语言使用中大量潜藏规则的冰山一角。将业已发现的潜藏规则知识应用于外语教学,无疑会促进学生建立正确的语言意识,掌握正确的规则知识。我们有理由相信,随着研究的深入发展和方法的改进,会有更多的潜藏规则被发掘、认知和描述,将它们变作适当的教学内容会有助于学生选择正确、得体、近似于本族语者语言的形式。

语料库语言学的研究发现是重要的教学资源,可用于不同层面的教学设计、教学输入和

过程活动。具体的应用程度、方法和技巧只能由教师根据教学对象、教学目的等因素确定。

3 语料库文本资源在外语教学中的应用

另一种重要资源是语料库的固有文本资源,可用作教学输入。但是,就目前语料库建设的情况来看,多数语料库的目的是研究性的,并非为了教学使用。所以,文本资源的首要使用价值在于教学研究。不过,部分文本仍不妨用作外语教学的辅助学习材料。

3.1 文本资源用作辅助学习材料

不同种类的语料库文本具有不同的用途。普通英语语料库的文类、题材、风格等特征使其适用于广泛种类和性质的英语教学课程。比如,经典的英语语料库 LOB、BROWN、BNC 等涵盖了十多个文类和数十个题材领域的内容,社会的、宗教的、艺术的、新闻的等等,可用于不同兴趣和目的的学生开展"自主学习"活动。而专门用途语料库的文类和题材使其适用于特定的学生群体使用。以 JDEST 的文本为例,该语料库目前库容达 4 000 000 余词次,涵盖了自然科学、工程学科、社会科学的 30 多个学科领域的多种题材内容,涉及的文类包括论文、研究报告、专著章节、教材、书评、研究成果简介等。研究者或外语教师可以根据教学大纲、学生兴趣、学生专业和难易程度选择 JDEST 的一些文本,用作高级英语学习者的辅助阅读材料,或者用于科技人员的写作训练。

由于多数语料库的建设目的是研究性的,文本的筛选、取舍至为关键。语料库研究者和外语教师至少需要考虑:1) 文本的趣味性,即它们在多大程度上能引起学生的兴趣?2) 文本的语域和风格,即这些体裁、题材与语言风格是否适合于现阶段学生的语言能力发展?3) 语言的难易程度,即文本所含的词汇量、句法结构等是否太难或太易?4) 过程活动,即文本内容有助于哪些学习活动的开展?具体来说,是适合于要点概括、逻辑推理、概念界定等活动的开展,还是适合于词义猜测、释义、观点讨论等活动的开展?选择合适的文本资源用作教学输入是有效利用语料库资源的关键环节;不合适文本的使用会导致整个努力失败,并产生对语料库资源的消极评价态度。

就教学而言,目前的许多语料库资源适合高级英语学习者。如利用 DDL 方法,高级英语学习者可以通过检索 BROWN, LOB, Frown, FLOB, JDEST, BNC 等本族语者语料库,获得词语、词组、结构等的典型用法,包括搭配、类联接、词块、语义韵等等。如果要使语料库资源充分服务教学和学习,还需要另外建立大型的、分类齐全的、分级的、为英语教学和学习量身定做的语料库,而且这类语料库必定是开放型的、动态的。

3.2 资源的共享与获得

文本资源只有用到了研究和教学实践中,才真正发挥了作用;束之高阁是一种巨大的浪费。目前存在的主要瓶颈是版权问题。建库者付出了数年的心血与劳动,也花费了相当的人力、物力和财力,自然有权保护自己的知识成果。但是,可以以一定的方式在版权拥有者和用户之间达成某种解决办法。根据我国语料库资源的具体情况,并借鉴欧洲的经验,不妨采用三种途径解决资源获得的问题:1) 建好的语料库实行商业销售,像 BNC 那样;2) 特别授权使用,即版权拥有者授予特定个人或群体使用库资源;3) 资源在线发布,即版权拥有者通过网络技术,将语料库转换为在线资源公开发布。

目前，国家社会科学基金项目"语料库技术与网络多媒体在外语教学中的应用"，已将JDEST、CLEC等语料库变为网络在线资源。现已实现700万词次的语料库数据的网络在线转换，用户可以实时、直接检索和下载。而且，该项目还为研究者和教学人员提供"在线词语索引"、"搭配词"数据、"扩展语境"等检索服务。用户可以无偿地获得部分资源和数据，运用于教学和研究，在一定程度上可提升资源使用的效益，缓解困扰我们的真实材料缺乏的困难，并尝试促进外语教学手段的改革。有关该项目的语料库在线资源情况，可参阅本书附录内容，或登录 http://corpus.sjtu.edu.cn 网站查询。

