

当代外语研究

CONTEMPORARY FOREIGN LANGUAGES STUDIES

10
2011

(总第370期)

- ❧ 思在昆仑山下——眼光与定力
- ❧ 创建子语料库，促成对比研究
- ❧ 语料库检索的模式、问题及启示
- ❧ 我国大学英语教学的问题与对策
- ❧ 《一千英亩》的生态女性主义解读
- ❧ 国学典籍翻译主体选择与译者能力培养研究

ISSN 1674-8921



9 771674 892116

当代外语研究

2011年10月第10期(总第370期)

目 录

大家治学

思在昆仑山下

——眼光与定力 钱冠连 1

语料库语言学

- 主持人语 梁茂成 5
- 创建子语料库,促成对比研究 许家金 梁茂成 6
- 语料库检索的模式、问题及启示 陈 功 10
- 中国英语学习者虚化动词 TAKE 词汇型式研究 刘国兵 14
- 面向信息处理的大规模句子语义知识库研究述要 周明海 亢世勇 刘金凤 20

语言习得与教学

- 我国大学英语教学的问题与对策 徐锦芬 26
- 试论翻译在外语教学中的重要性 曹燕萍 32

翻译研究

- 国学典籍翻译主体选择与译者能力培养研究 刘立胜 廖志勤 35
- 铃木大拙《文昌帝君阴骘文》英译本赏析 何立芳 39

文学理论与批评

- 《一千英亩》的生态女性主义解读 都岚岚 43
- 以《黑王子》为例浅析艾丽丝·默多克的女性意识 安 宁 袁广涛 47

书评

- 评苏珊·巴斯内特的新著《全球化时代的新闻翻译》 卢小军 53

专家访谈

From California to Beyond

—An Interview with Mr. Brook Thomas 曲 鑫 57

其他

- 《当代外语研究》来稿规范 4
- 第四届中西语言哲学国际研讨会征文通知 42
- 《当代外语研究》征集读者意见 56
- 《当代外语研究》接受网络投稿 59

英文摘要 60

期刊基本参数:CN31-2039/H * 1980 * m * 16 * 64 * zh-en * P * ¥8.00 * 4000 * 13 * 2011-10

本期执行编辑:甄凤超

主持人语

梁茂成

近年来,语料库越来越受到语言研究者的青睐,但随着语料库研究方法的推广和普及,各种新的困难和问题也不断地暴露出来,而解决困难过程中的更多思考也给语料库相关研究带来了新的思想、策略和工具。本专栏中共包含四篇文章,涉及语料库研究的不同方面。

许家金、梁茂成的研究介绍了其研究团队自行开发的子语料库创建工具 Sub-corpus Creator。在当代,语料库创建者对所采集的语料常常进行两种类型的标注。第一种标注是元信息标注(metadata markup),即将语料来源、产生时间、产出者身份、产出者性别等若干非语言信息以标签形式加入到语料库中,以便于使用者区分不同类型的文本,进而分析同类文本的共同属性和不同类型文本的差异。另一种标注是语言信息标注(linguistic annotation),即对语料库文本内部的语言信息(如语用信息、语音信息、篇章信息、句法信息、词类信息等)进行标记,以便于使用者提取特定语言信息。Sub-corpus Creator 可以利用以上两类信息,从大型语料库中抽取符合某种条件的文本创建子语料库。将子语料库与总库或其它子语料库进行对比,是语料库语言研究的常见方法。Sub-corpus Creator 的检索式完全支持正则表达式,极大地方便了子语料库的创建,也有利于我们开拓更多的研究维度。

陈功的研究主要涉及如何对语料库中的语言信

息进行有效检索。准确检索并析出所得结果直接关系到研究的可信度。语料库研究最令人信服的一个方面是其可重复性。也就是说,利用同一个语料库,采用同一种方法,无论谁来操作,所得到的结果应该是相同的。然而,在已有的研究中,一些研究者所使用的检索方法常常不够严密,使得析出结果与真正的语言事实大相径庭。基于此类挂一漏万的结果,一些研究者借用种种理论进行了所谓的“深度分析”。在如今语言研究不断科学化的今天,“不知不为过”绝不可成为研究不严密的借口。陈功的研究对常见的检索进行了分类,并对检索中容易出现错误成因进行了分析,对提高研究信度有意义。

