

CSSCI语言学类来源期刊
中文核心期刊要目总览(第六版)入编期刊
外语教育技术专业期刊

外语电化教学

COMPUTER-ASSISTED
FOREIGN LANGUAGE
EDUCATION



中华人民共和国教育部主管
上海外国语大学主办

CAFLE

外语电化教学

双月刊

2014年 第5期

(总第159期)

主 编 吴友富

副主编 陈坚林 胡加圣

编 委 (按汉语拼音顺序)

陈坚林 程东元 顾佩娅

顾曰国 何高大 胡壮麟

金 莉 吴友富 杨惠中

杨永林 张祖忻 祝智庭

编辑部主任 陈坚林

编辑部副主任 胡加圣

责任校对 柳华妮

编 务 王明德 贺诗熠

本期责任编辑 柳华妮

特 约 编 审 孟庆和

主管单位 中华人民共和国教育部

主办单位 上海外国语大学

出版单位 《外语电化教学》编辑部

联合出版单位 上海外语音像出版社

上海外语电子出版社

中国教育技术协会外语

专业委员会

地 址 上海市大连西路550号

366信箱 邮编: 200083

电 话 (021)35373318

E-mail wydhjx204@163.com

网 址 http://wydh.qikan.com

印 刷 上海新华印刷有限公司

中 国 标 准 ISSN 1001-5795

连续出版物号 CN31-1036/G4

广告经营许可证

许可证号 3101094000038

国外总发行:

中国国际图书贸易集团有限公司

北京市海淀区车公庄西路35号

国外代号 BM4383

国内邮发 代号 4-378

定 价 12.00元

目 录

语言学研究方法特约稿

- 3 民族志研究方法在语言测试研究中的应用 邹 申

网络教育技术研究

- 10 基于网络的第四代语料库分析工具 CQPweb 及应用实例 许家金 吴良平
- 16 英语语法引导式数据驱动学习平台的构建与成效 罗 凌 温善毅 朱永红
- 22 TPACK 框架下《高级英语》课程微群与课堂混合教学的行动研究 钟 维
- 28 基于 Wiki 的项目式协作性英语教学案例研究 杨东杰 王维倩 宁大利
- 34 基于微信的非英语专业大学生英语听说学习诉求的调查研究 柴阳丽

外语高校学科转型研究专稿

- 40 全球化背景下外语类高校人才培养模式研究
——以东京外国语大学为例 王志强 马 磊 王 寰

外语教学理论与实践

- 49 语言哲学与外语教学:回顾与反思——中国外语教学系列研究之一 曾建松 李洪儒
- 57 二语学术写作中的引用特征及与写作质量的关系 马 蓉 秦晓晴
- 63 EFL 读写结合测试任务中受试使用原文策略的实证研究 吴 越
- 70 影响工科学生英语听说能力的相关因素研究 马冬虹 刘 昕
- 75 大学英语听力教学研究现状与问题分析
——基于国内 CSSCI 期刊论文的分析(1994-2013) 柯应根 董 艳

会 讯

- 79 慕课(MOOC)时代的高等外语教学学术研讨会综述 向 平 卢东民

基于网络的第四代语料库分析工具 CQPweb 及应用实例

许家金¹, 吴良平²

(1. 北京外国语大学 中国外语教育研究中心, 北京, 100089; 2. 湖南商学院 国际教育学院, 湖南长沙, 410205 / 北京外国语大学 中国外语教育研究中心, 北京, 100089)

摘 要: 基于语料库已成为语言研究中广泛采用的研究方法。然而这一方法较依赖计算机手段, 这为运用该方法设置了技术门槛。本文介绍的基于网络的第四代语料库分析工具 CQPweb, 使得无技术背景的研究人员也能如浏览网页一般轻松使用语料库。CQPweb 具备词表生成、索引分析、词语搭配计算、主题词分析等当前主流单机版语料库工具的几乎所有功能。它更能按语料库预设的体裁、说话人性别、语言水平等分别检索和呈现结果。对语料库专业人士, CQPweb 还具有强大的高级检索功能。CQPweb 无疑是网络时代从事语料库研究的利器。

