绪　　论

当代语料库语言学的应用与发展

导　言

本篇章收入了论文 5 篇，论文 1 至 3 是作者近年来出席澳大利亚悉尼第 21 届和瑞典哥德堡第 23 届国际语料库语言学年会以及 2001 年英国兰卡斯特大学语料库语言学国际研讨会之后撰写的大会综述文章，主要介绍国外近年来语料库语言学在语料库建设，语料库检索工具开发和语料库应用于各个领域的最新发展。论文 4 侧重介绍国内外学者在应用外语语料库于外语教学领域方面的发展和成果；论文 5 则通过调查归纳国内近年发表的语料库研究论文内容来综述国内语料库语言学研究的现状。

从这批论文看出，目前国外对语料库的应用研究已经从最早期的词汇分析和词典编撰等发展到自然语言处理、人工智能和语言教育教学等多个领域。近 10 年来，利用语料库促进语言教育尤其是外语教育教学已经在国外形成了一种新的趋势（见 Granger，1998；Flowerdew，1996；Johns，1996；Kennedy，1992，2003；O'Keeffe & Farr，2003；Tribble，1990 等）。国内在这方面的研究也有长足的发展（见桂诗春和杨惠中，2002）。但综观近年来国内的研究成果，大多是集中于对学习者中介语语料库的研究，而且往往停留在外语学习者对目的语某些语言形式的过量使用、过少使用和错误使用等方面的对比研究。国外的语料库语言学研究虽然已经延伸到教学大纲设置、教材编写和课堂活动设计等方面，但似乎还缺乏比较系统的基础理论研讨，而且在应用原则和方法上也有待进一步的归纳和提炼。本书的理论篇与实践篇就是对这两个方面的着力探讨。

1 国外语料库应用介绍*

何安平

《用语料库研究语言》是 1996 年由英国 Lancaster 大学的珍妮·托玛斯(Jenny Thomas)和米克·肖特(Mick Short)联合编著出版的一本语料库语言学研究文集(以下简称《文集》)。它主要反映了 20 世纪 90 年代欧洲一批大学,尤其是英国 Lancaster 大学的专家学者在建设、开发和研究语料库等各个方面所取得的成果,同时也是来自英国、挪威、德国、瑞典各大学的 27 位撰稿人以及编者和出版商(Longman 出版公司)共同献给语料库语言学的开拓者之一,Lancaster 大学计算机语料库研究中心(UCREL)的创始人 G. Leech 教授 60 岁生日的礼物。G. Leech 和他所在的 Lancaster 大学从 20 世纪 70 年代开始就一直走在全世界语料库语言学学科发展的前沿。他们早在 1978 年就与挪威的 Oslo 大学和 Bergen 大学联合建立了世界上第一个百万词次的英国英语语料库 LOB;1984 年创建 Lancaster 大学计算机语料库研究中心成为国际上语料库语言学研究的重点核心基地并率先研制英语自动词性标注系统 CLAWS;1987 年与 IBM 公司联合建成有语音标注的英国口语语料库 SEC;1994 又与英国图书馆、Oxford 大学、Longman 出版公司等联合建成了上亿词次的英国国家语料库 BNC。正是由于他们长期不懈的、富于创造性的努力,才引导和促进了近 30 年来语料库语言学异常迅猛的发展。

20 世纪 90 年代以来是语料库语言学发展最快的时期。其趋势主要表现在以下三个方面:(1)大规模,多品种语料库的建设,(2)对语料库的深加工研究,(3)语料库在与语言相关的各个领域中的广泛应用。

首先是计算机等电子传播和储存技术的发展带动了语料库的建设规模迅速扩大。20 世纪六七十年代建立的第一代电脑语料库,如拥有百万词次的 BROWN 和 LOB 等在当今看来已算不上大型语料库,因为 20 世纪 90 年代建立的 BNC 就已经达到 1 亿词次,而 21 世纪初已实现通过因特网检索的在线语料库(WEBCORP)更是达到了 10~50 亿词次,平均每天更新的新闻语料网页就达 200 万(M. Pacey and A. Renouf, 2001)。现时国际流行的语料库类型

