[文章编号] 1003- 1553(2002)10- 0054- 04

基于网络的语料库及其在英语教学中的应用

武和平, 王秀秀
(西北师范大学 外语学院, 甘肃 兰州 730070)

[摘要] 本文概述了语料库的概念及其发展历程, 着重讨论了基于网络的语料库在英语语法教学、词汇教学、错误分析及培养学生自主语言能力中的作用。语料库不仅在技术层面上丰富了外语教学的手段, 而且使原来停留在理论和空想中的教学理念变为现实。

[关键词] 语料; 语料库; 英语教学
[中图分类号] G434 [文献标识码] A

一、语料、语料库及其历史发展

语料(corpus), 又称为语言素材, 是自然发生的语言材料(包括书面语与口语)的集合。从语料出发, 让语料说话, 是语言研究和语言教学的良好传统。历史上,《牛津英语词典》的编纂者 Murray 及《现代英语语法》的作者著名语言学家 Jesperson 都曾以很原始的方法认真积累真实的语言素材, 并且以这些素材为基础来发现规律和解释语言现象。这种语料无形中增强了著作及词典的权威性和可靠性。

语料库, 顾名思义, 就是按照特定的目的与方式建立起来的存储语言材料的"仓库"。在计算机出现以前, 语料库的建设是一项十分繁重而成效甚微的工作。世界上第一部英语词典的编纂者 Samuel Johnson 花了 8 年时间(1747～1755)才编就了这部词典, 其中大量的时间就消耗在寻找、整理语料与建立语料库。在上个世纪初, 强调运用语料来分析和描写语言的美国结构主义语言学的先驱者们, 如 Sapir 和 Baos 等, 不得不走访一些印第安部落来搜集语料。但是, 到了上个世纪 50 年代后期, 随着 Chomsky 生成语法学派的兴起,"基于语料"(Corpus- based)的描写语言学受到抨击。因为 Chomsky 认为语料只不过是对语言行为(performance)的取样, 无法客观地反映存在于人的大脑中的语言能力(competence), 所以不足以成为我们进行语言研究的依据和材料。到了上世纪 60 年代初, 人们对真实的语言素材的兴趣降到了最低点。而基于语言学家的语感和直觉所得到的"可能的句子"及"不可能的句子"成了语言研究及语法教学的主要数据来源。

尽管如此, 新一代语料语言学的奠基人不畏权威, 开始了自己开创性的工作。1959 年, R. Quirk 宣布要搜集大量不同文体的英语素材, 建立英国英语口语和书面语语料库, 即 SEU 语料库(The Survey of English Usage Corpus), 作为系统描述英语口语和书面语的根据。并在此基础上编就了著名的《当代英语语法》(1972)和《英语语法大全》(1985), 至今仍是英语语法的经典著作。1961 年, Francis 和 Kucera 在美国 Brown 大学建立起 Brown 语料库, 其中的语料语篇取自 1961 年美国英语出版物, 字数超过 100 万, 是世界上第一个机器可读语料库。

进入上世纪 80 年代以后, 随着计算机技术的发展与普及, 语料库的建设进入了一个空前发展的时期, 许多新的语料库相继建成, 对语料的处理也由较为简单的机器可读形式发展到人工或自动词形和句法分析的注释形式。计算机程序和软件的不断开发应用加快了语料库的建设, 提高了语料的处理能力和层次, 经过加注的语料又促进了语料研究和利用, 研究的深入转而又导致了更为先进的研究方法和研究模式的诞生, 许多先前需要手工处理的工作现在可以通过计算机程序及软件自动或半自动地完成。

同时, 经过 20 年的实践验证, Chomsky 生成语法学派对语料方法的责难被证明是片面的、错误的。对于 Chomsky 所倡导的唯理主义方法, 人们在跟从、应用和反思之后, 也逐渐发现其不足, 例如语言直觉的不可验证性等。因此, 80 年代以来语料库语言学的复兴, 在很大程度上反应语言学界的一种比较普遍的心理: 恢复语言研究中人工数据和自然数据的平衡。既然语料研究方法与内心的唯理方法各有所长, 为什

么不让二者共存或结合使用,发挥其互补优势呢? 许多语言学家与语言教学专家发出了这样的呼吁,“我不认为有这样的语料库:它能包括有关我要探究的英语词汇和语法领域的所有信息,不论其有多大……同时,每每有机会检索语料库,总使我获得一些用其他方法无法得到的语言事实,无论其有多小。我的结论是:两类语言学家相互需要。”