3.3 专门语料档案建设

专门语料档案建设是资源建设的重要组成部分。应当看到，由于种种因素制约，语料库资源的共享与无偿获得有其局限性。一般的外语教师有必要自己动手，建立适用于教学的各种语料档案。语料档案(linguistic archive)不同于严格意义上的语料库。语料库建设有一定的语言学标准、一套严格的语料抽样原则和相应的技术规范；它强调抽样的随机性、各种语料间的"平衡"、与同类语料库的"可比性"、整体语料的"代表性"等等。而语料档案则不同，没有太多的限制。具有初步计算机操作技能的外语教师皆可将具有一定价值的教学资料储存为机读文档、编制为某种结构，组成档案库。例如，供特定水平阶段学生使用的阅读资料档案库、某类题材或体裁的阅读或写作资料档案库、学生作文档案库、教师—学生课堂话语档案库、学生考试资料档案库，甚至各种教材资料的档案库等等。

由教师个人或较小群体合作者共建的语料档案，虽不具备语料库的许多特征和质量，但具有极强的针对性和实用性，可以直接使用于一定目的的教学，价值不可低估。在有限范围内用于教学的专门语料档案，避开了版权问题的困扰，可以弥补语料库资源的不足。

与语料档案建设相关的一个问题是网络资源的利用。当前的万维网提供了动态增长的海量语言资源。万维网语言资源不具备语料库资源的许多特征，语料具有一定的杂质性。但万维网语言资源也有其自身的特征：它高度动态化，语料不断更新；新词语、新用法往往最先出现；语料量之大，涉及语域之广，非任何语料库可比拟。所以，万维网语言资源必将成为外语教学重要的学习材料。关于这类资源的开发、利用，请参阅本书第二章。

4 过程活动的变化

语料库资源、数据与技术的应用，不可避免地要带来教学环节、手段和过程的变化。传统的外语教学基本上遵循一种自上而下的演绎式课堂活动模式；由于理念与资源的问题，课堂教学大体上由"陈述—操练—使用"(Presentation - Practice - Production)三个环节或步骤组成。语料库资源和技术有可能催生一种自下而上的归纳式学习模式。在这种模式里，学生首先接触的是大量的真实语言证据和事实，然后在教师的指导和启发下，识别语言现象，对其分类，最后归纳和概括(Identify - Classify - Generalize)。这就是 Tim Johns (1991)倡导的数据驱动学习方式。这种教学模式的优点之一是发挥学生的观察、判断和归纳能力，让学生自我发现、自我概括语言的意义和用法型式，属于认知性的学习风格。关于 DDL 的理念与方法更详细的介绍，请参阅本书第五章。

过程活动可采取多种形式。其中之一便是"课堂词语索引"(Classroom concordancing)。比如，用电化教学手段将"up"一词用作动词(表示"提升"、"增加"之意)的大量 KWIC 索引呈现给学生，让他们分析、归纳意义和用法：

1.	than his. Personally, my dander is	upped	by the ladling of sticky warm,
2.	Stewart. [p] Big-hitting John Daly	upped	his personal best by four yards at
3.	[p] ROYAL Bank of Scotland has	upped	its standard variable rate mortgage
4.	[p] British Rail, which earlier	upped	its 2.5 per cent offer, accused him
5.	Fridays. [p] Threshers and Beefeater	upped	market share by buying Berni Inns,
6.	off a takeover bid by De La Rue	upped	pre-tax profits in the half year
7.	for The Prudential said: We have	upped	premiums this year and will be
8.	and the United Arab Emirates had	upped	production to replace the loss of
9.	year ago. [p] The only sectors which	upped	production in the three months to
10.	time ago, and the cruelty stakes have	upped	so much since then. I mean, come on

"课堂词语索引"还可以采用"关键词屏蔽"技术，隐去位于索引行中间的关键词语，让学生猜测并填充。下列两组练习旨在提高学生对两个关键词及其半固定词组的用法意识：

(1) 关键词"largely"的屏蔽及填充

1.	e encounter between China and the West	_____	because of Hong Kong's unique cu
2.	s have received much attention. This is	_____	because of technological advances
3.	technology has become a policy issue	_____	because of the rising costs of
4.	The inability to find the effect was	_____	due to the manner in which social
5.	s through leaky modes. These losses are	_____	due to either the fibre design
6.	rrier, the good parts of this book are	_____	due to the efforts of the author
7.	r is much more potent, and is probably	_____	responsible for the bulk activa
8.	bonded electrons on oxygen. This is	_____	responsible for the distinction of
9.	Ohio. This vernacular architecture is	_____	responsible for the area's reju
10.	e of large landowning families were	_____	responsible for backward condition

(2) 关键词 "broadly"的屏蔽及填充

11.	oals which Hindus may hope to achieve.	_____	there are three types of religi
12.	f drilling machine in use but,	_____	, they may all be grouped into fo
13.	e comments made by the translators fell	_____	into five categories, concernin
14.	g from participation in the Net can be	_____	characterized as a shift from mo
15.	ever it can take other forms. It may be	_____	defined to include unreasonable
16.	principles of education are usually so	_____	expressed as to make their expli

17. ency and reducing jet exhaust noise". _____ speaking, in the original pure
18. operational even at lower temperatures. _____ speaking, this process is a b
19. ent of immune complex diseases can be _____ classified as endogenous, infect
20. cellular system. A cellular system, _____ defined, is composed of several

