中图分类号:G434 文献标识码:A 文章编号:1001-5795(2003)03-0027-0004

论个人教学语料库的构建

谢家成
(荆州师范学院 英语系,湖北荆州 434100)

摘 要:本文通过对语料库在外语教学中现状的介绍,论证了构建个人教学语料库的必要性和可行性,并结合教学实践需要,对个人教学语料库建设提出了一些建议。

关键词:个人教学语料库;检索软件;标注

On Creation of Personal Teaching Corpus

XIE Jia-cheng
(English Department of Jingzhou Normal University, Jingzhou, Hubei 434100, China)

Abstract: In this paper, the author justifies the necessity and feasibility of the creation of one's personal teaching corpus through the brief account of the current situation of corpus application in foreign language teaching. And the paper then goes on to offer some useful suggestions for the creation based on the real school teaching circumstances.

Key words: Individual Teaching Corpus; Concordancing Program; Annotation

语料库现在一般是指存放在计算机里的大量语言材料和定位检索管理软件的结合。语料库语言学近20年来取得了巨大发展。语料库功能十分强大,最突出的特点是能快速而准确地提供与一个或多个关键词有关的大量真实语料,并以KWIC(Key Words in Context)或其它形式灵活地展现出来,从而生动地揭示其丰富的"语言生态"(linguistic ecology),包括构词、表达结构、搭配、语境乃至话题、修辞等多方面的信息,对日常外语教学十分有用,甚至能改变教学思维和模式,然而它在日常外语教学中的运用正如Graeme Kennedy(2000)指出的那样"一直是语料库语言学的薄弱环节"。

在中国情况更是如此,至今仍有一些英语专业研究生课程未包含语料库方面的内容,本科阶段则几乎空白,语料库的普及应用任重而道远。相关文章对其介绍——无论语料库专著或外语期刊论文,也有一定局限性。以清华同方中文期刊全文数据库①检索的相关论文为例,约110篇论文中大都侧重于介绍国内外大型公开语料库或利用它们进行的研究。这给读者造成一种误解,似乎语料库主要是词典、词汇表等编纂的工具或者是少数研究者进行语言研究的工具,与广大教师的日常英语教学关系不大。

另一方面,针对其在教学中运用的文章语料来源大都局限于国外大型公开网上语料库(如Bank of English②、British National Corpus - BNC③,可免费试用)或光盘语料库(如LOB、Cobuild),这些语料库对于英语教学有较大局限性。比如免费网上语料库检索例句数目和长度均有限制(BNC检索条目为50条,Bank of

作者简介:谢家成(1969-)男,汉,讲师。研究方向:电化教学、语料库。
收稿日期:2003-01-20

English 每个例句长度为 80 个字符），由于语境不全，不符合“可理解输入”原则，只能使用部分语料，且对这部分语料也无法灵活选择，因此无法选择特定语域的语料进行检索。当然，最根本的矛盾是这些通用语料库建设的主要目的是词典等工具书的编纂和语言调查研究而非课堂教学，对于特点各异的教学对象难免缺乏针对性，比如适合初级学习者的语料就很少。光盘语料库也有类似缺陷，且语料一般无法更新，时效性差。再加上价格、市场等因素，这些语料库很难获得，因而不利于语料库的推广使用。

要想真正推动语料库在外语教学中的普及，笔者认为有必要建设和使用个人教学语料库。个人教学语料库能克服以上种种缺陷，并有着自己巨大的优势。突出表现如下：

（1）目标明确，语料收集针对性强，适合自己教学对象的难度和兴趣。

（2）语料能不断扩展，及时更新，具有开放性和时效性，语料反映时代、贴近生活。

（3）本地机操作，经济、方便、快捷，检索语料时选择自由，易于突出语言的语域特征。

（4）自主设计，灵活标注，功能更多，能提供更大语境，可同时发挥文本库的功能。

1 可行性

人们一般认为构建语料库是一个花费大量人力物力的工程，实际上现在个人建设教学语料库是完全可行的。语料库主要由两部分组成：以电子文本形式存贮于计算机中的语料和管理检索这些语料的定位检索软件。随着个人计算机性能的不断提升和价格的不断下降、网上和光盘电子文本的日益丰富、OCR 的广泛使用等，个人收集大量真实语料不再是困难的事情。另一方面许多功能强大、使用简便（不需专门计算机知识）的定位检索软件网上也能免费下载。再加上个人教学语料库规模灵活，可边建设边使用，建库要求也不是很严格，这一切使得个人教学语料库的建设成为现实。表 1 是部分网上免费检索软件功能及使用情况一览表。

