语 言 学

对小型语料库的初步研究

赵宏展
(山东大学, 山东 济南 250100)

[摘　要]　在语言教学和研究学界, 个人自建小型语料库因其孕育着广阔的应用前景而成了一个热门话题。语言教师或研究者自建小型语料库之前, 应首先对小型语料库相关理论原则进行思考。这些理论原则包括小型语料库的定义; 小型语料库的体积; 小型语料库的代表性和有效性等问题, 从而为将来自建语料库活动的展开作好理论准备。

[关键词]　小型语料库; 自建; 定义; 代表性; 有效性
[中图分类号]　H319　　[文献标识码]　A　　[文章编号]　1008－4053(2006)012－0214－02

"语料库语言学正在成为主流"(Svartvik)。在未来的语言教学实践中, 基于语料库的方法将越来越深入人心, 逐步成为语言教学和研究的必备手段。近年来小型语料库逐渐兴起, 个人自建语料库因其孕育着广阔的应用前景而成了一个热门话题。辛克莱 1997 年就曾指出,"语料库语言学虽然取得了诸多成就, 但在很多目前的做法上甚至是在理论基础方面, 尚无长期、稳定的方针"。由于语料库的设计和建设是在系统的理论语言学原则指导下进行的(杨惠中, 2002: 36), 这就要求我们在小型语料库的建设问题上采取审慎的态度, 在建库之前首先进行多方面理论原则上的思考。

一、小型语料库的定义

小型语料库的定义到目前为止尚未确立。在语言教学中, 大型语料库通常应用于教学大纲的编制和教材的编纂, 而应用于课堂教学的语料库则有所不同, 它一般来说是精心采集的, 旨在帮助语言学习者理解语言现象的小型语料库(Guy, 1995: 259)。在上述内容中, Guy 将语言学习者看作是小型语料库的使用者, 认为小型语料库的作用只是帮助语言学习者理解语言现象, 从而忽略了外语教师这一主流使用人群研究用目的和行为。实际上, 语料库用户如今是以语言教师和语言学习者共同为主流用户, 语言教师、学生、教材开发者、语言习得研究者使用并开发自己的语料库, 逐渐成为语料库的主流用户(杨惠中, 2002: 35－51)。所以, Guy 的定义对小型语料库用户的说明是不全面的。与 Guy 不同, McEnery, Xiao 和 Tono 在其所著 Corpus－based Language Studies 一书中将个人自建语料库称做 DIY corpus(2005), 认为其作用在于帮助研究者研究某一特定问题。这显然只描述了小型语料库在语言和教学研究方面的功用, 而遗漏了其教学辅助的功能。

Guy 和 McEnery 等人都忽略了一个重要事实, 即小型语料库分为两种类型: Learnable Corpus 和 Specialized Corpus, Learnable Corpus 记录的是语言学习者中间语, Specialized Corpus 属于特定用途的语料库(Ghadessy, Henry & Roseberry, 2003)。Guy 和 McEnery 等都没有能够涵盖小型语料库的这两种类型, 或是只将小型语料库描述成 Specialized Corpus, 或是只将小型语料库描述成 Learnable Corpus。所以, 他们对小型语料库的定义都有失偏颇。

McEnery, Xiao & Tono 从另一个角度提出了一个全新的观念。"我们不认为个人自建语料库就一定是小型语料库","机读语料的可及性严重影响着语料库的大小"(2005)。互联网为我们提供了大量机读语料的来源, 如果不考虑版权的因素, 个人完全可以利用相关工具在短时间内下载并处理大量语料。语料库常用工具软件 WordSmith V4.0 中的 WebGetter 工具也有在短时间内大量搜集语料的功能。

语料库语言学界目前虽然尚未在小型语料库定义上达成一致意见, 然而对小型语料库定义的研究, 恰恰可以验证辛克莱 1997 年所做的论断:"语料库的概念还在发展中"。