语料库资源与技术的应用，有待于教师开发出多种多样的教学形式和环节。但是，首要的问题仍是师生的认识与接受。任何新的教学理念、方法和技术的使用与普及都需要某种程度上的"教师训练"和"学生训练"。当教师和学生逐步接受并熟悉新的资源与技术时，其潜在的价值和作用就会实现，外语教学的环节与过程就会发生较大的变化。

5 学习者语料库数据的利用

学习者语料库数据对外语教学具有独特的使用价值。它忠实记录了学生语言产出的事实，含载了学生的总体语言能力状况及发展规律，特定学生群体在词汇、语法、语篇等领域的具体行为模式与错误特征、特征和问题产生的原因以及学生的学习策略等重要信息。目前，国内语料库研究者已开发出 CLEC（桂诗春，杨惠中，2003）和 COLSEC（College Learners Spoken English Corpus）（杨惠中等，2005）。前者库容为 1 000 000 词次，涵盖中学学生、大学英语学生和英语专业学生不同阶段的作文语料，含有 60 多类言语错误的赋码。后者库容 720 000 词次，涵盖大学英语学生的英语会话语料，含有 104 类主要的语音错误赋码和话轮转换、非言语声音等标注信息。学习者语料库为英语教学提供了决策依据，也为教学内容提供了资源和数据。

学习者语料库数据应用于外语教学，主要依据中间语对比分析的研究成果。近年，我们基于 CLEC 和 COLSEC 两个学习者语料库，对中国学生的词语搭配特征、类联接使用特征、强化语使用特征、主题词及词语网络特征等进行了一系列研究，对教学提供了有益的反馈和启示（参见本书第三部分）。通过对比分析学习者语料与本族语者语料的特征差异、学习者英语语料与母语（汉语）语料的特征异同、不同背景的英语学习者的数据异同，可以发现学生在诸多领域存在的问题和困难，有的放矢地设计教学内容和采取校正措施。关于学习者语料库的研究与使用，请参阅本书第三部分。

结论

本章从多个层面和视角讨论了语料库资源与技术应用于外语教学的有关问题。语料库语言学的理念、方法和研究成果业已确立了它作为一门独立语言学科的地位；它对外化语言的真实运作机制、语言行为模式的新发现和新描述对普通语言学和应用语言学的一些理论形成了挑战，尤其对描述语言学体系产生了巨大影响。同时，语料库语言学是一门应用性极强的学科。它对应用语言学和外语教学具有现实的使用价值。而且，众多语料库研究者本身就是应用语言学者和外语教师，他们的研究目的和研究内容直接相关于外语教学。从文本、研究成果到相关技术手段，语料库资源与技术在外语教学中的使用价值急待开发。以适当的方式解决版权问题、有选择地利用语料库文本，是开发资源、缓解外语教学真实材料缺乏的途径之一；语料库研究成果应用于教学过程，将有助于学生习得正

确的语言知识，建立恰当的语言意识。在外语教学大纲制定、教学内容设计等方面，语料库可提供不可替代的数据。我们有理由认为，语料库资源与技术，包括学习者语料库研究成果，在外语教学中具有极大的使用价值与开发前景；通过一定程度和阶段的教师训练和学生训练，推广语料库资源、技术与方法在外语教学中的应用，必将引起外语教学手段、环节和过程的变化，促进外语教学的改革与质量、效益的提升。

目前，万维网为我们提供了动态增长的海量语言资源，其语言使用特征和网络技术对外语教学的影响正日渐显现。如何应对新的影响和变化，在外语教学与研究中利用万维网语言资源，已是亟待研究的课题。我们将在下一章里讨论有关问题。

注释

① 不妨认为语料库语言学有两大特色方法，一为 Geoffrey Leech 的"基于语料库"的研究；该方法不太触动原有的语言理论框架与描述体系，主要利用证据详尽有力地证实理论和阐述范畴。另一为 John Sinclair 的"语料库驱动"的研究；该方法强烈地质疑原有理论的合理性，主张建立新的理论体系，认为语料库研究会给语言学领域带来大的变革。然而，Antoinette Renouf 与作者讨论时认为，二者并无根本的区别。

② Widdowson(2000)认为，语料库语言学成果不能直接应用于外语教学，因为语料库研究只是对语言事实的部分研究，有其局限性；语料库数据不能解决语言编码的可能性问题；语料库的文本数据是静态的和脱离语境的等等。他的这些观点有一定代表性，是本章内容无法回避的。但本章仅就紧密相关于语料库数据应用于外语教学的有关问题讨论，并不试图一一回应他的观点。我们会另撰专文，与 Widdowson 一一讨论有关问题。

③ 国内数家出版商与本章作者谈论过语料库建设事宜，旨在词典开发，但他们意想中的语料库远非一般意义上的语料库。

④ 语料库研究者 Louw、Stubbs 和本章作者将此类搭配现象归为 Semantic prosody（见卫乃兴，2002《外语教学与研究》第 4 期"语义韵研究的一般方法"）。但是，据本书另一作者李文中从英国寄来的讲义，伯明翰大学 Susan Hunston 不同意此种观点。Hunston 认为，单个词的语义韵并不存在。