* 本文为外语教学与研究出版社 2002年出版的《用语料库研究语言》的导读,此处稍有改动

有：

- 原始语料库(Raw corpora)，即将现实中使用的口头和笔头语用文字形式收集起来，按一定原则（语域、语体、历时、共时等）归类汇编起来的各种语料库，如 1999 年挪威 Bergen 大学新版的《ICAME Collection of English Language Corpora 》光盘就包含了 1 千 7 百万词次的世界各地的英语口语和书面语，成人语和青少年语，当代语和古代语等 21 个语料库（详细介绍见何安平，2001 : 75 ~ 78)。
- 附码语料库 (Annotated corpora)，即对原始语料进行了词性、语法、语音、语义、语篇乃至语用标记附码的语料库。如已作词性附码的 BROWN 和 LOB,已作语法附码的 ICE - GB （《国际英语语料库英国分库》),已作语音韵律附码的 LLC 和 SEC （《英语口语语料库》)等。
- 平行语料库 (Parallel corpora)，即两种或多种语言在句子乃至单词短语层面上实现同步对译的互动语料库，如英语与法、德、西班牙等语种的平行语料库 CRATER (McEnery and Oakes , 1996，见《文集》第 13 章)
- 学习者语料库 (Learner corpora)，即非母语学习者的口头和书面语料库，其中包括注有学习者拼写和语法错误标记以及改错提示的语料库。如，ICLE（国际英语学习者书面语料库）和 LINDSEI （《国际英语学习者口语语料库》） (Granger, 2000)
- 网格式语料库（Lattice corpora) ，即对自然语言（口语和手写语）进行自动语音和手写体识别处理之后生成的语料库 (Atwell，见《文集》第 10 章)。

可见，建立大型语料库已经不再是很困难的事，接下来的是如何建立各种有地域、语体、乃至语篇特色的大型语料库以及如何对现有的语料库进行深层次的开发与研究。语料库的深加工研究包括：(1) 对原始语料作多层面信息的自动附码处理，(2) 对多功能语料库检索工具的开发研制， (3) 对自然语言的自动识别和电脑化处理系统研究，等等。

20 世纪 90 年代以来语料库自动附码系统的研究可以说是从词性附码向着句法、语义和语篇附码的方向发展。以 Lancaster 大学语料库研究中心的研究成果为例,他们早在 80 年代初就已经研制出 CLAWS1 对 LOB 语料库做自动词性附码，而 90 年代中期的 CLAWS4 版本已经发展到可以自动标注句法分界和某一类词性（如动词这一大类别）底下的各个分支类别（Garside & Rayson, 1996，见《文集》第 11 章）。他们在 1993 年研制出了自动语义附码器 SEMTAG，使用它在已作词性附码的语料库基础上可以将语篇的每一个词或词

组自动归入预设的数十个语义场分类系列，包括对反义词、多义词的附码，对修饰词和被修饰词搭配关系的标注等等。经过语义附码的语料在分类检索和量化之后便可以得出该语篇的语义特征的总体相貌和分布状态。该中心同期研制出的语篇附码器 Xanadu 则可以自动标注某些语篇结构词如前指代词 anaphora 等，以用作文体、篇章结构和语用等方面的分析（Meunier，1998：25）。

语料库自动附码研究的发展带动了语料库检索工具的速度与功能的升级。除了当今较流行的 TACT 和 Wordsmith 之外，还有许多针对专门语料库设计的检索软件，如专作单词类别和频率分析的 VocabProfile（详见附录 1），专作口语语篇内容分析的 ACAMRIT（Wilson & Leech，1993）等。强有力的检索工具能够帮助研究者对语料库进行更加深入的多方面的调查分析。例如，1999 年出版的国际英语语料库—英国分库 ICE－GB 光盘就可以在已经完成词性、句法和语篇结构词附码的百万词次语料库里进行词汇、句法和语篇等多层面的英语研究和分析。该光盘里装有最新版本的检索工具 ICECUP3.1。其中的 Fuzzy Tree Fragments 系统可输入设定的句法模式做由下而上的或选择性的句型检索，透过检索和观察词汇或句型的语境共现，可以了解语法的抽象模式及其实际运用情况。另外其 corpus map 功能能显示有关语料库来源的社会信息，其口语语料部分还可以同步播放数码录音等等（Wallis 和 Aarts，2000）。至于其他一些作了语音语调附码的英语口语语料库如 LLC，SEC，COLT（《伦敦青少年口语语料库》）等，如今也已经将其配套录音转换成数码录音文件，实现了在电脑屏幕上口语文字、语音和音韵图像的共现，并且取得了一系列对语音语调和句法、语用等多层面交叉研究的研究成果（见 Wichmann，2000、2001）。

归结起来，有了前面谈及的各种类型的语料库以及相关的检索工具，人们就可以探索语言在历时的和共时的使用过程中呈现出来的各种规律性或倾向性的信息。以对英语语料库的研究为例，它可以检索出：

- 反映英语词汇在实际使用中的词频表
- 以英语某种句法、语义或词性为检索单位的分类信息
- 在英语中反复出现的习语或惯用搭配模式
- 在英语中反复出现的某些词汇与特定语法结构同现的信息
- 与某种语义场相关的词汇构成
- 反映某类语篇特征的词汇、语法结构和语篇结构词的使用特征
- 对附码语料库中两个／多个层次标记相互关系（如语音和句法的制约和互补）的信息
- 识别自然语言（语音或手写文字）的信号滤网模式