二、可及性和数据冗余

大型语料库的优点在于其所包含的语料样品多、代表性强、产出的数据复杂。但大型语料库带来的并不都是方便。

1. 语料库的可及性

大型语料库因为过于庞大, 价格昂贵, 其可及性不高(梁茂成, 2003)。例如, 英国国家语料库的世界版 CD 光盘的价格为 50 英镑, 购买时还须支付 10 英镑银行手续费和 7 英镑的运费、包装费, 而 BNC Baby 光盘也价值不菲。若要将教学科研工作需要的语料库都收集起来, 尤其是收集国外大型通用语料库并做到能够及时更新, 对英语教师个人来说是很大的负担。所以大型语料库对使用者个人而言往往遥不可及, 不如自建的小型语料库使用方便(梁茂成, 2003)。

2. 数据冗余

大型语料库的包罗万象有时会成为研究者的一种负担。在使用大型通用语料库进行研究的过程中, 经常会遇到数据冗余的问题。例如, 使用 BNC 第二版光盘对情态动词 will 进行 kwic 索引, 会得到大约 250, 000 条结果。这些结果覆盖了多种文类和信道, 纷繁复杂、舛误多变, 想在其中找出规律性的东西是仅靠人工观察异常困难。针对这样的问题, 英国艾塞克斯大学"W3－Corpora"工程专家组在其 World Wide Web Access to Corpus 一文中提到:"对于语料库的大小和研究所需的语料数量目前没有给定的定义, 重要的是要有'足够'的数据, 至于什么是'足够'的数据应该具体问题具体分析"(Arnold, 1998)。为解决数据冗余的问题, Arnold 等人建议:"如果研究个别现象, 使用某个小语料库或大型语料库的子库可能会更好一些"。

数据冗余另一方面的表现在于文本处理和分析工具软件能力不足。"目前语料库相关软件的文本分析功能单一, 且对分析过程和结果缺乏必要的说明和解释"(杨惠中, 2002: 59), 而且许多语料库应用工具对词语索引的处理能力都有上限。例如, WordSmith 第三版最多能够提取 16, 868 条词语索引, 这在处理大型语料库数据的时候是很不方便的(McEnery, Xiao & Tono, 2005)。虽然 WordSmith 第四版已经解决了上述问题, 但对个人来说, 频繁更换工具软件会进一步降低所用语料库及相关软件的可及性。

自建语料库的大小有赖于使用目的和一系列的现实考虑(McEnery, Xiao & Tono, 2005)。实际工作中语料库的大小之选并非唯"大"是举, 应该根据具体教学和研究工作的内容具体问题具体分析。

三、对小型语料库的代表性的争论

小型语料库虽然使用方便, 其语料代表性容易受到质疑: 小型语料库或语料库子库中的样品, 往往不足以代表全体样本(Arnold, 1998)。里奇曾经指出, 大量收集的机读电子文本是概率研究方法中获得"必须的频率数据的基础", 为获得必须的频率数据, 我们必须分析足量的自然英语或其它语言文本, 以便基于观测频率进行合乎实际的预测(Leech, 1987: 2)。但由于语料库的大小和研究所需的语料数量目前没有给定的定义, 目前谈到小语料库的代表性, 只是相对于通用大型语料库而言, 指小型语料库语料数量相对偏少或取样比例过低, 不具有全面的代表性, 因此认为基于小型语料库的研究也就不具备有效性。然而,"围绕某些可识别的文类与各种语体标准所提供的语料库材料, 其构成应以用户需要为基础, 即用户能够根据自己的学习和研究需要, 通过汇集(语料库材料)或把语料库重新切割成各个微型语料库, 获得自己的平衡和代表性"(杨惠中 2002: 57, 引自 Murison－Bowie 1993: 50)。所以在 Murison－Bowie 看来, 小型语料库

