·科技翻译与新技术·

小型翻译语料库的DIY*

赵宏展

(山东中医药高等专科学校基础部 烟台市 265200)

摘 要 自建小型翻译语料库在翻译教学和研究领域中孕育着广阔的应用前景,最近成为学界的一个热门话题。我国翻译教师和翻译研究者对自建小型翻译语料库大多缺乏必要的心理和技术准备,认为与自己的研究大异其趣且技术高不可攀。事实上语料库技术很多只是从属性的,作为开发者和应用者的个人只要把主要精力放在内容的选取和呈现方面即可,而不必在技术方面投入过多精力。通过使用相关工具软件自建小型翻译语料库,能让我们熟悉并掌握语料库这一先进工具、克服"技术恐惧症",让语料库真正走进我们日常的翻译教学和科研工作。

关键词 小型翻译语料库 翻译教学

Abstract Of recent, there has been a growing interest in using small translation corpus in translation teaching. The present author explains that it is actually not so difficult for translation teachers to overcome their technophobia and create their own small teaching corpus with the help of some software tools.

Key Words small translation corpus translation teaching

"语料库语言学正在成为主流"——Svartvik。

近年来小型语料库逐渐兴起,个人自建小型语料库因其孕育着广阔的应用前景而成为热门话题。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[1]。在翻译教学和研究工作中,翻译课教师或研究者个人可以充分利用互联网资源和多种工具软件创建小型翻译语料库以辅助自己的工作。

1 为什么要DIY小型翻译语料库

大型语料库公认的优点在于语料数量庞大、语料样品多、产出数据复杂而且代表性强。大型通用语料库的数据规模目前已经达到几亿形符,且有规模急速变大的趋势[2]。通用语料库虽然包罗万象但对个人而言可及性不高,而且还有数据冗余的问题。另外,由于大型语料库讲求平衡选取语料,在辅助教学、编写翻译教材和教辅材料方面往往不如临时自建小型语料库的内容更有针对性、即时性和新颖性。

1.1 大型通用语料库的可及性

大型语料库因为过于庞大,价格昂贵,其可及性(accessibility)不高[3]。例如,英国国家语料库(BNC)的世界版CD光盘(个人授权)的价格为50英镑,购买时须另支付10英镑银行手续费和7英镑的运费、包装费,而BNC Baby光盘也价格不菲①。若要将教学和科研工作需要的语料库,尤其是国外大型通用语料库一一收集起来并做到及时更新或取得全权在线使用权,对个人来说往往是个不小的负担。所以大型语料库虽包罗万象,但对个人而言却经常遥不可及,不如自建的小型语料库使用方便。

1.2 大型通用语料库的数据冗余

包罗万象的大型语料库有时会成为一种负担。使用大型语料库时,尤其是对某些特定词、词组或搭配(复现组合)研判时,经常会遇到数

① 资料来源于英国国家语料库(BNC)的官方网站,Obtaining BNC [M/OL]. http://www.natcorp.ox.ac.uk/getting/index.xml. ID = intro,2006-05-09.

* 收稿日期:2006—06—12/31

据冗余(too much data)②问题。例如,使用BNC第二版对情态动词 will 进行 kwic 索引,我们会得到总数约250,000条结果。这些结果覆盖多种文类(genre)和信道(channel),内容纷繁复杂、形式多变。研究者若想在结果中找出需要的材料或特定的内容,仅靠人工观察就犹如大海捞针。针对数据冗余问题,英国艾塞克斯大学"W3-Corpora"工程专家组③在其所著 World Wide Web Access to Corpus: Corpus Linguistics 一文中解释道:"对于语料库的大小和研究所需的语料数量目前还没有给定的定义,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"。为解决数据冗余问题,"W3-Corpora"工程的专家建议:"如果是针对个别(语言)现象,使用小型语料库或某个大型语料库的子库可能会好一些"。

数据冗余还表现在文本处理和分析工具软件的能力不足。"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"[2]。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第三版最多能提取16,868条词语索引,这对处理大型语料库数据来讲是非常不方便的[4]。虽然 WordSmith 第四版已经解决了上述问题,但对翻译课教师和研究者个人来说,频繁更换工具软件的高额费用显然会进一步降低语料库及相关工具软件的可及性,从而使语料库更加遥不可及。

小大由之,有所不行。选择语料库的大小取决于使用目的和一系列的现实考虑[4]。实际工作中在对语料库的选择上显然不能唯"大"是举,应该根据翻译教学和科研工作的具体内容做具体分析。