- 不同研究单位或学派对英语句法信息附码模式的对比
- 学习者对目的语的某些方面过量使用、过少使用和错误使用的倾向以及纠正的模式
- 双语或多语种语料对译时的匹配和非匹配信息

（参考 Atwell，1996。见《文集》第 10 章）

语料库研究的第一步虽然只能检索出语言的符号形式，但是由于它能够同时呈现出这些形式的语境共现（concordance or linguistic ecology ）和频率分布，这些形式符号就可以逐步变成有各种意义和功能的信息。人们正是在获得这些信息的基础上，也包括在建立语料库、研制语料附码和检索系统的过程中，逐步深化着对语言本质的理解，更新着对语言研究的思路和方法，构建着新的语言研究模式。因此语料库研究并不仅仅是一种语言研究途径和方法的更新，它更孕育着新的语言哲学观念。《文集》一书收入的 16 篇论文就是从不同的角度反映出这些新的认识、新的方法和新的语言研究模式。

《文集》所面对的读者，是那些对计算机略有所知但仍然需要学习了解如何使用语料库来进行语言研究的语言工作者、教师和学生。读者可以根据个人的需要和兴趣通读全书或选读其中的篇章。全书收入的论文分别归入 4 个部分：

1. 用语料库研究语言的意义和作用
2. 基于语料库的各层面语言研究个案
3. 语料库在计算机语言处理方面的应用
4. 语料库在与语言相关的其他领域中的应用

读者若想了解语料库语言学的发展历史，尤其是在欧洲或英国的发展历程的话，可以首先翻到《文集》的第 1 部分。这里收入的 4 篇论文主要是 G. Leech 的挚友和同事回顾 G. Leech 30 多年来开拓和发展语料库语言学的辛勤工作和令人瞩目的成就，也包括他们本人从事语料库语言学研究的经历。 J. Svartvik （LLC 语料库主持人）在文中引用另一位语料库语言学开拓者 J. Sinclair 的话语来总结 30 多年来语料库语言学发展所带来的观念转变：“处理数百万词次的语料在 30 年前被认为根本不可能；在 20 年前被认为有一丝可能但难以想象；在 10 年前被认为很有可能但仍然困难重重；而到了今天，这一切都已变得十分平常和普遍。”Svartvik 预言今后的计算机将会速度越来越快，体积越来越小，价格越来越便宜；而语料库的发展则是规模越来越大，质量越来越好，使用越来越普及。语料库研究正酝酿着对语言学理论和研究方法的重大

突破（见第一章）。

Lancaster 大学教授 G. Sampson 从事了 20 多年树型图句法研究，他这次从英语的口语和书面语，成人读物和儿童读物，通俗题材和学术论文等多种语料中检索到不少英语多重镶嵌句（central embedding）的实例。而这种现象却曾被语言学家认为是违背常人心理认知规律的，不自然的，不可能出现的语言形式（见第 2 章）。Sampson 用自己的语料库调查实践证明研究英语句法规律的最佳途径之一就是去观察和寻找大量在现实生活中出现的语例，而这正是用语料库研究语言的最突出特点之一。

在这一部分收入的另外两篇论文分别是伦敦大学 Quirk 教授等人利用语料库对外来词（obtrusive language）的研究和 Lancaster 大学 Knowles 教授对各语言学流派的研究方法的评述。Knowles 认为各种语言学理论的差异在本质上是其组织语言信息的方式不同，而语料库则是构建任何语言模式都不可或缺的重要基础（见第 4 章）。

读者如果对语料库研究的具体方法和原则感兴趣，可以细读《文集》的第 2 部分。这里收入了英国学者们利用语料库对英语的词汇、语法、语用和文体等多个层面的研究成果。首先是 S. Johansson（LOB 语料库的主持人）等对英语高频动词 get 的研究，他们通过检索口语和书面语料库对 get 的多种用法进行量化统计和实例分析，并同其他一些常见动词（如 have, be, keep）的用法作对比，结果发现 get 在使用形式上虽然没有语法化（如演变为助动词），但却要承认它已在英语的核心语法中占有重要席位（见第 5 章）。

接着的研究是 S. Greenbaum（国际英语语料库 ICE 的主持人）和他的同事从 ICE 中抽取等量的口语和书面语语料来分析英语谓语动词之后的补足语。通过对这类补足句的形式分类以及同其他从句（如状语从句，关系从句）的对比，归纳出这类句型的多种语法功能以及它们在不同语体中的使用差异（见第 6 章）。