2 个人教学语料库的构建

建立语料库，规划很重要，个人教学语料库也不例外。建立语料库要考虑的因素有建库目的、语料品种、取样标准、规模、代表性、设备、存贮方式与格式等（Graeme，2000）。其中建库目标最为重要，直接影响语料库类型和规模，并最终决定语料的选择和整理。“个人”和“教学”决定了语料库类型为兼收并蓄的全文语料库而不是抽样语料库，语料应尽可能包含与教学相关的语域。个人教学语料库虽然也需一定规模，但更应明白它是一个不断扩展更新的过程，是一定质的基础上的量。盲目追求语料数量不利于充分展示语言的特点，不利于检索的后期选择工作，因此应特别注意语料的选择、分类和整理工作。

2.1 语料的选择

为了保证语料的真实性，语料主要应从网上或光盘选取。网上语料十分丰富，不仅有大量网上报纸、杂志、书刊等电子媒体，还有一些专门的英语电子文本库（如 Gutenberg⑦，Alex⑧⑫等，内容从文学、哲学、宗教、法庭审讯、名言智语到菜谱等应有尽有）、数字图书馆、歌曲库，电演剧本库等。另外依靠关键词，如 online text，corpora，script/ transcript 等，也能搜索到大量所需文本。由于定位检索软件支持汉语显示或检索，收集英汉对照语料也能部分起到平行语料库（parallel corpora）的作用（丁树德，2001）。

语料筛选应有针对性，应考虑教学对象的英语水平、专业、兴趣、需求等，语料难度应适中，符合“可理解输入”或“i + 1”原则（Krashen 1985）。语料还应兼顾知识性，时效性，教育性。知识性指语料的信息量；时效性指语料应不断更新，贴近时代；教育性指语料能带给学生思想上的启迪，如收集大量 Quotation、Proverb、

表 1

软件名称	作者 版本	共同功能	使用评价
Microconcord④	Mike scott & Tim Johns，版本 1.0	多文件、多关键词同时检索；支持通配符、指定跨度检索、KWIC 灵活排序显示、更多上下文、检索结果多格式保存或打印	速度快，操作简单，英汉混合显示，但汉语不能作关键词检索
Concapp⑤	Chris Geaves，版本 6.0.1		速度快，功能强大，英汉均能作关键词检索
Concordance⑥	R. J. C. Watt，版本 3.0		检索前文本需预处理，速度慢，功能强大，支持表达式检索，自定义 Reference 功能，词表倒序排列，有 Web 功能。

注：以上检索软件各有特点，在实践中可根据不同情况灵活选用。

Witty statement 等语料，可以教书育人，相得益彰。

语料选择还应注意语料平衡性，突出教学特点。文学材料容易获得，但太正式，不宜过多（约翰·辛克莱著，王建新译 2000）。现代教学强调语言材料应反映其在实际生活中的使用特点。网上报纸和杂志，如 21 世纪报，因时效性强，贴近时代和学生生活，在语料中应占一定比例，这些语料还能克服教材信息滞后的缺限。口语体和实用性语料（如景点介绍，礼仪辞令、产品使用说明等）的收集也是必要而且可行的。网上也是口语语料收集的好去处，如 CHILDES⑨提供的儿童口语和成人口语（温志军 2001）（注：其初旨为提供儿童语言习得，现在其 contributing 部分也提供大量成人对话文本、声音素材）、CNN⑩的 transcripts 部分提供的时事评论，访谈等、还有各种电影电视脚本⑪和演说等。考虑到教学的特殊性，可适当收集一些写作范文、考试试题，条件许可也可将所用英语教材扫描录入（旺旺英语⑫之资料共享区提供许多教材文本），这样可以轻松获得任一词或表达式在教材中的所有出现语境，对师生都十分实用。利用 Concordance 的 WEB 功能还能将其制成网页，供学生通过校园网自主检索（具体用法请参见软件使用说明）。

