语料库在语言研究和外语教学中的应用

孟岫岩

（黑龙江大学，哈尔滨 150080）

提　要：语料库语言学的兴起和发展，为语言研究和外语教学带来了新的方法。随着计算机技术的发展，语料库规模和应用领域也不断扩大。基于语料库的研究方法为语言研究和外语教学提供了新的视角。本文依据现有的研究成果，讨论语料库的语言研究及其在外语教学中的应用价值。

关键词：语料库语言学；语料库；语言研究；外语教学

中图分类号：H359.3　　**文献标识码：A**

引言

随着计算机技术的快速发展，依赖于计算机强大的计算和信息处理能力的语料库语言学得以诞生并蓬勃发展起来，而且展示出愈来愈强的发展潜力和发展空间。随着人们对语料库语言学在语言研究和语言教学中作用认识的不断深化，语料库语言学发挥着越来越大的作用。基于此，有些研究者甚至断言："……语料库语言学已经成为语言研究的主流（J.Thomas 等人）。"（转引朱乐红 2000）因此，我们有必要对语料库语言学作全面的了解，对它在各个领域中的应用及所发挥的作用有完整和清晰的认识，从而更好地利用它来推动语言学研究和外语教学的发展。

1 语料库

关注语言学领域研究状况的人也许会发现，语料库语言学虽是最近 40 几年才发展起来的，但它却显示出惊人的增长能力。特别是从 20 世纪 90 年代以来，语料库（corpus）规模进入了飞速增长的时期。机读语料库的建立和发展状况本身就极好地说明了这一点。首先，语料库的规模越来越大。60 年代初建立的 Brown 语料库和 LOB（Lancaster-Oslo-Bergen）语料库，规模均只有 100 万词次。BCET（Birmingham Collection of English Texts）语料库建立之初，容量为 730 万词次，很快便发展到 2000 万词次。1994 年建成的英国国立语料库 BNC（British National Corpus），收录了 1 亿词次的语料，BCET 则扩大为一个开放式英语词库（Bank of English），到 1996 年底已增至 3.2 亿词次，且其规模仍不断扩大。进入 21 世纪，语料库增长更为迅猛，WEBCORP 语料库成为通过因特网检索的在线语料库，词次达到 10~50 亿。（潘璠等 2004；何安平 2004b：3）语料库语言学的主要奠基人和倡导者里奇（Leech）（1991）估计，"如果照这样发展，2010 年前会出现 1 万亿单词规模的语料库。"（转引朱乐红，2000）其次，语料库的类型越来越多。由单纯是口头语和书面语文本汇集而成的原始语料库，发展到附有词性、语音、语义等多项特征标注的附码语料库。从只收集一种语言语料的单语语料库，发展到包含两种或多种语言在单词、短语乃至句子层面实现同步对译的平行语料库，为帮助外语学习者更好地掌握外语而建立的学习者语料库。另外，还可从语料的语体、地域、语域等多方面对语料库加以

区分。研究者还建立了符合个人研究需要的各类小型个人语料库。

语料库发展的速度如此惊人，那么，什么是语料库呢？“语料库（corpus）顾名思义就是存放语言材料的仓库（或数据库）。”（黄昌宁等 2002：1—2）语料库这个术语并不是最近40 几年才出现的。“运用语料库进行语言研究可以追溯到 19 世纪末，但当时的研究手段还只停留在卡片制作和人工检索的阶段，其成果也仅用作编纂语法书或词典的参考。”（何安平 2004a：1—2）计算机技术的发展和应用，使语料库从传统的纸质版本发展成为机读电脑版本。因而，“近 40 年以来，语料库这个术语常指以电子形式保存的语言材料，并被广泛应用于语言研究和语言工程。”（黄昌宁等 2002：2）语料库的研究方法、研究对象和应用领域使其成为一门独立的学科：语料库在语言学。一般认为，“语料库语言学是在文本语料库的基础上进行的语言研究，”，“以语料为语言描写的起点或以语料为验证有关语言的假说的方法称为语料库语言学。”（朱乐红，2000）其实上述两个定义中，尽管对语料库语言学的研究基础、研究内容的界定大同小异，但从本质上来讲，两个定义是不同的。很显然，前者认为语料库语言学是一门语言研究科学，而后者则肯定语料库语言学只是一种研究语言的新方法。我国语料库语言学的研究者分别支持这两种不同的观点。持前种观点的有何安平、潘璠等。他们认为语料库语言学是“可与多门学科相交叉，为多门学科的语言研究提供研究工具的语言学分支。”（潘璠等 2004）而支持后种观点的研究者有朱乐红、黄昌宁、许家金等，他们则认为：“现代语料库语言学不是研究语言自身某个方面，而是一种以文本语料库为基础的语言研究方法。”（朱乐红 2000）；“基于语料库的研究方法（corpus-based approach）这一提法倒是更能准确地反映语料库语言学的性质和定位。”（许家金 2003）