J. Thomas 是本书的主编之一，擅长运用语料库来分析英语会话的语篇特征和语用功能（Thomas, 1995）。她这次和同事们尝试使用自动语义附码器 SEMTAG 将医生和病人交谈的语料进行语篇的语义对比分析（content analysis），也就是将不同的医生使用的词汇分别归入 SEMTAG 设定的语义场分类，检索之后归纳出各人话语的语义特征和语用功能差异。她对比以往语用分析常用的个案观察法和问卷调查法，指出利用语料库附码和检索的办法虽然在语言特征的判断方面无重大突破，但实施起来却快捷、准确、省事省力，而且证据充实，令人信服（见第 7 章）。

M. Short 是本书的另一位主编。他早在 20 世纪 80 年代初就与 G. Leech 合作构建了英语直接引语和间接引语的分类模式（含有话语行为 speech act 和思索行为 thought act 两个维度）。这次他又将该模式应用到小说体和非小说体的语料库中作验证分析，结果对原来的模式进行了扩充和修正（见第 8 章）。值得一提的是 Short 利用语料库做文体风格的研究一直在延续，他在 2001 年 Lancaster 国际语料库研讨会上公布的最新分类模式已经扩展到三个维度：话语行为，思索行为和书面表达行为 written act（Short，2001），从而使该模式更具概括性和解释力。这是应用语料库创建语言分析模式的又一力证。

倘若读者对发展语言的计算机处理技术颇有研究，《文集》的第 3 部分提供了来自 Lancaster 大学计算机语料库中心的研究成果作为参照。这里有从自动词性附码系统升级到句法附码系统的研究（见第 9 章）。有对口语和手写语等自然语言自动输入和自动识别的计算机处理系统研究（见第 10 章）。也有介绍该中心研制自动词性附码系统（从 CLAWS1 到 CLAWS7）十几年当中各时期各版本的设计原则和功能特征，以及将 CLAWS 运行于最新语料库 BNC 的试用报告（见第 11 章）。最后一篇还介绍了一种作为对上述多种附码器进行检验或补充的，基于即席语境的词性、句型和语义等特征的识别系统研究 template analysis（见第 12 章）。

读者的兴趣若是在语言的习得和教学方面，可以翻到《文集》的第 4 部分，那里有语料库研究成果在语言教学、测试和学习词典编撰等方面的应用。第一篇是 Lancaster 大学 McEnery 博士等人谈双语或多语语料库如何实现词句相互匹配和对译的问题，这对于机器翻译，双语对比研究和翻译教学都很有意义。第二篇是德国学者 Mindt 利用以英语为本族语者语料库来考察德国中学的英语教学大纲中某些语法项目的编排次序和讲解重点，结果发现许多英语教科书中的语言知识编排和讲解有脱离英语本族语者的真实使用情况的倾向，他指出现行教科书中这种脱离实际、过度简单化的知识传授的误导现象需要通过语料库调查和研究来加以解决。第三篇是 Lancaster 大学的 Alderson 对语料库应用于语言测试的可行性探讨。第四篇是 Longman 出版社的 Summers 谈如何运用语料库最基本的，也是最突出的检索结果——词型频率统计以及由其语境共现产生的词义——来编撰专供外语学习者使用的语言词典。

读者在阅读这些语料库的研究论文时，可以尝试抓住以下一些问题来掌握信息要点：

1. 研究的目标是什么（语言的哪个层面、范畴或是相互领域的交叉）？
2. 采用什么样的语料（语料库的名称，抽样的原则或方法）？

3. 使用了哪些检索工具（检索系统的名称，其功能特征和检索方法）？

4. 研究者对首次的检索结果如何做进一步的分析（如分类、归纳、统计或进一步的检索）？

5. 结果发现了什么？如何对这些发现进行解释（形式、意义和功能）？

6. 研究的结论是什么？

阅读完后，读者还可以反思自己从中学到了什么，还有哪些不明白的地方，可以通过什么途径来解决，例如尝试与作者用 email 交流，或通过书后的参考书目作进一步的阅读等等。

应该看到，《文集》一书由于受其出版的目的和时间的限制，在客观上也有局限性。一是区域性的局限。它仅仅收入了欧洲某些国家的，尤其是英国的，而且英国中又以 Lancaster 大学为主的语料库研究成果，而没有美洲，大洋州以及亚非洲等其他地区的研究成果。二是时间上的局限。它仅仅反映了 20 世纪 90 年代中期之前的语料库研究成果，而语料库研究的发展却是日新月异。尽管是这样，读者们仍然可以从中了解到语料库语言学的发展历史，语料库的建设原则、研究方法及其应用的深度和广度。特别是书中谈及的一些研究课题，如 Short 的引语分类模式研究和 Sampson 的树型图句法模式在语料库自动句法附码系统设计等方面的应用研究等等，如今在看到它们的最新成果时，却仍可以从《文集》中追溯到当年起始的背景和各个阶段性的成果，很能够形成一种研究思路的延续、衔接和完整性。