刘国兵的研究是一项基于语料库的实证研究。研究中考察了学习者语料库中的典型虚化动词 TAKE 及其搭配和类联接的使用情况。虚化动词不同于普通的实义动词,其意义常常取决于与其结伴的其他词。研究带给我们的启示是,语言教学中万不可孤立地教授虚化动词的用法,只有结合常见搭配和类联接进行虚化动词的教学才可能取得较好的效果。

周明海等人的研究首先考察了近年来国内外具有代表性的十二个句子语义知识库,特别关注了辅助语义角色的标注,并在此基础上,从语义角色的精细程度、数量设置、相对核心、标注格式等角度阐述了作者对构建句子语义知识库的几点认识。

北外语料库语言学研究团队介绍

北京外国语大学中国外语教育研究中心的语料库语言学研究团队(以下称“北外语料库团队”)经过数年的积累和发展,已成为国内语料库语言学研究的主要力量之一。该团队目前拥有专职研究员 4 人(王克非教授、梁茂成教授、许家金副教授、熊文新副教授),兼职研究员 1 人,全部拥有博士学位,专业领域涉及基于语料库的理论语言学、对比语言学、翻译学、应用语言学、计算语言学等方面的研究,是一个典型的跨学科研究团队。

北外语料库团队近年来积累了一系列成果。目前,该团队语料库语言学研究主要集中在双语语料库和学习者语料库两个方向。近年来完成或在研的省部级以上研究项目主要有:国家社科基金重大项目 1 项(由王克非教授主持)、国家社科基金一般项目 4 项、教育部人文社科重点研究基地重大项目 3 项、教育部人文社科一般项目 3 项,发表 SSCI、CSSCI、EI 级别的相关学术论文百余篇,专著十余部。

北外语料库团队的相关成果在全国产生了重要影响。由王克非教授主持建成的北外英汉-汉英通用

平行语料库是迄今为止国际上最大的英汉-汉英平行语料库。2010 年立项的国家社科基金重大项目“大规模英汉平行语料库的建立与加工”更将推出国际上前所未有的超大规模英汉、汉英平行语料库。

学习者语料库建设与研究是北外语料库团队相关研究的又一亮点。由文秋芳教授主持建立的“中国学生英语口语语料库”和“中国学生口笔译语料库”已成为中国学习者语料库相关研究的重要资源。基于学习者语料库分析需求而设计的 Collocator、Colligator、PatCount、PatternBuilder、Parallel Text Aligner、Readability Analyzer、TreeTagger for Windows、BFSU Stanford Parser 等已成为国内语料库语言研究者耳熟能详的分析工具。近期正在开发的 CROWN/CLOB 语料库项目、PowerConc 软件包项目更将有力推动我国的语料库语言学研究;以中心学者为主编写的《语料库应用教程》已成为国内语料库语言学的一本重要教科书。

我们衷心希望北外语料库语言学团队的研究能够得到广大语料库爱好者的更多关注与支持。

创建子语料库,促成对比研究

许家金 梁茂成

(北京外国语大学,北京,100089)

摘要:语料库研究中,常需要进行不同类型的对比,比如,英语专业与非英语专业学生英语作文的对比,双语语料库中源语文本与译语文本的对比等等。然而,语料库创建之初,文本多集中存储,研究者若要基于该语料库进行对比研究,需要根据研究目的化整为散,生成子语料库。本文以学习者语料库和双语语料库研究中的一些选题为线索,介绍子语料库的常见生成方法,同时藉此提出一些可供语言对比研究的分析视角。