无论对语料库语言学的定义如何，有一点至关重要的，即随着语料库语言学的兴起和飞速发展，它为语言学领域的研究提供了一种新的工具——机读语料库和一种全新的基于语料库的研究方法。语料库之所以能成为人们进行语言研究和语言教学的有用工具，是因为语料库内“所采用的不是通过臆想或杜撰的语言，也不是仅仅从某些著作抄录下来的若干例句，而是有目的、有系统地收集的大量在现实生活中使用的书面语和口语；并使用先进的电脑科技手段进行储存和检索，从而将语言研究建立在一个博大的、真实可靠的语料基础上。”（何安平 2004a：2）语料库是一个文本的集合，它建立的目的是“成为整个语言或在某一个特定时期语言的一个代表。”（黄昌宁等 2002：23）这就是语料库的代表性问题，也即是建立语料库的目的。只有保证所采集的语料具有最大限度的代表性，才能反映出具体语言在言语活动中的原貌，体现其使用中的特点和规律性。因而，语料库建立的原则之一便是取样范围要广，尽可能地涵盖具体语言在各个领域内应用的文本，尽可能多地考虑多种影响语言的因素和变量。我们能以早期建成的三大语料库对此问题进行说明。20 世纪 60 年代和 70 年代建立的两个语料库 BROWN 和 LOB 各采集了美国英语和英国英语 100 万个单词，其内容涵盖了新闻、宗教、技术、娱乐、小说、信函等 15 种书面语体。而在 80 年代建成的 London-Lund of Spoken English（LLC） 则是第一部英语口语语料库，汇集了 50 万词次，文本包括正式或非正式的个人演说、评论、双人或多人会话、讨论、采访和电话录音等口语语料。（何安平 2004a：2）随着语料库语言学的发展，对影响语言的因素考虑得越来越多，也更为细致，甚至“连说话者和听者之间的身体距离和社会地位差异都纳入考虑范围。”（潘璠等 2004）文本选择范围覆盖越大，就越能接近具体语言在言语活动中的真实状态，而且语料库中的语料（书面语和口语）都是在现实中可能的而非人工诱导的语言，这意味着依赖语料对语言进行研究更有可能反映语言在实际应用中的特点和规律，从而保证基于语料库导出的对语言的描述更为客观和科学。因而，里奇认为：“语料库研究方法是一种更为强有力的方法，因为其结果是可以验证的。”（转引朱乐红 2000）

语料库规模的飞速发展，语料库能与语言学领域的研究和实践紧密结合，这一切都与语料库自身的特点密切相关。语料库的特点决定了它能有广阔的应用空间，其特点主要体现在

真实性、代表性、灵活性、价值性。

语料的真实性是语料库的重要特征之一。语料库就是建立在真实可靠的语料基础上的。语料库采集的是现实生活中使用过的口头语或书面语文本，是鲜活生动的言语活动的实例，是人们在交际中应用的真实话语。语料的真实就在于文本不是单凭感觉或规则而主观臆造或杜撰出来的，而是来源于人们真实的言语活动。正是语料的真实性才能保证我们依赖语料而得出的语言规律具有科学性，并能对我们所归纳的语言规则进行验证。语料库的代表性既是建立语料库的目的，也是语料库的又一重要特征。如果语料库只具备真实性，而无代表性特征的话，那语料库的存在也就不会有太大的意义。代表性是一个相对的问题。语料虽然取自现实生活，但我们却无法涵盖人类浩如烟海的全部言语事实。因而，我们只能以某一时期、某一方面为重点尽可能多的选取语料，力图以有限的语料反映具体语言的特点和使用特征。这也是语料库规模不断增长的主要原因。语料库在多大程度上具有代表性，决定了研究者依赖语料库对语言描写科学性的程度。运用灵活亦是机读语料库的重要特征。运用计算机强大的处理能力和语料库检索工具，使用者可以在很短的时间内对几百万、上千万词次的文本内容进行检索、统计分析，这大大节省了使用者的时间和精力，提高了工作效率。与以前的人工语料库相比，机读语料库一旦建成，还具有可反复使用的优点，并且借助计算机技术还能达到资源共享。如果语料库只是原始文本的简单集合，其功能便十分有限，它的使用价值也不会很大。语料库要具有应用价值，必须对文本进行各层面的分析，并将所得结果标注到语料上去。语料库标注是“一种给口语和（或）书面语语料库增添解释的（interpretative）和语言的（linguistic）信息的实践。‘标注’也可以指这个过程的最终产品：即附加或分散在语料库中的语言标记。”（黄昌宁等 2002：139—140）一个语料库的标注越多，它的价值就越大，其应用也会更加广泛。目前的语料标注主要有词性标注、语音标注、语义标注、句法标注和学习者拼写、语法差错标注等等。增加标注的同时，还要注意降低标注的错误率，这样才可为语料库添加价值，更好地发挥语料库的优越性。