语料库语言学迄今在国内外仍属一门年轻的新学科。正如语料库研究专家 Sampson（2001）前不久所说“我们到现在为止仅仅是淌足语料库大海边上的一些小水泡，然而，如果没有像 G. Leech 那样一批老前辈引导我们来到沙滩海边，我们可能现在还停留在远离大海的陆地。”希望这本《文集》在国内的出版能够激起更多人对语料库这片浩瀚大海的研究兴趣，并且能够运用国外的以及国内自己建立的语料库取得更多更大的研究成果。

2 语料库语言学应用的最新发展

——ICAME2000 年悉尼国际语料库语言学研讨会综述 *

何安平

跨世纪的语料库语言学正在越来越广泛地应用于语言教学、词典编辑、职业培训、自动信息提取和自然语言生成等领域。2000 年 4 月 21 日至 25 日，国际语料库语言学协会（ICAME）第 21 届研讨会 ICAME2000 在澳大利亚悉尼的麦考尔大学召开。ICAME 创建于 1977 年，最初是由欧美一群语料库语言学的开拓者自愿组织起来的非官方学术团体，自从 1979 年起，每年轮流在世界各大学召开一届年会以交流语料库语言学的研究成果，包括论文、语料库和系统软件等等。出席本届年会的有来自欧洲、亚洲和大洋州等 17 个国家的 62 位语料库研究专家和学者。其中有语料库语言学的创始人和早期研究专家如 Lund 大学 Jan Svartvick，Oslo 大学的 Stig Johansson，Helsinki 大学的 Matti Rissanen，Goteborg 大学的 Goran Kjellmer，Bergen 大学的 Anna Stenström，New South Walse 大学的 Peter Collins，Victoria 大学的 Graeme Kennedy 等，也有一批成果丰富的中青年语料库研究专家如 Louvain 大学的 Sylviane Granger，Lancashire 大学的 Anne Wichmann，Macquarie 大学的 Pam Peters 以及来自 Zurich 大学的 Sebastian Hoffmann 等一群博士后研究员。出席大会的中国学者除笔者之外还有分别在伦敦学院和悉尼大学从事自然语音处理和语料库句法附码研究的方成玉博士和乔洪亮博士（音译）。

今年 ICAME 大会的主题是：语料库语言学应用的新发展。在这次大会上发表的 47 篇论文大致可分为 6 个方面的议题：

1. 语料库在社会语言学领域的应用
2. 跨语种、跨地域的语料库研究
3. 口语语料库的研究
4. 语料库与语言教学
5. 语料库软件演示

* 本文原载于《外国语》 2001 年第 2 期，此处稍有改动。

6. 语料库的历时性研究

从以上分类看出，多领域、多层面、多功能的语料库应用研究是这次大会的一个显著特点。大会的论文既有较传统的对已有现成附码的语料库作语音、词汇或语法等层面的描述，也有对目前尚难于附码的语意、语用和文体修辞等方面的研究；研究热点不仅是基于纯文本文字构建语言理论模式，同时也开始研究语料本身所负载的各种社会信息（如语料来源者的社会背景、身份、性别、年龄等）；研究的方法既使用基于语言规则的检索分析，也运用概率的分析途径乃至借用其他领域学科的方法等等。以下分别对各议题中的部分文章作一介绍。

1. 语料库在社会语言学领域的应用

来自惠灵顿 Victoria 大学的 J. Holmes 和 R. Sigley 合写了 *What's a word like girl doing in a place like this? Occupational labels, sexist usages and corpus research*，他们调查了以 60 年代语料为主的 BROWN 和 LOB 语料库和以 20 世纪 90 年代语料为主的 FROWN 和 FLOB 语料库中的 woman, man, female(s), male(s), girl(s), boy(s)等词，发现 90 年代英美笔头语料库中有女性标记的词频在增加，男性标记的词频减少，尤其是以 girl 合成的词频大幅度减少，而由 female 或 male 合成的词（尤其在表示行业和职称方面的用词）在增加，这就从一个侧面反映出社会对性别差异的注意度在上升，即女性对社会参与程度和享有职业平等机会在增加。然而，对职业称谓中的 -girl, -boy 的合成词的语境和含义分析则表明 -girl 仍然比 -boy 更多地被用来指称成年人，尤其是侧重外貌长相的职业称谓如 dancing girl, chorus girls, callgirls, streetgirls，而 -boy 则更多被用来指称具有某种职业或工种能力的少年如 bellboy, busboy, cabin boy, cookboy, messenger boy, newsboy 等等。

挪威 Bergen 大学 Anna Stenström 的论文 *From slang to slanguage: a description based on teenage talk* 利用伦敦青少年口语语料库来探讨“俚语”和“非俚语”的分界线。她发现“俚语”的定义远不如以往词典所给定的那么简单，诸如“起源于大都市中心区”，“流行于一般教育水准之下的青少年的口头习语”之类，她认为应该根据真实语料建立起描写框架，其中应含有以往不属于俚语语言特征的 slanguage。