关键词:子语料库,对比分析,学习者语料库,双语语料库

[中图分类号] H0

[文献标识码] A

[文章编号] 1674-8921-(2011)10-0006-04

1. 问题缘起

对比分析(contrastive analysis)是语料库语言学的常见方法和分析视角。国内语料库学者谈及对比方法,最容易想到“中介语对比分析”(Contrastive Interlanguage Analysis,简称 CIA)和“英汉对比研究”。中介语对比时,需要比较英语母语和学习者中介语的语言特点,或比较不同母语背景的外语或二语学习者的中介语,还可能会比较语言水平较低和较高的中介语表现(Granger 1998)。英汉对比时,可能会比较翻译的双方向(汉译英、英译汉)语言表现(Baker 2000;王克非 2006),也可能对比英汉语料内部的不同语类,如小说类语料或者科技类语料。在进行对比研究时,语料通常是分别存储的,研究者只需对两批语料分别检索,再对比数据和语言特点即可;若试图在语料库内部再做对比(比如四级和六级学生作文的对比、小说类科技类对比),则需将大的语料库拆分成若干小语料库,或者从大的语料库中抽取符合条件的子语料库(sub-corpus)。

比如,“英国国家语料库”(British National Corpus)是一个集英国英语口语和笔语的大型通用语料库。Rayson, Leech 和 Hodges(1997)利用该语料库进行了一项社会语言学研究。在该研究中,研究者对英国不同性别的对话口语进行了对比和分析。因研究问题所需,先从通用语料库中抽取对话口语,然后再区分对话者的不同性别,建成了两个子语料库——英国男性对话口语语料库和英国女性对话口语语料

库,以满足对比的需要。

子语料库的创建正是本文要讨论的主要问题。在实际研究中,视研究的需要,常常依据某种条件从一个大的语料库(如通用语料库)中抽取若干个符合条件的文本构建一个子语料库,以达到对比的目的。对语料库的拆分和抽取,可以拓展语言对比的研究层面,同时引发新的研究选题。

2. 语言对比的层面

语言对比涵盖范围很广。本文以探讨学习者语言对比和英汉语言对比为主,并着重讨论创建子语料库的操作方法,从而服务于语言对比研究。以往的检索工具多数不能对语料库中的特定文本,或文本中的特定内容进行检索,因而限制了多层面的语言对比研究。这是基于语料库的语言对比研究中缺失的一环。如能从技术上简化创建子语料库的方法,则能有效弥补这缺失的一环,同时还能扩展语言对比研究的层面。

以下将通过一些具体的研究选题,展示对比在语言研究中的重要性,以及语言对比研究的层面。之后,本文将就其中一些选题,简述如何通过创建子语料库,促成对特定语言内容的考察和对比。

利用语料库,可进行多个层面的语言对比分析,比如:

- A) 出现较多动宾搭配错误的学生语言特点研究;
- B) 使用较多复杂句的学生语言特点研究;
- C) 英美新闻语料中有关中国崛起(或中国威胁)的报道;
- D) 学习者语料库中成绩居前 30%与后 30%的语言特点对比研究;
- E) 中国学生英语口语复述与其英语会话之间的话语特征比较;
- F) 不同主题作文之间的语言特点的异同;
- G) BNC 中少年与成年话语的对比研究;

作者简介:许家金,北京外国语大学中国外语教育研究中心专职研究员、副教授。主要研究方向为话语分析、语料库语言学。电子邮箱: xujiajin@bfsu.edu.cn

梁茂成,北京外国语大学中国外语教育研究中心专职研究员、教授。主要研究方向为语料库语言学、计算语言学、第二语言习得。电子邮箱: frankliang0086@yahoo.com.cn

本文是 2008 年教育部人文社科项目“基于语料库的中国大学生英语口语话语特征研究”(编号 08JC740002)的部分研究成果。

H)《时代周刊》中2001年9月11日(911事件)前后新闻报道的异同;

I)新闻语料中的社论、评论与现场报道话语的异同;