在此背景下,利用语料对语言进行研究的成果不断出现,很多成果已被用于辞典编纂和语言教学等实际工作。例如,1980 年由 Sinclaire 主持的一个语料库工程,即 Cobuild 计划,这是 Collins 出版公司与 Bingmingham 大学的一项合作计划。他们搜集了大量的英语口语与书面语言素材,逐一分析每个单词的语法、语义、语体、语用特征,并将这些资料输入到计算机数据库中,建立起了一个 750 万词的语料库,后来这个语料库扩大到 2000 万词。根据这个数据库中的语料,他们陆续出版了 COBUILD 辞典和语法等工具书。这些工具书中的例句取自真实的语言素材,辞典中的释义排列顺序由语料库中得出的统计结果来决定,更加客观地反映了英语的使用情况。此外,词的释义方法更利于语言学习和语言教学。因此,一经问世,他们就受到语言学界及语言教师和学生的青睐。

进入上世纪 90 年代以后,随着网络技术的广泛使用,基于网络的语料库使用日趋广泛,并呈现出新的特点。首先,语料库的规模越来越大,其次,基于网络的语料库的出现,使得语言教师、语言学习者和研究者们可以直接以较低的成本在线查询、下载网上语料库。最后,语料库的种类也呈专门化、多元化趋势。某些用于特殊目的的语料库也相继建成。如专门用于研究儿童语言学习的语料库 CHILDES (http://childes.psy.cmu.edu),专门用于研究中国大、中学生英语学习的语料库 CLEC (http://www.elasscom.cn/lexi/lexi.html),专门研究科技英语的语料库等等。

二、基于网络的语料库在外语教学中的使用

基于网络的语料库为外语教学提供了海量而又鲜活的语言原料,是编写辞典、语法书及各种教材的重要语料来源。事实上,当代一些对外语教学有重要影响的辞典和语法书就是以基于语料库的方式编写的。这些著作深刻地影响着外语教学。除此而外,语料库在外语教学的思想、理论、方法、内容及手段等方面都起着十分重要的作用。下面我们从外语教学的几个侧面,来说明语料库在外语教学中的广泛应用。

1. 网络语料库与语法教学

例证是语法教学中常用的手段。英语教师在教学中所用的例句往往是凭语感造出来或是从语法书中查到的。由于辞典或语法书中的例句都很短,而且也不能给出足够的语境供参考;此外,辞典和语法书上的例句有时是过时的、不准确的,因此,利用这些方式得到的例句不能全面地反应英语语言的使用情况。语料库所提供的例句是以真实的语篇为基础的。这样得到的例句既真实又生动,具有时代感,说服力强。只要语料库达到相当的规模,可供选择的例句的数量是相当客观的。

以前,我们在规定主义语法教学思潮(prescriptivism)的影响下,语言教师常常受语法规则的束缚,不厌其烦地告诉学生正确的英语句子“应该”是什么,并把这些规则作为金科玉律让学生去记诵,而不是让他们自己在实际的语料中,去发现英美人“实际上”是怎样说英语的。例如,有一条广为人知的“规则”:定语从句的先行词前有限定词 all, any, every, only, some 或序数词或形容词最高级的修饰时,其后的关系代词要用 that,而不能用 which。这一“规则”在考试中也成为热点项目。但是,我们从 COBUILD 语料库中查询的结果却告诉我们,英美本族人在其口头和书面言语中却没有“遵守”这条规则。下面是我们从这个语料库中撷取的这条规则的“反例”:

(1) and they ascertained that all which the devil had revealed to him was

(2) We discussed in detail beforehand everything which might cause conflict in the

(3) emerge before the Council holds it's first meeting which is due to take place

(4) appreciate and welcome all activities which accord with the progressive

(5) e picked up all those which are currently undecided, Clinton

(6) whose character is thus marked by every act which may define a tyrant, is

(7) of do I retire or can we come to some arrangement which lets me carry on

另外,我们还利用“英国国家语料库”(British National Corpus)查询了 that 和 which 在这种搭配中分别出现的频率。如下表所示,虽然 which 出现的频率没有 that 高,但这绝不能说明在先行词为不定代词或接受特定词修饰时只能用 that 而不能用 which,更不能作为考试命题的依据。

由此我们可以看出一个问题:我们在课堂上交给学生的东西、我们的各种考试要求学生掌握的语言规则,是不是当前大多数英美人正在使用的语言? 对语料库中的材料进行统计和分析,发现哪些表达方式是

现在英美本族语人士所使用的,就把教学的重点放在这些项目上。这样做,可以减少教学的盲目性,让学生学到自然的、地道的语言。

先行词接受 all, any, only, first, every 及 best 修饰时关系词 that 与 which 在 BNC 语料库中出现的频率分布

修饰词 / 关系词	All	Any	Only	First	Every	Best	%
That	2380	2368	2628	1851	568	288	74.5%
Which	1390	1172	442	323	83	37	25.5%