2. 跨地区、跨语域的语料库研究

这一部分的论文首先集中了几篇来自不同英语国家和地区对英语情态动词的使用研究。其中瑞典 Vaxjo 大学 Hans Linguist 的论文“*Modals in spoken British, American and New Zealand English*”通过调查英国、美国、新西兰等国数百万词次的新闻广播语和日常口语语料库中一系列情态动词的肯定／否定式使用频率，推断出英语情态助动词是表现语言历史变迁和词汇语法化过程的有力证据。

意大利 Verona 大学 R. Facchinetti 的论文“*From teens to twenties : the use of modal verbs by British young speakers*”通过对英国口笔头语料库和青少年口语语料库中情态助词的社会变体和语用功能分析，发现 can, would 和 will 的使用最为广泛；must 和 shall 逐渐减少；不过 will 和 may 在私人会话中明显减少，would 则在公众演说中频率最高。还有年轻人普遍使用情态助词来表达试探性语气，尤其是用 might 和 should。然而在非试探性的语境场合里，女孩却比男孩更多使用 can, shall, will 和 must。

澳洲 Macquarie 大学的 S. Butler 介绍利用东南亚各国区域性英语语料建成的英语方言语料库用于编撰英语方言词典和其他语言参考书。使用者只需键入一个检索词便可立即获得该词在某个（些）国家和区域的使用地情况，包括在不同语体或篇章类型中出现的频率和例句等资料，对于区域性英语变体的研究很有价值。

跨语种、跨区域方面的语料库研究论文还有《对“that's what”及其在德语和挪威语中对应语的多元对比分析》，《跨文化会话中的模糊语言研究》，《世界各地英语中情态功能的形容词短语》，《美国英语和英国英语中的零主语关系从句》等等。

此外，还有相当数量的论文是探讨英语语法的各个层面，例如研究英语不定代词复数形式的单数处理现象，复合分词的前置修饰语，现在完成时的时间规定，wh-从句的统取现象，whom 的消亡及其相关问题，以及组词构建的语言模式，论多义词的翻译等等。

3. 口语语料库研究

对口语语料库的建设和开发研究是这次大会的热点话题之一。 Bergen 大

学 Knut Hofland 在会上介绍了如何将伦敦青少年口语语料库的录音带转换成数码录音文件，并由英国 Softsound 公司协作把它制成可在电脑屏幕实现以 5～15 秒为单位的口语文字及其语音共现，因而可以对大规模的语料库检索结果进行语音，语调及其词汇和句型等多层面调查研究。

英国 Lancashire 大学 Anne Wichmann 在会上所作的题为“ *Investigating the intonation of interpolation in the ICE GB corpus* ”的报告则对即席会话中的“插入成分”（parenthetical element） 进行话语特征→韵律模式→话语结构→语用功能→交际手段的系列研究。由于“插入语”通常具有 backtracking 的特征，即重复插入语之前的部分词句，所以她首先从语料库中检索一批具有该特征的话语，截取其中的插入语片段作语音图像分析，从调幅和语速等方面归纳出三种主要韵律模式。然后将之推及其他具有类似语调特征的话语，结果发现这些话语大多具有“自我修补”（self-correction），“扩展性修饰”（extended modification）；“诱发听者反馈”（eliciting response）“提供参照信息”（establishing reference）等语用功能。在此基础上，她进一步探讨这种“插入性的语调特征”到底在多大程度上反映出说话者的在线思索及其记忆的局限性，又能在多大程度上帮助实现话语意义的表达和维持双方交流的会话目的。

关于口语语料库的研究论文还有：英国 Lancaster 大学和英国 Telecom 电话公司的合作研究，分析接线员与用户的对话语料 transactional dialogues）与日常会话的区别，对话的目的对其用语方式的影响以及该类对话的词汇和语篇特征等等；德国 Wuerzburg 大学的 Ilka Mindt 讨论英语降调的句法和篇章功能；还有对荷兰口语语料库和德国口语语料库的建库报告等。

4. 语料库与外语教学

语料库研究在外语教学中的应用近年来受到越来越多学者的重视。国际英语学习者语料库（ICLE）的主持人，比利时 Louvain 大学的 Sylviane Granger 对近 10 年来语料库语言学在英语教学中的应用做了回顾和总结，其中谈到英语学习者语料库的建库原则，学习者的背景分类，错误标码方式以及如何与二语习得和外语教学理论相结合构建分析模式等问题。详情可参阅由 Granger 于 1998 年编著出版的 *Learner English on Computer* 和 Granger 等人 2002 年主编出版的 *Computer Learner Corpora , Second Language Acquisition and Foreign Language Teaching* 及网址 http://www.fltr.ucl.ac.be/FLTR/GERM/ETAN/CECL/cecl.htm（有关于学习者语料库的目录）和 http://data.geogetown.edu/FLIRT/docs/in-

dex. htm（可下载 ESL语料库）。目前，另一个大型国际英语学习者口语语料库（LINDSEI）已在筹建之中。

关于这个议题的论文还有《对日本英语学习者的连词使用调查》，《高级英语学习者的作文研究》，《语料库语言学中的计算机语言交际》等。笔者这次发表的论文也是针对中国学生英语作文中过量使用连接词 SO 的现象，研究 SO及其汉语粤方言中对应词‘瞰’的语法化和语用化表现，从而探讨它对英语教学的启示。