2.2 语料的分类

语料收集并不是语言材料的简单堆砌，从一开始就应按一定原则（语域、语体、来源、难度等）进行科学分类。分类可适当参考一些大型语料库分类标准，但个人教学语料库有自己的特点，语料分类应以“我”为主，考虑实际教学需要。比如笔者就先将语料分为书面语和口语两大类，然后再以“语域为主，来源、时间兼顾”的原则进行细分。由于定位检索软件支持多达数百文件同时检索，语料分类越细越好。比如新闻就可先分为时事、财经、体育、热门话题、科技、娱乐等，再根据新闻来源如 China Daily 、21 世纪报、CNN 等细分。这样既可区分难度，显示语料来源和语域，又能方便相同新闻事件的用语对比。

分类整理后的语料应选择“纯文本”格式保存，文件名也应保持统一，并尽可能多地体现分类信息（通过重命名，扩展名也可充分利用）。比如笔者对新闻语料命名统一为“news + 类别（如体育）+ time + 来源（如 CHD）”。这样便于多个同类文本快速选择，如通过“newssports * ” 或“ * . CHD”（“ * ”为通配符）可分别快速选中所有体育新闻或 China Daily 新闻语料。

2.3 语料标注

语料整理另一项重要工作是做标注。所谓标注就是把语料的有用信息用一组符号一一标注出来。本文所指的标注不是人们通常所指的词性标注（tagging）和句法标注（parsing），而是根据教学需要并结合软件特点，对部分语篇所做的关于背景信息、主题或语篇类型、修辞、功能、写作等多方面的创造性标注。

笔者在实践中摸索出一套有效的标注方法，这种方法分为两种。一种类似 COCOA 模式，在语篇篇头对有关语篇背景信息，如作者、标题、日期、语篇来源、语篇类型、文体等进行标注。篇头标注由两部分组成：其一是代表语言特征名称的附码（如用 C 代表“类别”），其二是具有该特征的语言单位，两个部分放置在中括号内。这样一个语篇的语域信息就可以标注为类似 < C GOLF > 。这种标注符合 Concordance 软件自定义 reference 标准，其中语域信息标注十分实用，这样不仅能检索到所有相关语域的语篇，而且使用其自定义 reference 功能还能生成任一特定语域的词频统计表，比如听力课便可利用做了诸如股票，网球等语域标注的各类新闻分别生成各自语域的词频统计表（wordlist），使学生能迅速掌握该语域的常用词汇。

另一种标注则更灵活。只需在某一关键词或添加的关键词前加上诸如“ ~ ，# ”等能与其它文字明显区别开来的特殊符号，以此作为关键词检索，如“ ~ success”，就能与庞大语料中其它未加标注的“success”区分开来，仅显示前者。不过应避免使用检索语句可能使用的“ * ，@ ，、，/ ， + ”等符号。利用这种方法可对部分语篇，在适当地方做主题、功能等多方面的标注。

主题（topic）标注是将语篇所包含的一个或多个与教学可能相关的主题进行标注，如“ ~ happiness， ~ 校园生活， ~ gene”等，适合作主题标注的语篇一般不宜过长，网上新闻、杂志有许多这样的小品文或科普文章。主题标注符合整体教学法原则，有助于拓展课文主题（话题），扩大知识面，充分发挥语料库作为一个大的文本库的功能，使教师真正成为信息的主人。另外，对口语类素材，如大量电影脚本，也可做诸如“安慰”、“留有余地”、“停顿语”等语篇功能标注。这样能快速查找相关功能的多种表达方式，且语境丰富，语言地道，大大优于一般口语教材。当然对特殊语言现象、修辞、语法、写作（如神态、景物描写等）、文化（如颜色词，动物寓意等）等各方面均可作类似标注。可见语料标注非常灵活，几乎不怕做不到，就怕想不到或不想做。