关键词: CQPweb; BFSU CQPweb; 第四代语料库分析工具; 多语种语料库; 语料库

中图分类号: H319.3

文献标识码: A

文章编号: 1001-5795(2014)05-0010-0006

如今语言学研究中, 与语料库相关的学问, 即便不能称为显学, 也是不折不扣的热门。然而, 由于语料库分析涉及较多计算机操作技术方面的要求, 这无形中将相当一部分研究人员拒之于门外。基于网络的第四代语料库分析工具(以下简称“第四代语料库工具”)的逐步兴起, 使得缺少技术背景的研究人员也能如平日浏览网页一般轻松使用语料库, 在降低语料库使用技术门槛的同时, 它也允许语料库专业人员深入、充分、便捷地挖掘语料库资源。本文将以 CQPweb(特别是 BFSU CQPweb)为例, 介绍如何利用第四代语料库工具开展语料库研究。

1 第四代语料库分析工具的设计特征

1.1 概述

受“网络语料库”(WaC, Web as Corpus)和“云计算”(cloud computing)等思维影响, 部分学者相继开发

出在线检索语料库工具, 如 WebCorp(A. Renouf)、Phrase in English(W. Fletcher)、BYU 系统(M. Davies)、Sketch Engine(A. Kilgariff)、BNCweb(S. Hoffmann 等)和 CQPweb(A. Hardie)。McEnery & Hardie(2012: 37-48)在对语料库分析工具进行时代划分时, 将这些基于浏览器-服务器(B/S)模式的语料库检索工具归为第四代语料库分析工具。

我们认为第四代语料库工具有两大设计特征:

(1) 语料库与分析工具合二为一。在线语料库工具通常将语料库文本按特定格式建成索引, 存储于服务器。用户在任何地方只要联网, 即可进行语料库分析操作。而且, 使用第四代语料库分析工具几乎没有技术门槛。

(2) 采用索引技术(indexing), 响应时间快。第四代语料库工具通常利用关系型数据库(如 SQL Server、MySQL)或专用型语料库检索系统(如 CQP、SARA)预

作者简介: 许家金: 男, 博士, 教授, 博士生导师。研究方向: 话语分析、语言对比研究、语料库语言学。

吴良平: 男, 博士生, 讲师。研究方向: 语料库语言学。

收稿日期: 2014.03.20

基金项目: 本文得到国家社科基金“基于双语语料库的汉语复杂动词结构英译研究”(项目编号: 12CYY060)和教育部“新世纪优秀人才支持计划”(项目编号: NCET-12-0790)的资助。

先对话料库进行数据索引,这使得对上亿词级的语料库的复杂检索通常也只需几秒或更短时间即可完成,为开展大规模语料库研究提供了可能。

1.2 与第三代工具的关系:继承与发展

第四代语料库工具的出现,并不意味着对第三代工具的完全摒弃,而是给用户在网络时代多一种选择。总体来说,第四代语料库工具可完成第三代工具的几乎所有功能。

当前主流语料库工具属于第三代,其中以 Word-Smith、AntConc、MonoConc、Xaira 和 PowerConc (许家金、贾云龙,2013) 等为代表。索引分析 (concordancing)、词频表 (wordlist)、主题词 (keyword) 是这些工具的几大核心功能。第四代语料库工具部分或全部实现了这些功能,如 WebCorp 可进行索引分析和生成词频表,而 Sketch Engine、BNCweb、CQPweb 等则实现了第三代软件的大部分功能,在某些方面还有所创新。

第四代语料库工具区别于第三代工具的根本特征在于基于网络的客户端-服务器 (C/S) 交互模式 (McEnery & Hardie, 2012: 44)。由于利用了网络,客户端计算机的处理能力和操作系统变得不再重要,语料库检索实现了“瘦客户” (thin client) 和“平台无关” (platform independent)。普通用户可使用个人电脑、平板电脑或智能手机随时随地进行语料库检索而无需知道语料库的具体存储位置,其操作系统可以是 Windows, Mac OS, Linux 或其他平台,这大大扩展了第四代语料库工具的用户基数。

这种软件架构克服了 PC 处理能力的局限,主要运算集中交由服务器端解决,软件开发者只需专注于服务器端软件的开发,用户刷新页面即可得到最新版本的软件。同时,语料库也可以即时更新,用户能够快速查询升级后的语料库而不必更换软硬件设施,降低