[收稿日期] 2006－07－06
[作者简介] 赵宏展(1970－), 男, 山东大学外国语学院 2004 级研究生。

在平衡和代表性上反而会增强。

对小型语料库代表性的怀疑还表现在对语料“自然性”的疑问上。例如,由于学习者语料库是通过收集语言学习者各种书面语和口语的自然语料而建立起来的,学习者语料库中的语言运用材料的自然性就不同于一般意义上的语料库(杨惠中,2002: 60—61)。目前国内建成的学习者语料库都是收集学习者作文。一般来讲,自由作文更符合语言学习者语言运用的常态,所以数据的自然性就强;试卷作文虽然能够反映学习者目前的写作水平,但考场中的压力和焦虑感是他们的作文并非常态的语言运用,其数据的自然性就弱。无论来自自由作文、指导作文、试卷作文还是其他方面,学习者语料库语料都不免带有诱导特性,不能算是纯粹的自然语料(杨惠中,2002: 61)。不纯粹的语料自然性就差,进而代表性就不高。在学习者语料库语料的自然性问题上,目前并没有很好的解决办法。Granger 认为学习者语料库的数据只能尽量追求自然(杨惠中 2002,引自 Granger 1999: 2)。

四、小型语料库的有效性

如今一个外语教师或语言研究者为满嘴及特殊目的建立一个专用语料库已不是一件很困难的事(Meyer 2002;梁茂成 2003)。如果某外语教师在教学中决定自建小型中间语语料库,语料来源是他所教班级全体同学的指导作文。以这种方式建立起来的中间语语料库体积不会很大,与大型通用语料库相比是小型语料库,基于这种个人自建的小型语料库的研究其有效性就一定不高吗?统计研究表明,决定语料代表性的主要因素不外乎样本抽样的过程和语料量的大小,语料库建设中可以通过控制抽样的过程和语料比例关系来缩小偏差,增强语料的代表性(杨惠中,2002: 36)。事实上,如果我们从抽样率的角度分析该小型语料库的代表性,会发现该教师对自己学生的抽样率为百分之百,这远远高于大型通用语料库的抽样比例。因而,建立在该小型语料库之上的研究,对该教师个人教学工作来说具有高度的代表性。

小型语料库的代表性已经得到过验证。Coats(1983)曾利用 Lancaster Corpus 和部分的 London Corpus 对英语情态动词进行研究,这两部分语料库的总型符数为 1,725,000。Mindt 于 1995 年采用了 Brown 和 LOB 语料库以及部分的 London—Lund 语料库开展了相似的研究。虽然两人所采用的语料库的大小差别甚大,他们的研究结果却惊人的相似”(Meyer,2002: 13)。Meyer 由此得出结论:“对于某些经常出现的语法结构,采用小型语料库就可以可靠地进行研究”。

五、结语

小型语料库的建立,可以加深广大外语教师语言研究人员对语料库的认识,方便他们的教学和研究。然而辛克莱曾指出“在语料库研究的许多领域,情势仍很不稳定、难以制定和实施明确、严谨的标准”。在小型语料库建设的问题上,我们首先进行一些理论原则上的思考,谋定而后动,能使我们自己在小型语料库的建库过程中少走不少弯路。◇

参考文献

1 Arnold Doug. Corpus Linguistics: Introduction[M/OL]. http://www.essex.ac.uk/linguistics/clmt/w3c/corpus_ling/content/introduction.html, 2006—05—09
2 Ghadessy, M., Henry A. and Roseberry R. L. Small Corpus Studies and ELT: Theory and Practice[M]. Amsterdam: John Benjamins, 2003
3 Guy Aston. Corpora in Language Pedagogy: Matching Theory and Practice, in Principles & Practice in Applied Linguistics[M]. New York: Oxford University Press, 1995
4 McEnery. A., R. Xiao & Y. Tono. Corpus—based Language Studies: An Advanced Resource Book[M]. London: Routledge, 2005
5 Meyer, C, F. English Corpus Linguistics: An Introduction[M]. Cambridge: Cambridge University Press, 2002
6 梁茂成. 利用 WordPilot 在外语教学中自建小型语料库[J]. 外语电化教学, 2003, 94(06): 42—45
7 杨惠中. 语料库语言学导论[M]. 上海: 上海外语教育出版社, 2002

[责任编辑: 一然]