2 小型翻译语料库的建立

在创建自己的语料库前,首先应根据该语料库的用途确定一些基本原则[2]。外语教师日常工作中经常需要临时编制一些专门的翻译教材和教辅材料,这时使用小型翻译语料库就得心应手。小型语料库的建立迄今未有明确的标准和规范,但一般包括语料采集(data capture)、标注(markup)和赋码(annotation)的过程。

2.1 语料采集和格式转换

翻译教学工作中教师常常要扮演教材设计者的角色,需要针对特定学生群体提供专门的教材和教辅材料。例如,某大学翻译培训班的学员大部分来自一个特定行业,他们希望所学教材除了能提高他们的基础翻译能力外,还能对其所属行业有相当的针对性,而现有翻译教材不敷使用。这时就可以临时建立小型语料库以补充编制教材和教辅材料的内容。

构建语料库所需语言材料的采集不是随意的。对小型语料库建设者来说,语料需要便于搜集、格式统一、内容要根据实际情况,有针对性。目前,因特网和各种大型电子文库无疑为我们提供了方便可靠的机读数据来源,然而把网页内容逐页下载的做法费时费力,可行性不高。在实际工作中可使用一些免费的小型工具软件(如 WordSmith 和 HTTrack 等)来辅助语料搜集,先进行关键词搜寻,然后将含有关键词的网页的文字性内容一次性下载。常用的语料库工具软件 WordSmith 4.0 有一个 WebGetter 辅助工具,能够进行关键词网页搜索并可就网页内容的语料语言、网页的最小字数、语料的最小字数等条件进行定制,定制完成后就可一次性多线程下载相关网页。WebGetter 主界面如图1所示:

可在 WebGetter 的"Setting"对话框中自由调整下载内容的存贮目标文件夹、最小字数、最大线程数、语言选择等项目,设定完成后单击"go"即可自动进行语料收集。另一个重要的语料来源是各种大型电子文库,如光盘版的《大英百科辞典》和 ENCARTA 等[3]。通过上述方式得到的语料一般是 HTML(Hypertext Markup Language)格式,需要将该格式的内容转换成纯文本(SCII)或 XML 格式,否则一些语料库通用软件工具如 WordSmith 和 WordPilot 等无法识别。对于大批量语料的格式转换最好采用 MLCT(Multilingual Corpus Toolkit)④工具包,该

② 资料来源于 Doug Arnold. Corpus Linguistics: Introduction [M/OL]. Http://www.essex.ac.uk/linguistics/clmt/w3c/corpus_ling/content/introduction.html,2006-05-09.

③ Arnold Dough 是著名的 W3C-Corpus 工程的负责人,该工程已经于1998年完成。

④ 该工具包可以到英国兰开斯特大学(Lancaster University)主页地址下载:http://www.lancs.ac.uk/staff/piaosl/research/download/mlct_public.zip,2006-05-19.

工具包运行于 Java⑤ 环境下(JRE version1.4 或更高)。由于很多外语教师不熟悉 Java,在实际工作中还可以采用一些批量转换工具或者将语料文件在 IE、Nescape 等浏览器中打开并另存为纯文本文件(*.txt)的做法。

图 1:WebGetter 主界面和搜索前的定制

在网页上采集语料虽然简单易行,但是要特别注意语料版权的问题。版权问题有时要比语料收集更为复杂。美国版权法中有"合理使用"的条款,指用于非商业性的研究工作时,可以使用受版权保护文章的部分或全部内容。然而,该"合理使用"不包括将受版权保护文章 2000 字以上的节选用于语料库发行的情况[5]。所以,对于个人制作的小型翻译语料库要严格限定其使用范围,最好仅供建库者个人使用,以避免版权纠纷。

2.2 语料的赋码

以上述方式获得的语料还要清除杂质和多余符号,并统一语料的格式和存放方式。语料最好是每一个文本作为一个独立文件单独存放,这样,研究时就可得出每个文本的统计特征及整个语料库的总体统计特征。如果语料库是一个文件,那么就只能检测出整个语料库的总体统计特征[2]。语料赋码工作的第一步通常是加注篇头(header),即给语料标注篇名、作者、文本、领域、语体、时代、出版信息、文本字数等[6],一般做法是将上述信息分别填入尖括号中并放置在文本第一句前面。例如:<ST 6> <SEX ? > <Y ? > <SCH GIFL> <AGE ? > <WAY 1> <DIC 2> <TYP 2>⑥。加注篇头目前还没有自动工具软件,在建立个人小型语料库的过程中加注篇头会耗费相当的时间和精力。考虑到小型翻译语料库的用途,笔者建议不进行篇头加注,一般情况下只进行词性赋码(POS tagging)即可。