当然,还有基于复合条件的语料分析。比如,

J)不同年龄女性话语中的被动用法。

综上,有关语料文本内部语言特征(如A、B、C)的研究,有基于语料文本外部特征(也称为社会语言学变量)(如D—I)的研究,以及综合语言特征和社会语言学变量(如J)的研究。

3. 子语料库的创建

要获得上面选题所需的研究语料,有不同的思路。一些在线语料库检索系统(如<http://corpus.byu.edu/>)可按文本外部特征分别检索和呈现结果。比如,《时代周刊》语料库(1923-2006)就提供了按年代进行检索的设置。因此,检索terrorism, bin Laden就会得到2001年前后明显不同的语言分布,进而还可以分析2001年前后这两个词的典型搭配的情况。然而,在线语料库的使用只能是“因陋就简”,网站提供怎样的选项设置,使用者就只能在给定的范围内使用。比如,若要对BNC中男性与女性话语的特点,利用<http://corpus.byu.edu/bnc/>就无法实现。在线语料库检索界面的设计者,不可能考虑到各个用户的不同需求。因此,利用语料库中的元信息等创建子语料库,才能使我们的研究需要得到满足。

3.1 语料库元信息的结构与内容

如果用户电脑上存储了相关的语料库(如BNC、CLEC、SWECCL等),希望从这些语料库中提取所需的语料,创建子语料库,本文提出一个简便易行的方法。其核心是获得所需文本的文件名。文件名可以通过两种方式获得,即“基于元信息”和“基于内容检索”。两种方法都需要通过检索实现,前者的检索针对非语言特征,后者针对语言特征。

虽然,语料库的格式多种多样,主要有:纯文本格式、SGML格式、XML格式等,但一般都有元信息标注,常见标注格式如下:

CLEC

<ST 3> <SEX ?> <Y ?> <WAY 1>
<TYP 1> <SCH 2703> <DIC ?> <TITLE
Getting to Know the World Outside the Campus>
<SCORE 7> <ID 430124>

SWECCL

书面语:<STU2><ARG17><YEAR06>
<GRADE1><TIMED>

口语:<TEM4><GRADE2><YEAR03>
<GROUP130><TASKTYPE 1 2 3><SEX

T1=F, T2=F, T3=F, F><RANK=3>

北外通用汉英平行语料库

<TEXT_HEAD><TITLE>GoneWith-
TheWind_1/63</TITLE><AUTHOR>Marg-
aretMitchell</AUTHOR><LANGUAGE>
English</LANGUAGE><SL>English</SL>
<TYPE>1</TYPE><FIELD>1</FIELD>
<STYLE>1</STYLE><MODE>1</MODE>
<TIME>2</TIME><PUBLISHER>?</
PUBLISHER><SIZE>36705</SIZE></TEXT_
HEAD>

CROWN 语料库^①

<Country>US</Country>
<Publication_year>2009</Publication_year>
<Publisher>The Christian Science Monitor
</Publisher>
<Author>Jane Arraf</Author>

从上面的各种标注格式可见,元信息可以接排,也可分行排列。一般来说,学习者语料的元信息常包含:学生类型(如ST 3,大学低年级)、性别、入学年份、年级、作文得分(如SCORE 7及RANK=3)等。双语语料库的元信息会包含:当前文本的语言(如<LANGUAGE>English</LANGUAGE>)、源语语言(如<SL>English</SL>)、原作者及译者等。通用型语料库一般需标明出处(如<Publisher>The Christian Science Monitor</Publisher>)和作者(如<Author>Jane Arraf</Author>)等。