因为语料库能方便地将具有同一特征的语言项目作为查询结果返回给用户,所以,我们可以有效地利用语料库来发现过去被忽略的语言规律。Pierce 曾对 3 万个词的连续文本进行过分析。他首先把文本中出现的全部动词和名词分别列出来结果他发现,列出的动词 90% 以上与所列出的名词相同。其他类似的分析得出的百分比也支持这一结论。我们都知道英语中有些词既可以作名词,也可以作动词,但如果不是借助语料库,我们不会估计到这类词在实际交际中出现的频率会如此之高。因此,Peirce 得出结论,我们应该把这类词叫"动- 名词"(verb- noun)而不是单纯的动词或名词。他认为,这样分类对语法教学大有裨益。另外,Sinclair 在对英语中 of 一词的搭配在语料库中的分布频率作了研究,对它的词类归属也提出了质疑,认为不应该将其归为介词。因为其他的介词一般位于名词前,构成介词短语,而 of 却是对于它前面的名词更为敏感。

2 语料库与词汇教学

在我国外语教学中,词汇教学是"单"词的教学,向来有只见树木不见森林的流弊。但是,人类语言中的词汇与其他词之间存在着各种音、形、义之间的联系。如果孤立地学习一个词,是无法看到这种联系的。但目前,我们可以利用现代计算机技术,可以迅速方便地从包含数百万上千万词的语料库中,把某个词或短语在这个语料库中出现的全部实例检索出来,并且统计出该词或短语出现的频率。这样,我们更准确更全面地建立词汇之间的关系,认识各语言形式在实际交际中的意义和用法。

我们以名词 survey 为例。常用的英语词典中都列出了 make a (general) survey of 这样的搭配。但我们通过对 200 万词的语料以 survey 和 surveys 为关键词进行检索,就会发现 make 和 survey 这种搭配竟然一次都没有出现,而返回的检索结果显示,使用最多的搭配是 conduct/carry out a survey,并且被动用法远远高于主动用法。

循序渐进是语言教学中的一条重要原则。但应循何序却是一个见仁见智的问题。传统的英语教科书以语法结构为序,流行于上世纪七、八十年代的功能-意念法却以语言的功能为序。而当代的交际法以语言的实际使用作为语言教学的途径和目的,所以其词汇、语法、话题都是以它们在实际语言使用过程中的频率为序的。但是,在语料库出现以前,对语言使用频率的排序是一件非常繁复的工作。但有了语料库之后,这一工作则变得非常轻松而又准确。例如,在传统的情态动词教学中,一般是循着"can- may- must- shall- will- should- would"的顺序来进行的,但在经过查询国家英语语料库,我们所得到的结果却与这一顺序大相径庭(见图 1)。

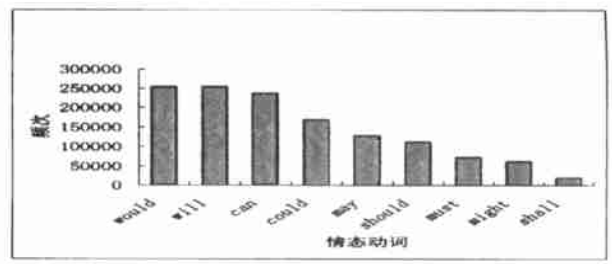

图 1 英语情态动词在英语国家语料库中出现的频次

图 1 显示,英语中最常见的情态动词并非 can,而是 would。如果我们在教材编写及教学设计中,能考虑到这些词在自然语言中的实际分布,将最常出现的词先教,那么学生学到的语言及使用的语言也就更接近自然状态。这样更有利于激发学生的学习动机,使语言学习和言语交际更紧密地结合在一起。

3 语料库与学生语言错误分析

错误分析曾盛行于上世纪六、七十年代。其目的在于发现学生产生言语错误的原因及语言学习的规律。但在语料库出现以前,我们只能从零散的学生作业或口语中获得不系统、不全面的言语错误。我们无法看到学生错误的全貌。语料库可以帮助我们发现学生错误中一些深层次的问题。以下是我们从中国学生英语学习语料库(CLEC)中查询 have 一词后所返回的结果。

(1) it. On the ground have more than one hundred country's

(2) in the Beijing road have many people here and there

(3) to Conghua. There are have many person.

W inter. . .

(4) because in there have many choices to. . .

(5) the same use? Or there have some different. . . ?