2 语料库的应用领域

鉴于语料库的这些主要特点，基于语料库的研究方法在语言学领域已得到了越来越多的应用。总体上来讲，语料库在语言研究和语言教学，尤其是外语教学两方面所发挥的作用较为显著。下面我们就基于语料库的方法在这两方面的应用分别加以论述。

2.1 语料库与语言研究

语料库是由来自真实语言的大量文本构成的，丰富的语言材料可以为词典编纂提供充足的例证。引述实例是词典编纂的主要方法，计算机语料库的出现彻底改变了利用语料的方式。这一工具问世之前，语料的收集、制作、整理和索引都处于人工完成阶段。这使语料在数量、规模和代表性方面都受到很大的限制，而且利用起来也费时费力，不能重复使用。例如，“1890 年德国的 Kaeding 为了收集德语的字母和单词频率信息，以发展德语速记模式，曾动用了 5000 多名助手花了数年时间处理了 1100 多万词次的口、笔头语料。”（何安平，2004a：2）而现代词典编纂学家只需坐在计算机前就能无遗漏地将指定的词、短语或句子从上千万甚至上亿的词次中快速检索出来。检索的结果可以直接显示在屏幕上，甚至更进一步，研究者可以用多种方式对索引的结果进行排序。采用这种基于语料库的方法，不仅意味着词典的编纂和修订速度能大大加快，而且词语的释义排列顺序由语料库中得出的统计结果决定，从而使结论更加客观，科学性更强。

利用语料库可以对语法进行多层面的分析和研究。语音方面能够依赖口语语料对语音和语调进行研究；词汇方面能够对词类、词出现的频率、语境、邻近搭配等方面进行研究；句法层面能够对句子结构、各类句型的使用情况进行分析；篇章层面可以对语篇的连贯与衔接、各类篇章的语言特点进行研究。“在 20 世纪 80 年代之前，语法的经验研究不得不主要依靠

定性分析。这类研究可以提供语法的详细描写，但很大程度上难以超越频度和罕见度的主观臆断。”（黄昌宁等 2002：155—156）语料库为语法研究提供了定量分析的工具，但我们认为，语言分析不应完全依赖于量化分析。“Leech 认为语料库的应用应看作是语料加直觉的问题，而不是语料或直觉的问题”。（许葵花等 2003）因而，运用语料库时我们应注意采用定量定性分析相结合的方法，这样才能保证分析的结论具有客观性、科学性，这样的分析结果才能作为编写语法参考书的依据。“Biber 等 5 位学者编著的《朗文英语口语与笔语语法》便是其中的一个典范。基于语料库的语法书的最大特色是重视频率、搭配和词组合、语言变异、语法中的词汇、语言材料的真实性。”（胡春雨 2004）

语料库在语义研究方面的重要作用在于“它可以为词项赋义提供客观的准则。”（黄昌宁等，2002：157）一个词究竟是用作何种意义，常常与词语的搭配、句法结构、乃至语篇密切相关，而语料库恰恰为词的语义研究提供了客观依据。如以 certain 一词为例，通过检索 LOB 语料库发现，“‘人称词+be certain’的结构，在小说中远比社科文本中普遍，其含义是‘确信’、‘确切’等，其他部分结构，如 a certain、in certain 和 of certain 则在社科类语篇中较多，其含义常常是‘不确切’。”（朱乐红，2000）。

对语料库在语用研究中的作用，研究者持不同的观点。一种观点认为，“通过语料库对语言使用和语言交流进行研究，可以发现实际交际中用词、表示方法、习惯用法等规律。”（朱乐红，2000）掌握语言在实际运用中的特点和规律，弄清词语的常用搭配、习惯用法和使用频度，可以对我们的研究给予指导。而另一种观点则认为，语用方面基于语料库方法的研究很少。语料库采集文本时对其长度有所限制，因而文本方面的语境被删除了，而语用学中对意义的研究则主要依赖于语境。（黄昌宁，2002：158）