上面所列不同格式之间可以互相转换,一般推荐格式为XML格式(见上文北外通用汉英平行语料库和CROWN语料库的标注格式)。元信息的标注可简可繁。目前常见的元信息标注,以BNC最为繁复。BNC采用TEI标注规范,记录多达三十多种信息。语料的元信息多数记录在语料文档的开始部分,因此元信息又称“头部信息”(head、text head或header)。也有的元信息记录在文件名中,比如USWRAR01.txt这个文件名表达的是美国人的书面语议论文第一篇(ARgumentative WRiting of US speakers)。再如,CROWN语料库的命名(如A01A.txt)中,根据BROWN语料库的取样和文件命名原则,第一个A表示“新闻报道”类,01表示新闻报道类44个文本中的第1号,最后一个A表示该语料属于美国英语,而A01B.txt则是对应的英国英语文本。此外,还有一种称之为“分离式标注”(stand-off annotation)的元信息标注法,即相关信息单独存储为一个文件,而不是存放在文档头部。这种标注方法使用较少,因此,本文不做重点讨论。

3.2 利用 Sub-corpus Creator 创建子语料库

本文介绍的子语料库创建方法,主要依靠获得

含有特定元信息以及含特定语言特征的文本的文件名,然后批量拷贝出所需语料。为此,我们编写了软件 Sub-corpus Creator[®],方便研究者自行创建子语料库。下面通过实例简单介绍利用 Sub-corpus Creator 创建子语料库的过程。

如果希望从 SWECCL 的口语语料库(SECCL)中获得谈论“我”和“老师”之间交往的口语叙事语料,可以先选中 SECCL 所在语料的目录(Browse Directory)(见图 1),然后设定检索条件。因为 Sub-corpus Creator 支持正则表达式检索,我们可以检索同时含有 teachers?(?表示前面的 s 可有可无)和 I|me|my(表示 I、me 或 my 三者有其一)的文本,点击 OK 后所有相关文本的文件名即会在软件界面的中间窗口中列出。若希望在结果中得到 2000 年以后学生的英语口语语

料,可在 File names match 下面的检索框里,增加一个限制条件:0\S+,即以 0 开头的文件(SECCL 的口语文本的文件名方式为“年份-组别-序号.txt”,其中的“年份”以 00、01 等两位数表达),重新点击 OK,就可得到筛选后的文件名列表。在这个例子中,1148 个文件中有 687 个符合条件的文本。同时,为便于检查文本是否确实是我们所需要的文本,还可单击中间窗口中的文本名,从而预览文本内容。

按同样方法,可以得到 2000 年以前的相关语料。如果按类似方法将 SWECCL 语料中每年的相关语料分别抽取出来,甚至可以分析十年中国英语专业大学生谈论自己同老师之间的交往时所用语言存在怎样的变化。