标注应有系统性，应注意对上层概念的标注。比如一篇关于生物芯片的文章便可标注为：~ new tech，

~bio，~chip 等，这样检索上层概念如 ~new tech 时，能清楚地展示所有有关新科技的文本。标注还应多层面，这也是语料库发展的方向。比如一篇幽默故事，四个神职人员的母亲在谈论自己的儿子时，前三位分别吹嘘说自己的儿子是牧师、主教和红衣主教，深受尊敬。第四位说她的儿子身材特别高大，每次一出现，人们便会惊呼“Oh，my God”，可分别作如下标注 ~religion，~clergy，~my god，~god，~bragging 等。另外，对于不同类型的标记最好采取一定方法区分开来，如 ~tchip，~demotion 分别表示话题/芯片，描写/情绪。这样通过“~d*”作关键词便可显示所有有关描写的标注。

经过以上收集、分类和整理后的文本存放在计算机的某一目录下，就构成了个人教学语料库所需的语料，再通过前面介绍的定位检索软件，便可开始使用自己构建的教学语料库了。收集、分类、整理，特别是“标注”工作，需投入一定的时间和精力，但这些努力是值得的。标注越多，越能发挥个人语料库的功能，况且这本身也是一个不断学习、发现和研究的过程。

3 结语

构建和使用个人教学语料库目前还是一个少有人问津，但却急待探索的领域。个人教学语料库的建设不仅必要，而且简单易行。当然这并非一定要“个人”独干，可以小组协作，只要目标明确，规范统一即可。语料库只是一个工具，并不能代替使用者思维（tools don't do the thinking）。对语料的收集、整理和使用都要求使用者具备语言学、外语教学等多方面的知识。教师还应更新教学观念，树立实证思想，对语言在实际生活中的使用频率，语域特征，搭配规律等应经常求助语料库，做到心中有丘壑。总之，笔者希望本文能激起广大外语教师对语料库在语言教学中的兴趣，并真正行动起来，动手构建自己的教学语料库，这必将大大促进语料库在外语教学中的推广普及。

附录：以下为文中部分内容网址，http:// 略，限于篇幅，其它网站请通过搜索引擎查找

1. ① 清华同方期刊全文数据库 www.cnki.net

2. ②③⑦⑧⑨⑫参考 Barlow 主页相关链接 www.ruf.rice.edu/~barlow/corpus.html

3. ④⑤⑥均可从 The Virtual CALL 的 multilingual 下的 textual analysis 获得：www.sussex.ac.uk/langc/CALL.html

4. ⑩CNN 节目文字录稿 www.cnn.com/TRANSCRIPTS/

5. ⑪电影电视脚本库 www.script-o-rama.com

6. ⑫ www.wwenglish.org/bbs □

参 考 文 献

[1] Graeme kennedy. 语料库语言学入门[M]. 外语教学与研究出版社，2000.

[2] Krashen，S. D. The Input Hypothesis：Issues and Implications[M]. NY：Longman，Inc. 1985.

[3] 丁树德. 浅谈西方翻译语料库研究[J]. 外国语，2001，5.

[4] 约翰·辛克莱著，王建华译. 关于语料库的建立[J]. 语言文字应用，2000，2.

[5] 温志军，胡瑰玲. 开发利用世界上最大的儿童语料库——CHILDES[J]. 外语教学与研究，2001，9.

[6] 施 敏. 英语语料库、教学方法和教师角色的转变[J]. 宁波大学学报教育科学版，2000，8.

（上接第 21 页）

[11] 何克抗. 建构主义——革新传统教学的理论基础[J]. 电化教育研究，1997，3.

[12] 张正东. 外语教学技巧新论[M]. 北京：科学出版社，1999.1.

[13] 杨 梅. 阅读理解·翻译[M]. 武汉：武汉测绘科技大学出版社，2000.10.

[14] 王道俊，王汉澜. 教育学[M]. 人民教育出版社，1989.12.

[15] 章志光. 心理学[M]. 人民教育出版社，1992.10.

[16] 杭宝桐. 中学英语教学法[M]. 上海：华东师范大学出版社，1993.3.

[17] 林君芬，余胜泉. 关与我国网络课程现状与问题的思考[Z]. http://www.vschool.net.cn/kecrz/k0003.htm. 2002.6.