了语料库的维护成本。

2 第四代语料库分析工具的功能对比

BYU 系统、BNCweb、SketchEngine、Wmatrix 等代表性的第四代语料库工具和 CQPweb 的功能对比,详见表 1。

通过对比可知,第四代语料库工具中 CQPweb 所能实现的功能最全,其基本功能可概括如下:

(1) 在线生成词频表 (frequency list);

(2) 查询 (query) 字词、语言结构等,以获取大量语言实例或相应结构的出现频次,并可按语体、年代、章节、学生语言水平级别、写作题材等分别呈现查询结果;

(3) 计算特定词语在语料库中的典型搭配 (collocation);

(4) 计算语料库的主题词 (keywords)。

与其他工具相比,CQPweb 优势明显。相比广泛使用的 BYU 系统,CQPweb 支持更复杂的高级检索,与 SketchEngine 和 Wmatrix 比较,CQPweb 用户界面更友好。值得一提的是,SketchEngine 的后台程序 Manatee 是 CQPweb 的后台程序 CQP 的克隆,两者在查询语法上有很多相同之处 (Rychly, 2007; Evert & Hardie, 2011); 而 BYU 系统的开发也受到了 CQP 数据结构的启发 (Davies, 2005)。

3 CQPweb: 以 BFSU CQPweb 为例

3.1 背景

1994 年德国斯图加特大学自然语言处理研究所 (IMS) 因计算词典学研究开发了 CQP 语料库查询工具 (Christ, 1994), 不久即被欧洲语言工程标准专家咨询小组 (EAGLES) 采用,同时它也是著名的 Xkwic 的后台程序。1998 年该工具的维护由 Stefan Evert 等人接手,并转向以 Linux 为主要开发平台,2005 年开始向

表 1 主要第四代语料库工具功能对比

特征	CQPweb	BYU 系统	BNCweb	SketchEngine	Wmatrix
免费使用	是	是	是	否 *	否 *
开源	是 **	否	是 **	否	否
支持正则表达式	是	否	是	是	否
支持任意标注格式	是	是	否	是	否
允许用户上传本地语料库	否	否	否	是 ***	是 ***
支持三大语料库核心功能	是	否	是	否	是
支持平行语料库	是	否	否	是	否

* 一个月内免费使用; ** 基于 GPL 协议; *** 对上传文件大小有限制。

公众开放源代码 (Evert, 2008)。次年 Hoffmann & Evert 将 BNCweb 和 CQP 结合, 替换掉 BNCweb 原有的后台程序 SARA, 推出了基于 CQP 的 BNCweb (Hoffmann et al., 2008: xv), 使得 BNCweb 的用户界面更加友好、操作更加灵活。但该系统只能检索 BNC, 不适用于其他语料库的检索。

为突破这一限制, 兰卡斯特大学的 Andrew Hardie 以 BNCweb 为模板, 用 PHP 语言彻底改写了 BNCweb, 从而于 2008 年 1 月推出了 CQPweb 的雏形 (其命名也反映了其 CQP + BNCweb 的渊源)。至 2013 年 12 月版本号已累积至 3.0.16, 目前, CQPweb 全面支持 UTF-8, 因此理论上可处理世界上任何语言的语料库。

促成开发 CQPweb 的另一重要原因是困扰语料库研究的版权问题。囿于版权保护相关法律法规, 语料库研究者往往不能将创建的语料库与其他研究人员共享。这意味着他们辛勤劳动的成果仅能供少数人使用, 更糟糕的是因此减少了其他研究者复制和验证相关研究的机会, 进而影响到研究成果的推广 (McEnery & Hardie, 2012: 44)。CQPweb 将语料库文本按特定格式建成索引后存储在服务器上, 并不把语料库整体复制到第三方计算机, 搜索结果只显示有限的上下文语境, 和 Mark Davies 的 BYU 系统一样, 用户无法根据检索结果重构原语料库 (Hardie, 2012)。这种做法和 Google 等搜索引擎应对版权法的做法一致, 是目前商业界和学术界的通行做法。

3.2 应用现状

CQPweb 采用 GPL 协议对公众开放源代码, 允许

任何第三方对其使用、修改和分发。自发布以来, 由于其能处理多语种语料库的便利, 世界上很多研究机构将其作为语料库查询平台。至 2012 年 11 月, 已有 10 多个研究机构采用了 CQPweb。从已上线可供查询的语言来看, 涵盖印欧、汉藏、阿尔泰等世界几大主要语系, 详见表 2。