5. 语料库光盘及检索软件演示

来自 Bergen 大学的 Knut Hofland 介绍了最新推出的语料库光盘《ICAME Collection of English Language Corpora》。它汇集了当今在国际上广为使用的 20 个英语书面和口语语料库，共达 1 700 百万词次，比 1992 年出版的语料库大全 ICAME CD-ROM 增加了 17 个语料库，其中的书面语料库除原有的 BROWN 和 LOB（tagged/untagged）之外新增了 20 世纪 90 年代以后建成的均为上百万词次的 FROWN（美国），FLOB（英国），ACE （澳大利亚），WCE（新西兰），EAC （东非）等；口语语料库除了原有的 LLC 之外新增了 SEC（英国），COLT(伦敦青少年口语)，WSC（新西兰），EAC （东非）等；历时英语语料库除了原有的 Helsinki 之外，增添了 HCOS （古苏格兰英语），ICAMET（14-17 世纪英文书信）等；另外还有两个已作句法附码的 LOB 和 Polytechnic of Wales 语料库。尤其值得一提的是该光盘内装有多个检索系统如 WordSmith，TACT，Wordcruncher 等，可用来检索光盘内或光盘之外的其他语料库（详情可查 http://www. hit. uib. no/icame/cd 或电邮 knut. hofland@ hit. uib. no）。

伦敦大学的 S. Wallis 和 B. Aarts 在会上演示了最新版本的 ICECUP3. 1。这是一个比 1998 年版的 ICECUP3. 0 更强有力的检索工具，主要用于研究已做了句法标注的语料库如 ICE—GB （《国际英语语料库 英国子语库》）。其中的 Fuzzy Tree Fragments 系统可输入设定的句法模式做由下而上的或选择性的句型。使用者可以学习语法的抽象模式及其实际运用情况，透过 corpus map 可显示语言现象的社会分布情况，可观察词汇或句型的语境共现，可做频率统计，还可同步播放数码录音等等。作者最后指出 ICECUP 下一步的发展方向是帮助使用者从语料库中探索知识及进行语法教学（详情可查网页 http://www. ucl. ac. uk/english-usage/ice-gbl 或电邮 icecup@ ucl. ac. uk）.

随着网上语料库的日益增多和加大，如何快捷有效的进入目的语料库作研

究也成为本次大会的另一个热门话题。来自苏黎世大学的一批年轻学者 Sebastian Hoffman，Peter Schneider 和 Hans Martin Lehmann 等设立了网上语料库检索站 Corpus Navigator，不但可以进入各种语料库文本做检索，还可以让使用者获得检索结果之外的一些信息，如说话者或作者的情况及篇章的类型等。此外，他们还设立 BNCweb 以用于 2 亿词次的英国国家语料库 BNC2.0 的在线检索。

6. 历时语料库的研究

利用历史语料库对语言做纵向研究历来是语料库研究的一个重要分支。Helsinki 历时英语语料库的创始人 Matti Rissanen 的在会上发表的论文主要是通过对不同时期的英语语料（从中古期到现代期）调查，找出了介词 notwithstanding→（in）despite（of）→in spite of→despite 的语法化演变过程，他认为当某个词的使用频率显著升高的时期往往就是语法化进程的开始。

Innsbruck 大学的 Manfred Markus 则研究 14 和 16 世纪两个不同时期的英文书信语料库中女性书信在语用和修辞方面的特征，其中包括自我崇尚的表现方式，强调语式，感情表达方式，套语形式以及通过感叹词所反映出来自我表现和祈求话语的语用标记等等。结果发现 16 世纪的书信要比 14 世纪的更广泛地使用个人指称（deixis）的表达型式，流露出个人自主意识的复苏。此外，与男性书信相比较，女性书信中有更频繁的强调语式表达，而且与读者的交流手段更为活跃，更有创意和更具合作性。