2.2 语料库与外语教学

正如杨惠中教授在 2003 上海语料库语言学国际会议开幕式的发言中所指出的，我国的语料库语言学研究从 20 世纪 80 年代中期第一个语料库起，就与外语教学结下了不解之缘。外语教学始终面临着两个问题：教什么，怎么教。传统的外语教学采取的是以教师为中心的灌输式的教授方式，完全忽视了学生的学习积极性、主动性。再有，教材的编写过多地依赖于语言直觉，忽视了对言语活动中鲜活生动的言语的描写。许多研究表明，外语教材中的内容与本族语者在实际中使用的语言之间常常存在着很大的差别。研究者对口语中表示将来时态的 will 和 be going to 作过词频统计。统计结果表明， will 的使用频率远远大于 be going to，而恰恰与此相反，我国某些教材却将二者学习顺序颠倒了。“例如北京出版社 2000 年出版的九年制义务教育小学《英语》中，be going to 出现在第 4 册第 7 单元，will 出现在第 5 册第 6 单元。学生学习、运用 be going to 的频率远远大于 will。”（许葵花等，2003）基于传统或直觉编写教材，虽然注意到了语法结构的系统性和完整性，但结果有时并不符合语言事实，造成对教学的误导。而计算机语料库恰恰弥补了上述不足。语料库不仅为我们提供了丰富的真实语料，而且随着它功能的不断完备，研究者可对其进行统计和分析，从而为外语教学大纲、教材内容编排、教学重点和教授顺序的选择提供科学依据。

语言是一个庞杂的、不断发展变化的系统。如果将陈旧的语料或过时的模式搬到外语教学中，肯定不会达到学以致用的目的，从而误导学生。而语料库恰恰很好地解决了这一难题。语料库可作为一种资源在以下几个方面可提供适时帮助：1）语料库仅呈现真实的语例，因为凭空臆造的语例很容易误导学生；2）语料库可以使人们认识自己的语言本能，即清楚地看到自己平时不经意但却是经常使用的语言形式；3）语料库可以提供词语使用的语境；4）语料库为语法提供语义基础。通常一个具有两种意思的词也意味着它有两种语法结构；5）语料库提供大量语言使用的变体以及新鲜而有创意的形式，体现出语言的最新发展和变化。（何安平 2004：46—47）语料库中的语料来自交际现实中的自然语言，可以向我们展示语言在实际中的使用情况，从而使学习者能够准确地掌握语言，达到学以致用的目的。

语料库可以为教学提供科学依据，借助语料库可以发现语料库发现语言特征、规律，在此基础上，进行统计和分析，可以为教学大纲、教材内容的选择、编排顺序、教学重点提供科学的依据，作为编写教材的参考方面。词频统计是计算机语料能够提供的最基本的功能之一。研究发现，“词语在语言中的分布是极其不平衡的，为数不多的高频词（常用词）占据了语言的主体，而数目庞大的低频词只构成语言的一小部分。”（潘璠等 2004）研究者对 730 万词次的 BCET 语料库的词频情况作了统计，结果如下：

总词数	132452
只出现 1 次的词	68302
出现 2-9 次的词	42255
出现 10-19 次的词	8152
出现 20-29 次的词	3450
出现 30-39 次的词	1839
出现 40-49 次的词	1251

我们可以发现，在这个规模不算小的语料库中，出现 20 次以上的词语只有 6000 多个。由此可见，利用此种分析方法可以帮助我们确定某种语言的常见词语，以此为基本词汇来编排教学内容，定能节省学习的时间、提高学习效率，从而取得事半功倍的效果。

许多文献（陈建生 2004；胡春雨 2004；许葵花等 2003）中都提及一种把语料库用于课堂教学的主动学习方法，称之为数据驱动学习（data driven learning），简称 DDL。DDL 的深层理据是：“卓有成效的语言学习实际上是一个探索语言的过程，而语境共现则为激活归纳式的学习策略提供了前所未有的条件。”（何安平 2004a：58）DDL 能激发学生自主学习的动机，使他们能带着问题到真实语料中找寻答案，学生们更容易记牢通过自己努力发现的知识。尽管从理论上看，DDL 是一种“发现式学习”的方法，但在实际运用中，目前尚存在许多问题。最突出的问题是：非常耗时。无论对教师还是学生来说，在课堂教学实践中运用 DDL 方法都需要花费大量的时间。首先，教师编写 DDL 课堂教材或练习时相当费时。其次，对学生来说，既要学会使用语料库检索软件查找所需的语料数据，同时又要掌握基本的分析数据的方法，这无疑会加重学生的学习负担。鉴于此，目前情况下，我们尚不能提倡在教学中普及这一方法，而只能建议采用这一方法时，要有针对性。一是针对特定的学习内容，将我们的重点放在较重要的语言问题上，力图在有限的时间内取得更为显著的学习效果；二是针对特定的学习者群体。DDL 方法并不适宜所有的外语学习者，对外语初学者和非外语专业的学生运用 DDL 方法，似乎有些强人所难了。