语料进行词性赋码前应先确定赋码方案(Tagset)。从简化赋码的角度出发,可直接使用词性赋码软件中集成的方案,而不必另寻它路。目前进行词性赋码比较流行的工具有 AnnoTool 和 GoTagger 软件。GoTagger 是 Goto Kazuaki(日本)以 Delphi 写成的一款简便的词性赋码软件⑦,所以不需要运行 ActiveX 或 Dll 文件就可以直接在 Windows 上运行。其自带的赋码方案原来是针对法语的,我们必须将其赋码方案先替换成英语赋码方案,然后再进行赋码。具体方法如下:⑧

第一步:下载 GoTagger 软件 http://uluru.lang.osaka-u.ac.jp/~k-goto/GoTagger.zip. 并解压。

第二步:在 http://research.DIYsoft.com/

⑤ 该软件可以到 SDN 公司主页 http://java.sun.com/j2se/1.5.0/download.jsp 下载,2006-05-19.
⑥ 标注来源于中国学习者语料库(桂诗春、杨惠中.中国学习者英语语料库.上海:上海外语教育出版社,2003)随书光盘 ST6 子库首页第一行。
⑦ GoTagger 可以在以下地址下载:http://uluru.lang.osaka-u.ac.jp/~k-goto/GoTagger.zip
⑧ 具体方法在 http://uluru.lang.osaka-u.ac.jp/~k-goto/index.html 有详细介绍。

~brill/下载 Brill 赋码方案，另外一个下载地址是 http://www.cs.jhu.edu/~brill/RBT1_14.tar.Z

第三步：解压赋码方案文件，拷贝"Bin and Data"文件夹中的 10 个方案文件并将文件粘贴到 GoTagger 文件夹中的"G data"子文件夹中。

第四步：点击 GoTagger 图标，这时 GoTagger 软件就可以正常使用了。

GoTagger 不兼容中文，界面中文标记的文件夹会显示成日语片假名，所以电脑中的相关文件夹以英文命名会更方便。GoTagger 主界面如图 2：

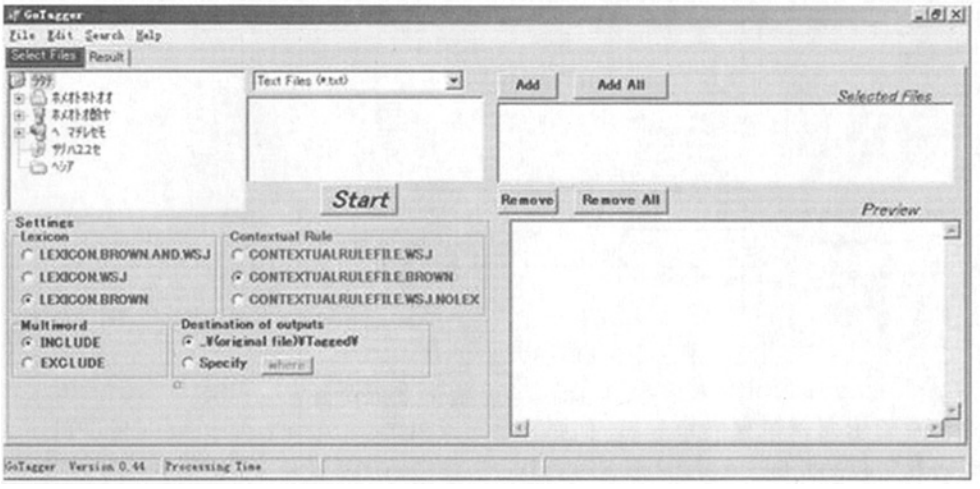

图 2：GoTagger 主界面

点击图标打开 GoTagger 的主界面后，需在左侧窗口选择赋码文件所在的文件夹，进入文件夹后选蓝文件，点击"Add"按钮然后单击"Start"，这时赋码就自动进行。赋码前，还可在主界面"Destination of outputs"对话框中预先选择结果的保存文件夹，一般是在硬盘上先建立一个新文件夹并命名为"Tagged"，经过赋码的文件会自动保存其中。GoTagger 软件的优点在于它有批处理功能，将需要赋码的大量文件一次性处理完毕。