对语言历时性研究的论文还有《历时分析中介词与关系从句的统取现象》，《从 LOB 到 FLOB 看当前英国英语变化的新动向》等等。

历时 4 天的大会在语料库语言学创始人之一，瑞典 Lund 大学的 J. Svartvik 教授幽默而意味深长的诗歌中结束。 2001 年 5 月份的 ICAME 年会将在比利时 Louvain 大学召开，主题为：语料库语言学的研究工具和方法论。征文的题目指南包括：语料库建设，词法和句法附码，语料标注，口语语料库的注释，人工检索和电脑自动检索相接合的研究方法，受理论驱动和受数据驱动的联合研究模式等等。

3 口语语料库、平行语料库、学习者语料库

—— 第 23 届国际语料库语言学年会 ICAME2002 综述 *

何安平

2002 年 5 月 22－26 日，由哥德堡大学英语系主办、Karin Aijmer教授主持的第 23 届国际语料库语言学年会（ICAME2002）在瑞典哥德堡阿肯国际会议中心举行。出席大会的有来自亚、非、欧、美及大洋洲 23 个国家的 112 位学者，其中特邀专题发言人有：美国南加洲大学的 John Du Bois，挪威伯根大学的 Helge Dyvik，澳洲悉尼大学的 Michael Halliday，英国兰卡斯特大学的 Geoffrey Leech，利物浦大学的 Michael Hoey 和伯明翰大学的 John Sinclair 等著名学者。大会的主题是英语语料库的理论与应用，会上发表的 70 余篇论文中尤为突出的是对口语语料库、平行语料库和学习者语料库的建设、开发与应用。本文就这 3 个专题分别作重点介绍。

1. 口语语料库

目前在国际流通已有数百万词次的大型口语语料库，其中包括英国口语语料库 LLC，美国口语语料库 SALT，英国广播语料库 SEU，伦敦青少年口语语料库 COLT 以及英国国家语料库 BNC 和国际英语语料库 ICE 中的口语子语库等。建库的规模越来越大，然而，正如 Halliday 在题为 The spoken language corpus at the foundations of grammatics 的演讲中指出，书面语料库汇集的是人们经过思维策划的、比较符合书面语法的文字语料，但是如果光是用书面语法规则和文字注释的规范来建立口语语料库，则会扭曲了自然真实口语的语料，因为自然口语中除了字、词句以外，还包括韵律和语调、说话者边想边说的种种措辞的心理表征以及语法和语用的标记。这就对建设大型口语语料库提出了新的挑战。Du Bois 在“*What is 'natural' discourse? Implications of spoken corpus research*”的演讲中也指出，一个语言片段就是一段真实生活的再现（To take a

* 本文原载于《国外外语教学》 2003 年第 1 期，此处稍有改动。

slice of language, we must take a slice of life. ），其中有很多内容是无法用文字表现的，所以研究者一旦为了某个目标开始某项语料库研究的时候，他其实已经开始运用某种语法模式在描写语言和创建语料库。这个过程本身已经自觉不自觉地引入了人类学和社会学的研究方法，即通过观察和录音的方法努力把这种生活语篇的真实性和自然性表现出来。实现这个目的的途径之一就是建立音文同步的口语语料库。在会上，挪威伯根大学的 H. Knut 演示他们研制的音文同现的 COLT（伦敦青少年口语)和 SALT （美国口语）语料库。其方法是先将口语语料的录音制成可供电脑处理的声音文件，再用语音分析软件 PRAAT（见附录 1）一面听声音，一面观看语音分析图像（音高，长短，停顿等），一面作文字录写，之后便可进一步实现音文语料同现。

本次大会发表的关于口语语料库的研究论文就有十余篇，例如以下几篇（括号中的汉语是该论文的主要内容或特点，下同）：

- What does(n't)*or* mean?（美国口语语料库中 or 的语义和语用功能）
- Pragmatic noise: the *ers*, *ums* and *aha* of Early Modern English dialogues（英语对话中语气词的形式及语用功能）
- A corpus based analysis of the sociopragmatic distribution and use of *let's* in spoken language (let's 在口语中的分布使用及社会语用功能）
- A comparison of *cos* and *pues* in English and Spanish teenage talk（对比英国与西班牙青少年口语中的语篇结构词）
- Signalling spokenness in personal ads on the Web : the case of ESL countries in south-east Asia （从网上的“个人介绍”看东南亚各国英语说话者的口语特征）
- How to do things with prosody: requests with *please* in ICE – GB（分析口语中 please 的韵律特征与功能）
- Analysing professional oral reading on the basis of corpus data（对比职业与非职业朗读者的英语音韵特征）
- Hyperbolic expressions in speech （口语中的夸张语的类型、程度及语境特征分类）
- Evaluating a pragmatic model of speech act assignment using spoken language corpora （用口语语料库检验话语行为的语用分析模式）
- Spoken vocabulary development in a corpus of children's English（从儿童英语语料库看 6 ~ 12 岁儿童口头词汇的发展）
- New application of child language corpora （英国儿童口语语料库的词汇和语