从表 2 可见, 基于 CQPweb 的语料库既有大型通用语料库, 也有规模较小的专题语料库, 适合不同研究需求的语料库处理。

3.3 运行体系与流程

理解 CQPweb 的运行体系有利于更充分地对其利用和研究。从逻辑结构来看, CQPweb 分前端、后台两部分, 前端为呈现在浏览器的用户界面, 接受用户查询输入等, 与用户进行交互; 后台由 CWB (Evert & Hardie, 2011) 和 MySQL 数据库构成, 负责建立索引和向用户界面输出分析结果。其中 CWB 包含创建索引的辅助程序和用于查询的 CQP 程序。

一个完整的查询过程可分解成几个步骤: 首先, 用户通过浏览器输入查询语句, CQPweb 收到该语句后送交 CQP (Corpus Query Processor) 或 CEQL (The Common Elementary Query Language) 进行解析, 然后从 CWB 预先处理好的索引文件中提取符合该查询条件的结果返回给用户界面。用户若需要对查询结果进行后续处理, 如计算搭配、查询词频体裁分布等, CQPweb 则利用 MySQL 数据库, 对存储其中的查询结果进行排序和计算, 并将结果返回给用户。以上过程可以图 1 示意。

表 2 CQPweb 使用情况一览表 (不完全统计)

大洲	国家	研究机构或项目	语种特色	网址
亚洲	中国	北京外国语大学	多语种, 英语、汉语、德语、日语、俄语、阿拉伯语、冰岛语等 7 个语种, 33 个语料库	124.193.83.252/cqp 用户名: test; 密码: test
		台湾师范大学	现代汉语 (繁体)	140.122.83.250/cqpweb
欧洲	英国	Lancaster University	多语种, 40 多个语料库	cqpweb.lancs.ac.uk
	德国	Universität Duisburg-Essen	德语议会语料库	polmine.sowi.uni-due.de/cwb
	瑞士	Universität Zürich	瑞士立法语料库	server.linguistik.uzh.ch/cqpweb
	葡萄牙	University of Lisbon	当代葡萄牙语语料库	alfclul.chul.ul.pt/CQPweb
	西班牙	Universitat Autònoma de Barcelona	当代西班牙语语料库	sfncorpora.uab.es/CQPweb/cea
	土耳其 *	TS Corpus	土耳其语语料库	tscorpus.com/eng
美洲	加拿大	BMH Linguistics	生物医学英语语料库	cqpweb.wetware.ca

* 土耳其横跨欧亚两洲

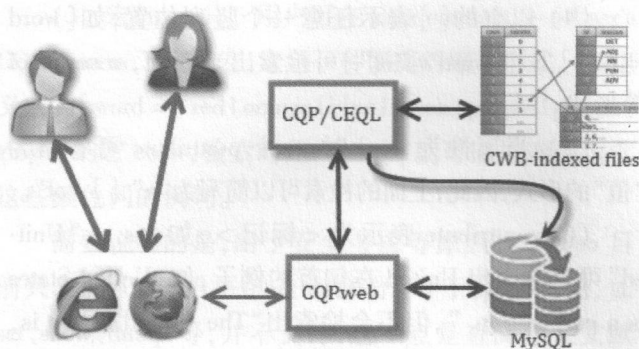


图1 CQPweb 运行体系与流程

需要指出, CQP 是一种专用型语料库查询语言, 由编译器写作程序 Flex 和 Bison 写成, 具有目前一般 SQL 语言所不具备的正则表达式查询能力, 但其排序功能不强, 所以 CQPweb 将其与 MySQL 数据库结合, 扬长避短。

3.4 语料格式规范

一个 CQPweb 语料库一般由两部分构成: 语料库文本和对应的元信息 (metadata)。

元信息文件是 CQPweb 限定条件查询 (Restricted query) 的基础。限定条件查询即选定一个或多个限制条件 (如, 语体、年代、章节、写作题材等) 进行针对性的检索。限定条件的有无、多寡, 取决于语料文本的元信息。因此, 创建语料库时, 应尽可能详细地提取文本包含的社会语言学信息, 并用元信息记录下来, 从而可以丰富研究的层面和深度。元信息可以 XML 格式存储在文本头部, 也可在文本之外单独存储。单独存储时, 表示为制表符分隔的文本。表 3 为单独存储的 Crown 语料库 (Xu & Liang, 2013) 的元信息片段。