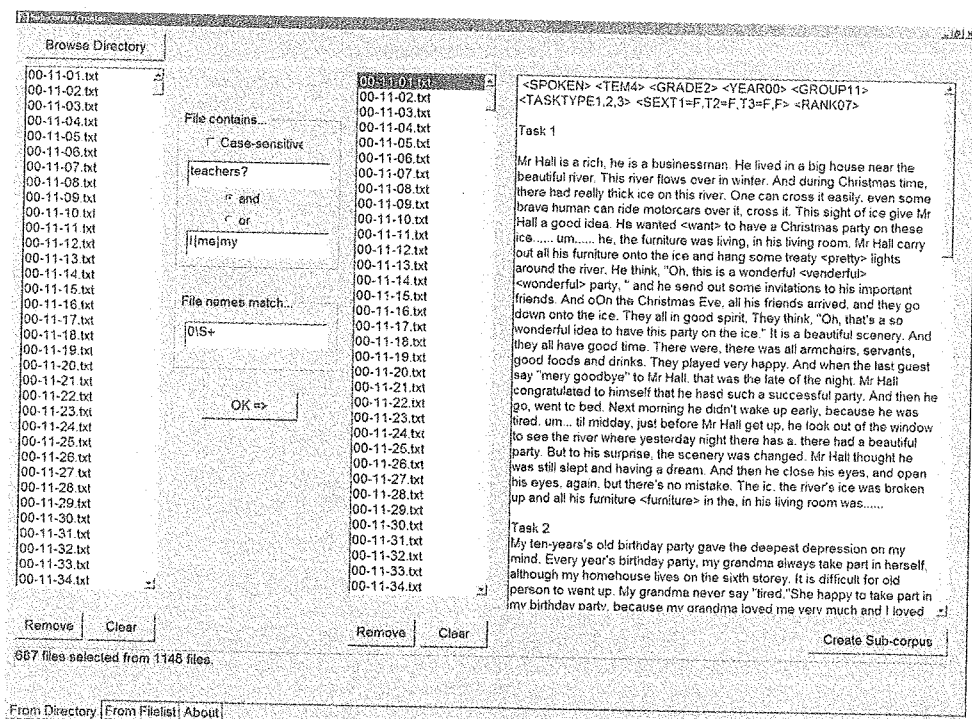


图 1 Sub-corpus Creator 主界面

上面这个例子主要是通过检索文内信息,并结合文件名中所含的年份信息抽取出子语料库。研究中,还会遇到这样的案例:有 120 篇学生作文(这 120 篇作文成绩单单独存储在一个文件中),若要得到 120 篇作文中成绩靠前 25%(即前 30 名)和成绩靠后 25%(即后 30 名)学生的作文进行对比,可以先将学生成绩降序排列,选取前 30 名学生的文件名。然后,将文件名存为一个纯文本(*.txt)文件,比如,可以将其命名为 top30.txt。同理,可以创建一个成绩靠后学生的 poor30.txt 文件。

Sub-corpus Creator 还设计了另一种利用已准备好的文件名列表创建子语料库的方法。如果已有所需语料的文件名列表,则可利用 Sub-corpus Creator 左下方的第 2 个选项卡 From Filelist(见图 2)进行

子语料库创建。首先,通过 Open a Filelist 读入准备好的文件名列表(如 top30.txt 和 poor30.txt),点击 OK;再点击 Create Sub-corpus,按弹出窗口提示,为子语料库创建一个新的文件夹,命名后,就可得到成绩靠前 25%(即前 30 名)和成绩靠后 25%(即后 30 名)学生作文的两个子语料库。需要提醒的是,通过 Sub-corpus Creator 软件加载文件列表时,列表中的文件名应包含完整路径。

基于给定文件名列表创建子库这种方法,还特别适合从 BNC 中抽取特定语料。比如,可以利用在线版 BNC Indexer[®]得到所有男性说话人和女性说话人的文本的文件名列表。通过导入这两个文件名列表,就可以从 BNC 语料库中生成两个子语料库,一个是男性话语子库,一个是女性话语子库。