2.3 语料的整合

语料赋码完成后，要用语料库工具软件将所有语料整合起来。这项工作交给 Compulang WordPilot 完成。Compulang WordPilot 是香港科技大学开发的一款旨在提高学生英语口语和写作水平的小型应用软件⑨，由于具备强大的检索功能和软件自身开放性的特点，可以将它作为小型语料库的建库工具使用[3]。WordPilot 主界面如图 3：

语料库建库过程的具体方法如下：

第一步：在 WordPilot 所在安装文件夹 Compulang > WordPilot 2002 > libraries 子文件夹中创建一个名为 translation 的文件夹，将所搜集的纯文本语料拷贝其中。

第二步：点击 WordPilot 桌面图标，然后点击 File 中的 New 选项，出现新建对话框，选择 text library 选项。

第三步：点击 Edit 的 Add text 选项，在系统弹出的对话框中打开存放在 translation 文件夹的纯文本格式文件。

第四步：点击 File 中的 Save as 选项，在对话框中键入 translation，保存为.clb 格式文件即可。

这样小型翻译语料库就制作完成。以这种方式建立的小型语料库是开放性的(open corpus)，可以根据工作的需要不断增加新的语料或建立新的子语料库。由于库引擎软件 WordPilot 具备如 kwic 检索、练习题和试题自动生成等多项教辅功能，除用作语料库引擎软件之外，还可以充分利用这些功能辅助我们的翻译教学。WordPilot 不能读取经过赋码的语料，对保存在上述 Tagged 文件夹的语料需要利用其它

⑨ 软件可以在 http://www.compulang.com 下载，网页上还有使用介绍。

工具软件如 MicroConcord、WordSmith 进行分析　　和研究。

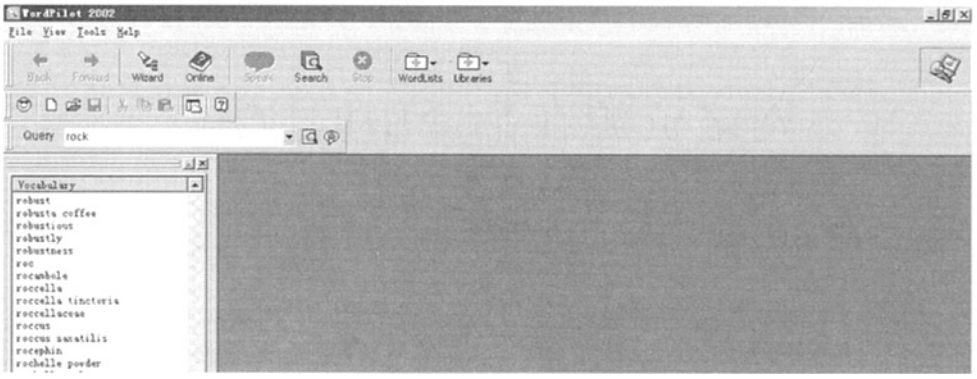

图 3：WordPilot 2002 的主界面

3　结　语

语料库的建立在全国范围方兴未艾，语料库及其应用软件为翻译教学和科研提供了一个全新的思路和方法。通过个人建立小型翻译语料库，广大翻译课教师和研究人员可以加深对语料库这一新的研究领域和科研方法的认识，方便自己的教学和科研工作并能有效克服所谓的“技术恐惧症”。辛克莱(2004)认为：“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”。小型翻译语料库作为一个全新的领域，有太多有待解决的问题，即便是小型语料库的定义目前语料库语言学界也尚未达成一致意见。以上建立小型翻译语料库的方法仍有不成熟、不完备的地方，在此仅供学界参考。

4　参考文献

1　Guy Aston. Corpora in Language Pedagogy: Matching Theory and Practice, in Principles & Practice in Applied Linguistics. New York: Oxford University Press, 1995,259

2　杨惠中. 语料库语言学导论. 上海：上海外语教育出版社，2002 ,30,59,63,153

3　梁茂成. 利用 WordPilot 在外语教学中自建小型语料库. 外语电化教学，2003，94(06)：42-45

4　McEnery. A., R. Xiao & Y. Tono. Corpus-based Language Studies: An Advanced Resource Book. London: Routledge, 2005

5　Meyer, C, F. English Corpus Linguistics: An Introduction. Cambridge: Cambridge University Press, 2002,61

6　王克菲. 新型双语对应语料库的设计与构建. 中国外语，2004，25(06)：73-75

(上接第 9 页)