表 3 CQPweb 语料库元信息样例

文件名	语体	美/英	出处	年份
A05AA	Press	A	The New York Times	2009
A06AA	Press	A	Los Angeles Times	2009
B18BA	Press	B	Guardian	2010
B18BB	Press	B	Independent	2008
B18BC	Press	B	Observer	2009

CQPweb 要求文本符合一定的格式规范, 这一规范继承自 CWB, 是一行一单词 (token) 的竖排格式。了解这一结构对在 CQPweb 平台上安装和查询语料库都很有意义。设想有一微型语料库文本包含如下文

字: “Welcome to the world of CQPweb.” 经 CLAWS 词性赋码后结果为 “< s > Welcome_VV0 to_II the_AT world_NN1 of_IO CQPweb_NP1 ._. </s>”。其 CQPweb 数据格式可用图 2 表示。

#word	pos	lemma	p-attribute
<text id="sample">			
<s>			
Welcome	VV0	welcome	s-attribute
to	II	to	
the	AT	the	
world	NN1	world	
of	IO	of	
CQPweb	NP1	CQPweb	
.	.	.	
</s>			
</text>			

图 2 CQPweb 语料库文本数据格式图

CQPweb 的数据格式中, 每行是一个单词, 单词、词性 (pos) 及该单词的其他属性 (如, lemma, 原形词) 间用制表符隔开, 统称 p-attribute (positional attribute)。而以 XML 方式标记的文档及文档内部结构属性统称 s-attribute (structural attribute)。任何其他格式的语料库若要经 CQPweb 进行检索, 须通过预处理满足这一格式。

同时, CQPweb 检索语法也与这一结构紧密相关, 只是需要将竖排格式理解成横排格式。以下 3.5.2 节将详述这一点。

3.5 案例: BFSU CQPweb

BFSU CQPweb 是利用开源的 CQPweb 架设的多语种在线语料库检索平台。目前 BFSU CQPweb 平台上已安装英语、汉语、德语、日语、俄语、阿拉伯语、冰岛语等 7 个语种共 35 个语料库。该平台可按规定的文本格式不断增加语料库的种类和数量。

3.5.1 简单检索

CQPweb 的检索可分为简单查询模式 (Simple query mode) 和复杂检索模式 (CQP syntax)。简单查询模式下, 可输入单词、短语等进行检索, 如通过 BFSU CQPweb 在 Crown 语料库中检索单词 “the”。简单模式本质上使用的是 CEQL 语法, 所以除单词、短语外, 对经 CLAWS 词性标注过的语料库还可以输入 “_JJ *”

来查询所有形容词。有关 CEQL 的详细用法可参考 Hoffmann et al. (2008)。BFSU CQPweb 对检索结果还可进行包括搭配计算等在内的后续操作(见表4),具体操作步骤可参阅许家金(2014)编写的使用手册,此处不赘述。

表4 查询结果后续操作分项功能表

New query	新查询
Thin	查询结果随机抽样
Frequency breakdown	频数分解
Distribution	查询结果的分布展示
Sort	查询结果排序设定
Collocations	搭配计算
Download	下载保存查询结果
Categorise	对检索结果手工分类标注

3.5.2 复杂检索

CQPweb 支持基于 CQP 语法(Evert et al., 2010)进行复杂检索。CQP 语法是 CQPweb 检索的精髓,它可以实现许家金、熊文新(2009)所描述的各种类联接的高级检索方式,这与 CQPweb 的数据结构紧密相关。仍以前面提到的语料文本为例,在检索时可按检索的顺序性将其从竖排格式变成横排格式以方便理解。

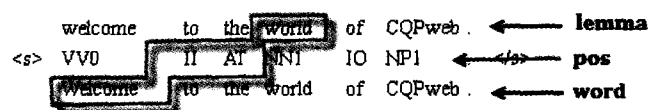


图3 CQP 语法逻辑结构图

CQP 的基本检索单元为 p-attribute (如图3中的 word、pos、lemma) 和 s-attribute (如表示句首位置的 <s> 和句尾位置的 </s>)。其检索模型可按如下几个要点来理解:

(1) 检索是线性的,从左至右(若是竖排格式则是从上至下),也可走“之”字形路径。

(2) 一个 p-attribute 对应图3中的一行(<s> 和 </s> 除外),检索时同一列的 p-attribute 作为一个检索单元,单元分割符为 []。

(3) 一个 p-attribute 表示为 [属性 = “值”] 的形式,同一列的 p-attribute 可用布尔运算符(和“&”、或“|”、否“!”)进行合成运算,如 [word = “lights” & pos! = “NN”] 表示查询所有竖列处词性不是名词的单词 lights。

(4) 以空的 [] 表示任意一个竖列位置,如 [word = “a”] [] [word = “of”] 可检索出 a lot of, a series of 等形式。

(5) 当属性为 word 时,一个 p-attribute 可省略为“值”的形式,因此,上面的检索可以简写为“a” [] “of”。

(6) s-attribute 表示为 < 标记 >, 如 < s > “United” 可以检索出 United 在句首的例子,如“United States is a corporation.”,但不会检索出“The United States is, indeed, a republic, not a democracy.”。

(7) 组合检索:查询语句 < s > “Welcome” [pos = “II”] [pos = “AT”] 可以检索出“Welcome to the Edge of the Wilderness.”,“Welcome to no man’s land.” 这样的句子。

同时,CQP 语法的另一重要特点是同时支持在检索单元内(即 [] 内)和检索单元间(即 [] 间)两个层面进行正则表达式查询。例如 [word = “bea. *”] 表示检索以 bea 开头的任何单词,而 [word = “as”] [] { 1, 3 } [word = “as”] 则表示检索 as...as 结构且中间有 1 至 3 个单词的情形。

表5 CQP 正则表达式简明语法

符号	意义
.	点号匹配任意单个字符
[]	字符集,如 [1-5], 表示 1, 2, 3, 4, 5
?, *, +, n, { m, n }	重复个数, 分别表示 0 或 1, 0 或更多, 1 或更多, n 个, m 到 n 个
()	群组符号, 如 (ha)? 表示 ha 或 haha
	选择符号, 如 be was 表示 be 或者 was

此外,CQP 语法还有几个特殊的扩展规则,能实现常规正则表达式查询所不能实现的一些功能。

(1) 标签(label)。标签可以限定远距离词项间的关系,以 label: 表示。例如 a: [pos = “N. *”] “are” [pos = “N. *” & word = a. word] 表示第一个检索词和第三个检索词必须词形相同,中间为 are, 能检索到 women are women, concepts are concepts 等同义反复结构。

(2) 零宽度条件判断(zero-width assertion)。以 [::] 表示, [::] 内匹配结果不作为结果返回,仅作判断条件。例如, [pos = “VB. *”] “that” [: pos! = “JJ. *” | “N. *” :] 表示检索 be 动词后的补语从句。

(3) 宏(macro)。用户可在 User controls -> User's CQP macros 下定义宏,这是一种便捷的缩写形式。如定义宏为[word = "red|orange|blue|green|purple|purple|grey"]的宏 color,检索时用/color[]就可以检索到包含这些颜色词的例句。

需要说明的是,出于安全考虑等原因,CQPweb 目前只支持 CQP 语法的查询部分,对于命令动词,如 set,show,dump 等,并不支持,这一点是查阅相关文献时要注意的。

3.6 CQPweb 的局限性

CQPweb 安装架设相对复杂,说明文档欠缺,因而影响了其使用和推广。好在安装工作可由技术人员完成,无需一般用户专门学习。本文也是对其说明文档进行补充的一个尝试。

CQPweb 不支持用户上传语料库,只能由系统管理员统一管理。由于数据索引过程会耗费大量计算资源,影响系统运行,绝大多数第四代语料库分析工具(包括 CQPweb)都是采用预先索引好的语料库供用户查询,体现为语料库与工具密不可分。部分系统允许用户上传自己的语料库,但对上传的文件大小通常有严格限制。

CQPweb 的另一个问题是,因为用户无法获得语料库文本,一些专门化的研究(如 Biber 倡导的多维度分析)以及个性化的研究需求(如基于相对完整篇章的定性分析)都无法实现。然而,终究瑕不掩瑜,CQPweb 已然是一款十分出色而全面的大众化语料库工具。