(6) can do some sport of have a good rest or repair

这 6 个来自中国中学生实际语言的例子显示, 中国学生在初学英语的 have 一词时, 常常受到英语中 there be 结构的干扰。因为在英语中, 存在性"有"与所有性"有"是两个词, 但在汉语中, 则只有一个词。由于母语的干扰, 就会出现这两种结构混用的现象。

何安平运用 Vocaprofile 和 MicroConcord 两个检索系统调查分析了我国广东省中学生高考作文语料库(约 13 万字)中拼写错误的类型与频率。她的研究发现约一半以上的中国中学生拼写错误很可能与其发音失误有关, 而且辅音失误造成的拼写错误要比元音失误所造成的更为严重。

对比分析是帮助学习者发现语言学习问题时的一种有效手段。而语料库的出现使这一手段更为完善。如果把英语学习者语料库中的材料和操本族语人士的语料库中的相关材料进行对比, 就可以发现一些问题。例如, Granger 发现英语本族语人士出现的最多的 5 个连接词 however, therefore, so, then, thus; 而英语学习者语料库中最常见的 5 个连接词则分别是 in fact, so, indeed, for example, for instance。从这个对比可以看出, 学习者在写作中出现的一个问题是对某些连接词使用过多, 而对另一些连接词则使用不足。这些错误虽然不是语法错误, 但正是因为这些错误, 学习者的语言才显得不够地道。利用语料库可以帮助我们发现这些问题。

4. 语料库与学生自主语言学习

语料库可以帮助学生形成自主语言学习能力, 达到由"学会"向"会学"的转变。在以往的英语教学中, 经常见到的是教师总结出语言规律, 学生在笔记本上记下这条规则, 然后死记硬背在考试中"运用"这些规则。如果利用语料库, 可以让学生利用关键词查询技术, 找出某个词或结构在语料库中出现的例句, 再让学生研究这些例句, 从中自己归纳、发现出语言规律。

语料库可以帮助我们建立一个全新的自主语言学习模式。如图 2 所示, 我们可以在对学习者个体特征的分析、个体学习者的语料库的基础上形成对学习者需求的分析, 再根据这一分析, 从学习材料语料库中选取适合这一学习者的语言学习材料, 为这一学生设计出根据他本人的特点"定制"的语言学习模式, 然后, 让学生自主学习语言, 最后对其学习效果进行评估。自主学习及评估的结果又可以及时地反馈到学习者特征组块及学习者语料库中, 使得这一自主学习过程越来越精确地贴近学生的学习需求, 体现了以学习者为中心的思想, 帮助学生形成自主语言学习能力。

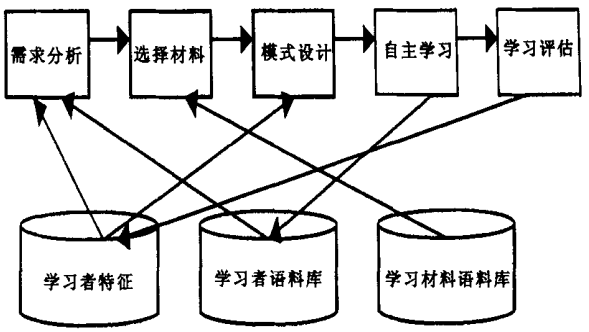

图 2　基于语料库的自主语言学习模式设计流程图

语料库的出现, 为外语教学带来了福音。特别是基于网络的语料库的出现, 不仅可以使我们快捷方便地使用这一丰富而又具有生命活力的语言宝库, 在外语教学的各个领域施展身手, 更为重要的是它深刻地改变了我们的语言教学思想, 使原来仅仅停留在理论上或空想阶段的语言教学思想变为现实, 如以学生为中心的思想、使用真实语言材料的思想等等。作为 21 世纪的外语教师, 我们要学会利用这一工具, 为我们的外语教学服务。

[参考文献]

[1]陈建生. 关于语料语言学[J]. 当代语言学, 1997, (1).

[2]引自 C. J. Fillmore (1992), "Corpus Linguistics", 载 J. Starvic (ed) (1992) Directions in Corpus Linguistics, Berlin: Mouton de Gruyter.

[3]如杜秉正, 董眉君. (1991) 大学英语语法与练习[M]. 上海: 上海外语教育出版社; 章振邦. (1989) 新编英语语法(修订本)[M]. 上海: 上海译文出版社.

[4]J. E. Pierce (1985) "The Nature of English Grammar." English Teaching Forum, Vol. XXIII, No.

[5]J. Sinclair (1991) Corpus, Concordance, Collocation. Oxford: Oxford University Press.

[6]陈建生. 定位检索软件辅助英语教学[J]. 外语教学与研究, 1997, (2).

[7]何安平. 学生英语拼写错误分析"[J]. 外语教学与研究, 2001, (3).

[8]S. Granger (1994) "The Learner Corpus: a revolution in applied linguistics" English Today, Vol. 10, No. 3.