译文所传递的信息而使信息无法得到正确的接受而耗散或重构。比如“心主神志”是中国传统文化的一个重要理论，翻译成英语即：The heart controls the mind。不了解中国文化或中国医药的西方读者读了这个译文后，一定会感到莫名其妙。根据他们所接受的现代科学教育，人的思维在大脑，不在心。对于这样的信息，他们可能认为是错误的而不予接受，或认为是译者笔误而自作聪明地发挥修正。

对外介绍中国文化时，如何解决这种因文化差异而导致的信息耗散和重构现象呢？只有一个途径，那就是加强中国的对外文化交流，努力将中国传统文化系统、全面、深入地传播到世界各国，让海外人士了解中国人的文化观念和思维方式。当然这个工作不是一蹴而就的，需要我们从事文化交流工作的人不断努力推进。

从我个人多年的对外文化翻译交流实践来看，在很多情况下，信息的耗散与重构都是由于我们自己在翻译时对原文的误释和误译所引起的。之所以会误释和误译，与我们自身的文化素养和知识结构有很大的关系。

明人梦醒龙在《古今谭概·无术部第六》讲了这样一则笑话：

魏博节度使韩简，性粗质，每对文士，不晓其说，心常耻之。乃召一士人讲《论语》，至《为政篇》。明日喜谓同官曰：“近方知：古人禀质瘦弱，年至三十，方能行立。”

一个封疆大吏居然将孔子“三十而立”曲解为“三十岁才能站立起来”，确实可笑。其实这样的笑话在我们理解和翻译中国古典文献时，并不罕见。这也就是翻译过程中信息不断被耗散和重构的根本原因所在。《大学》说：“本乱而末治者，否矣”。强调个人修养对立身务业的影响。修身在于务本，为学亦在于务本。根枯，枝无以立；枝萎，叶无以茂。我们在从事文化翻译的时候，首先要完善自身的文化素养，即要务本，本立而道自生。

小型翻译语料库的DIY

作者： 赵宏展， ZHAO Hongzhan
作者单位： 山东中医药高等专科学校基础部,烟台市,265200
刊名： 中国科技翻译PKU
英文刊名： CHINESE SCIENCE & TECHNOLOGY TRANSLATORS JOURNAL
年，卷(期)： 2007,20(2)
被引用次数： 8次

参考文献(15条)

1. 查看详情

2. 查看详情

3. 查看详情

4. 桂诗春;杨惠中 中国学习者英语语料库 2003

5. 查看详情 2006

6. 查看详情 2006

7. Arnold Dough是著名的W3C-Corpus工程的负责人,该工程已经于1998年完成

8. Doug Arnold Corpus Linguistics:Introduction 2006

9. ObtainingBNC 2006

10. 王克菲 新型双语对应语料库的设计与构建 2004(06)

11. Meyer C F English Corpus Linguistics:An Introduction 2002

12. McEnery A;R Xiao;Y Tono Corpus-based Language Studies:An Advanced Resource Book 2005

13. 梁茂成 利用WordPilot在外语教学中自建小型语料库[期刊论文]-外语电化教学 2003(06)

14. 杨惠中 语料库语言学导论 2002

15. Guy Aston Corpora in Language Pedagogy:Matching Theory and Practice 1995

引证文献(8条)

1. 赵晶 基于小型双语平行语料库对政治文献翻译显化的探讨——以近十年政府工作报告中"搞好"的翻译为例[期刊论文]-鲁东大学学报（哲学社会科学版） 2010(4)

2. 吉晓霞 国内翻译教学研究十五年回顾与思考——基于13种外语类核心期刊论文的统计与分析[期刊论文]-南京晓庄学院学报 2010(4)

3. 袁卓喜.何佩祝 网络资源与语料库方法在商务翻译中的运用[期刊论文]-怀化学院学报 2010(8)

4. 赵宏展 专题性英语口语教学语料库的建设[期刊论文]-山东教育学院学报 2009(3)

5. 纪可 广西-东盟平行语料库建设与地方翻译人才的培养[期刊论文]-东南亚纵横 2009(11)

6. 王正.孙东云 利用翻译记忆系统自建双语平行语料库[期刊论文]-外语研究 2009(5)

7. 王连柱.王兰英.张瑞君.雍文明 语料库及检索工具在医学英语词汇教学实践中的应用研究[期刊论文]-中国医学教育技术 2008(5)

8. 周杰 小型学习者语料库的建设与应用[期刊论文]-贵州大学学报（社会科学版） 2007(6)