4 结语

CQPweb 优势明显:它基于广域网,语料可以实时更新;利用信息检索的索引技术,响应速度快。不足在于:安装难度大,不支持用户上传自己的数据。综合看来,未来很长一段时间内,第三代和第四代语料库工具将并存发展(许家金、贾云龙,2013)。

CQPweb 采用开源模式,一般用户、高级用户和程序员都可以参与进来,可以分别提出功能需求、帮助发现问题或直接贡献代码,这有利于集思广益和充分利用人力资源,前景看好。可以预见,CQPweb 的推出使得语料库工具基本功能从头开发的需求下降,但软件定制化的地位将日益突出。□

参 考 文 献

- [1] Christ, O. A modular and flexible architecture for an integrated corpus query system [A]. In *Papers in Computational Lexicography* [C]. Budapest: COMPLEX, 1994.
- [2] Davies, M. The advantage of using relational databases for large corpora: speed, advanced queries, and unlimited annotation [J]. *International Journal of Corpus Linguistics*, 2005(3).
- [3] Evert, S. & Hardie. A. Twenty-first century Corpus Workbench: Updating a query architecture for the new millennium [A]. In *Proceedings of the Corpus Linguistics 2011 Conference* [C]. Birmingham: University of Birmingham, 2011.
- [4] Evert, S. & The OCWB Development Team. CQP Query Language Tutorial [OL]. http://cwb.sourceforge.net/files/CQP_Tutorial.pdf, accessed 11/10/2012, 2010.
- [5] Evert, S. Inside the IMS Corpus Workbench [R]. Presentation at the IULA, Universitat Pompeu Fabra, Barcelona, Spain [OL]. http://cwb.sourceforge.net/files/Evert2008_InsideCWB.pdf, accessed 11/10/2012, 2008.
- [6] Hardie, A. CQPweb—Combining power, flexibility and usability in a corpus analysis tool [J]. *International Journal of Corpus Linguistics*, 2012(3).
- [7] Hoffmann, S., Evert S., Smith N., Lee D. & Prytz. Y. *Corpus Linguistics with BNCweb—A Practical Guide* [M]. Bern: Peter Lang, 2008.
- [8] McEnery, T. & Hardie. A. *Corpus Linguistics: Method, Theory and Practice* [M]. Cambridge: Cambridge University Press, 2012.
- [9] Rychly, P. Manatee/Bonito—A modular corpus manager [A]. In RASLAN (ed.). *Recent Advances in Slavonic Natural Language Processing* [C]. Brno: Masaryk University, 2007.
- [10] Xu, Jiajin & Liang Maocheng. A tale of two C's: Comparing English varieties with Crown and CLOB (The 2009 Brown family corpora) [J]. *ICAME Journal*, 2013(37).
- [11] 许家金,贾云龙.基于 R-gram 的语料库分析软件 PowerConc 的设计与开发[J]. *外语电化教学*, 2013(1).
- [12] 许家金,熊文新.基于学习者英语语料库的类联接研究:概念、方法及例析[J]. *外语电化教学*, 2009(3).
- [13] 许家金.多语种在线语料库检索平台 BFSU CQPweb 使用简明手册[OL]. http://www.bfsu-corpus.org/static/corpus_tools/CQPweb_guide.pdf, 2014 年 8 月 10 日读取, 2014.

(下转第 56 页)

(上接第 15 页)

Web-based Fourth Generation Corpus Analysis Tools and the BFSU CQPweb Case

XU Jia-jin¹, WU Liang-ping²

(1. National Research Center for Foreign Language Education, BFSU, Beijing 100089, China; 2. School of International Studies, Hunan University of Commerce, Changsha, Hunan 410205, China / National Research Center for Foreign Language Education, BFSU, Beijing 100089)

Abstract: This paper describes CQPweb, a web-based fourth generation corpus analysis tool, which enables technically less competent researchers to exploit corpora just like browsing web pages. CQPweb accommodates almost all the functionalities characteristic of mainstream PC-based corpus tools: wordlist generation, concordancing, collocation extraction, keyword analysis, and so forth. Furthermore, it is capable of searching corpora and displaying results according to such metadata as pre-defined genres, speakers' gender and/or language proficiency. For corpus specialists, CQPweb also offers more advanced search options to qualify itself as an exceptionally powerful tool for corpus study in the Internet age.

Key words: CQPweb; BFSU CQPweb; The Fourth Generation Corpus Analysis Tools; Multilingual Corpora; Corpus