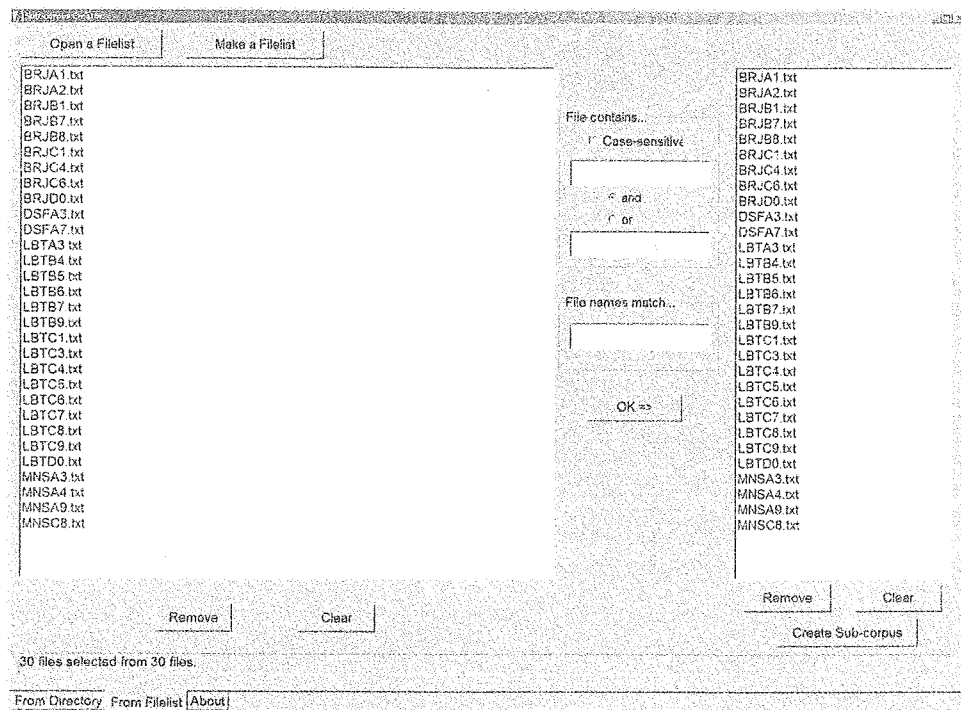


图2 Sub-corpus Creator 基于给定文件名列建库界面

第2节谈到的研究选题,绝大多数可以在 Sub-corpus Creator 中通过检索元信息或语料内容得到相应的子语料库。这里补充谈一下,有时面对得到的大量文本,需要再做随机抽样。比如得到四级作文 2000 篇,由于无法分析上千篇文本,于是必须对这 2000 篇文件进行随机取样。如果只需随机抽取其中的 100 篇,则可以利用 Concordance Randomizer^④ 对 2000 个文本的文件列表进行随机取样,即可实现这一目的。

除了随机取样,如果语料文本命名有规律,比如文件名以 0-9 结尾的话,在 Sub-corpus Creator 的 File names match 框里输入 \S+5\ .txt, 就可以得到所有文件名为 * 5. txt 的文本。这样就实现了通过分层取样(stratified sampling)的方法建立子语料库。

对于双语语料库,可以采用类似的操作。以“北外通用汉英平行语料库”为例,如果要聚焦在语料中的小说部分汉译英的译文语言特点上,则可通过检索语料库中含有 <LANGUAGE>English</LANGUAGE><SL>Chinese</SL> 的文件,按上面的步骤即可获得相应的文本。同样,还可依照类似的方法获得中文及英文的原生语言与翻译语言的子语料库,以便进行翻译共性研究。

4. 结语

本文介绍的子语料库创建方法,虽然主要以中国学习者英语对比语料和英汉语对比语料为例,但显然这一方法广泛适用于各类语料库的子语料库生成。需要提醒的是,为方便创建子语料库,拓展对比

研究的层面,语料库建库之初,规范统一详细的元信息标注十分关键。同时,语料库文本的命名应尽可能有规律,尽可能记录元信息。本文着重介绍的是创建语料库的思路,而非软件的使用。任何一款软件都不可能满足所有的研究需要。只有掌握了方法,打开了思路,才有可能得到更多的研究选题。

附注:

- ① CROWN 语料库是许家金、梁茂成主持创建的一个 200 万词级的英语语料库。
- ② Sub-corpus Creator 下载网址: <http://www.fleric.org.cn/pub/scc.rar>.
- ③ BNC Indexer 网址: <http://ucrel.lancs.ac.uk/bncindex/form.html>.
- ④ Concordance Randomizer 下载网址: <http://ishare.iask.sina.com.cn/f/13929531.html>.

参考文献

- Baker, M. 2000. Towards a methodology for investigating the style of a literary translator [J]. *Target* 12(2): 241-266.
- Granger, S. 1998. The computer learner corpus: A versatile new source of data for SLA research [A]. In S. Granger (ed.). *Learner English on Computer* [C]. New York: Longman, 3-18.
- Rayson, P., G. Leech, & M. Hodges. 1997. Social differentiation in the use of English vocabulary: Some analyses of the conversational component of the British National Corpus [J]. *International Journal of Corpus Linguistics* 2(1): 133-152.
- 王克非. 2006. 语料库翻译学——新研究范式[J]. 中国外语 (3): 8-9.

(责任编辑 甄凤超)