(上接第 213 页)2. 言语交际中的共知性有时体现为谈话双方的共知,这些共知不为第三者所了解,即存在一种双方共同理解并接受的预设,这样交际才能正常进行。请看话剧《春雷》中的一段对白:

蘩漪: 怎么这两天没看着大少爷?
四凤: 大概是很忙。
蘩漪: 听说他也要到矿上去,是么?
四凤: 我不知道。
蘩漪: 你没有听见说么?
四凤: 我不知道。——他说,他问太太的病。
蘩漪: 他到是惦记着我。〈停一下,忽然〉他现在还没有来么?
四凤: 谁?
蘩漪: 〈没有想到四凤这样问,忙收检一下〉嗯——大少爷。
四凤: 我不知道。
蘩漪: 〈看了他一眼〉嗯?
四凤: 我没有看见大少爷。

上述台词的中心话题是周萍。蘩漪“进攻”、“试探”,四凤则“防守”、“闪避”,双方心中都有“鬼”,这就构成了存在于双方之间的语用预设。正是由于这种预设的存在,双方的语言交锋才如此精彩。

从这段对话中我们推导出:蘩漪想从四凤的口中了解周萍的情况,但为了避嫌连周萍的名字不敢提,只能用长辈关心晚辈的语气来询问四凤,而四凤明知蘩漪的目的却装作不懂。如果我们不了解周萍、四凤、之间的三角关系,根本就不懂这段对话深层的含义。可见预设的“共知性”对交际双方理解对方的谈话目的有着决定的作用。

(三)“主观性”

是指预设总是倾向于发话者的经验和知识背景。所谓共知信息,实际上首先是发话者的主观认知状态,在受话者对发话者的设定没有认可之前,它总是相对于发话者而存在的(陈意德 2005: 31)。例如:

A: 今天买了一台电脑,明天打算买个“猫儿”。〈上网用的调制解调器〉
B: 什么?你喜欢养猫吗?

这个笑话之所以产生,是因为说话人主观认为“猫儿”这个词的所指是大家都清楚的,便没多做解释,而受话人头脑中预先并没有这个知识,“预设”不能被双方共同理解,于是产生了误会。

(四)“隐含性”

是指言语中人们总是从经济原则出发,总是想用最少的言语来表达最多的意义。这种隐含性要通过听话人结合一定的背景知识去理解推导出说话人的真正意图。在许多广告用语中,预设的隐含性被运用得恰到好处。例如:

我跟我媳妇说“你也弄瓶贵点儿的呀”,可人家就认准大宝了(任万芳 2003: 26)

这则广告里隐含着许多预设:该产品的价格合理;“人家”者代词在这里除了代指“她”而外,还预设了“我媳妇”在选用化妆品方面的权威性,用“人家”做主语预设其行为一定是明智的、优越的、值得效仿的。

三、结语

总之,预设是人们言语交际中常见的语言现象,尽管有时人们并没有明显的意识到,但在话语理解和会话交流中都有重要的交际价值:有利于语言的简洁,有助于传达更多的信息,有利于发话人实现其隐蔽的意图,有助于实现发话人原本难于达到的交际目的。利用语用预设的特点形成的交际策略,对提高语言质量,增强交际效果大有裨益。因此,我们在交际过程中要善于利用语用预设,成功的实现交际。◇

参考文献

1 Levinson, S. C. Pragmatics[M]. 北京: 外语教学与研究出版社, 2001
2 Peccei, J. S. Pragmatics[M]. 北京: 外语教学与研究出版社, 2000
3 Yule, G. Pragmatics[M]. 上海外语教育出版社, 2000
4 陈意德. 认知、预设及预设推理[J].《中国外语》, 第 5 期, 2005
5 何兆熊.《新编语用学概要》[M]. 上海: 上海外语教育出版社, 2000
6 何自然, 冉永平.《语用学概论》(修订本)[M]. 长沙: 湖南教育出版社, 2002
7 任万芳. 预设语境歧义[J].《张家口师专学报》, 第 4 期, 2003

[责任编辑: 文哲]