3 结束语

尽管语料库在语言研究和外语教学中的应用价值，但还应在肯定语料库积极作用的同时，对语料库的局限性有所认识。语料库是特定语言文本的集合，是这类文本的代表，但语料库究竟在何种程度上具有代表性，还需实践证明。尽管语料库在建立时考虑到了对语言产生影响的诸多因素和变量，但言语事实浩如烟海，即使是规模再大的语料库也无法对其完全涵盖。因而，“只能在一种非常有限的意义上才可以说语料库是‘真实的语言’。”（胡春雨 2004）这样，我们对基于语料库统计分析出的语言规律难免不会产生一丝怀疑。再有，选择语料库文本时对其长度的限制，在一定程度上也使语言脱离了语境。此外，基于语料库的教学方法尚存在许多难题。一方面，这种新的教学方法是否已得到教师们的普遍认同，已经习惯了传统课堂教学的教师，是否愿意接受这种新方法、新工具。另一方面，愿意利用这一工具的教师或学生，是否具备了必需的设备和条件。毕竟，目前在我国各大高校内，一人一机的教学条件尚未完全实现。这些问题若不解决，还不能提到在教学中普及基于语料库的教学方法。我们认为，对于语料库，我们在充分认识到它在语言研究和外语教学中所产生的深刻

影响的同时，还应对之“采取审慎的态度，警惕语料数据可能会误导我们做出错误的结论。”（胡春雨 2004）我们应明智地利用语料库，既不能对它过分依赖，也不能处之漠然，善于利用好这一新工具、新方法促进语言研究和外语教学的发展。

参考文献

[1]陈建生 2004 语料库语言学与英语教学[J]，解放军外国语学院学报，第 1 期。
[2]何安平 2004a 语料库语言学与英语教学[M]，北京：外语教学与研究出版社。
[3]何安平 2004b 语料库在外语教育中的应用：理论与实践[C]，广州：广东高等教育出版社。
[4]胡春雨 2004《语料库与应用语言学》评介 [J]，现代外语，第 3 期。
[5]黄昌宁 李涓子 2002 语料库语言学[M]，北京：商务印书馆。
[6]李赛红 2002 解构英国国家语料库[J]，外语教学与研究，第 4 期。
[7]李文中 濮建忠 卫乃兴 2004 上海语料库语言学国际会议述评[J] ，解放军外国语学院学报，第 1 期。
[8]潘璠 冯跃进 2004 语料库规模增长原因探查[J]，外语学刊，第 3 期。
[9]许家金 2003 语料库语言学的理论解析[J]，外语教学，第 6 期。
[10]许葵花 张卫平 2003 论语料库语言学在外语教学中的应用[J]，外语与外语教学，第 4 期。
[11]徐曼菲 何安平 2004 图式理论、语料库语言学与外语教学[J]，解放军外国语学院学报，第 6 期。
[12]邢富坤 2006 语料库：值得教育技术学关注的新型学习资源[J]，解放军外国语学院学报，第 2 期。
[13]朱乐红 2000 语料库对语言研究及外语教学的作用[J]，外语与外语教学，第 3 期。

The Application of Corpus in Language Study and Foreign Language Teaching

MENG Xiu-yan

（Heilongjiang University, Harbin, 150080, China）

Abstract: The development of corpus linguistics brings forth new methods for the study of languages and foreign language teaching. With the rapid development of computer science and technology, corpus has been expanded in its scale and field of application. The research method with corpus as its basis provides a new angle of view for language study and foreign language teaching. Based on the exiting fruits of relevant scientific research, the paper discusses the application of corpus in language study and foreign language teaching.

Key word: corpus linguistics; corpus; language study; FLT (foreign language teaching)

收稿日期：2007-03-07

作者简介：孟岫岩（1971－），女，黑龙江哈尔滨人，黑龙江大学俄语学院 2005 级硕士研究生。主要研究主向：俄语语义学。

[责任编辑：薛恩奎]