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

1
2013
CAFLE

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
2013年 第1期
（总第149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：上海出版印刷有限公司
国际标准刊号 ISSN 1001—5795
国内统一刊号 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

大学外语教学改革探索
——课程模式和评价研究

外语教育技术理论与实践

语言技术研究

会讯

期刊基本参数：CN31-1036/G4*1981*b*A4*80*zh*P*￥12.00*8000*13*2013-1

语言技术研究

基于 R-gram 的语料库分析软件 PowerConc 的设计与开发

许家金，贾云龙
（北京外国语大学 中国外语教育研究中心，北京 100089）

摘 要： 在继承以往语料库分析软件优点的基础上，本研究开发了具有独立知识产权的 PowerConc 语料库分析工具。PowerConc 对传统的词汇索引、词表生成、主题词计算等功能进行了重构、扩展和优化。整个软件以基于正则表达式（regular expressions）的 N 元组（N-gram）为基础。二者的有机结合即本文所提出的 R-gram。R-gram 这一概念大大增强了检索和匹配的灵活性。同时我们设计了兼容正则表达式的简易输入语法——Smart Input，降低了用户使用的难度，提高了软件的易用性。PowerConc 软件基于面向对象的思想开发，核心功能被封装在不同的类中，与界面分离，具有很好的扩展性和可维护性。PowerConc 的开发将有效促进语料库语言学研究的开展。

关键词： 语料库分析工具；PowerConc 软件；R-gram；语料库
中图分类号： H319.3 **文献标识码：** A **文章编号：** 1001-5795（2013）01-0057-0006

1 语料库分析软件的开发背景

1.1 引子

语料库研究需要对大量文本进行计算机分析，其中语料库分析软件的作用十分关键，且很大程度上决定着研究数据的准确性和可靠性。没有良好的语料库工具支持，语料库研究便难以有效开展。目前常用的语料库分析工具有：Mike Scott 设计的 WordSmith Tools（以下简称 WordSmith）、Laurence Anthony 设计的 AntConc、Michael Barlow 设计的 MonoConc Pro 和 R. Watt 设计的 Concordance 等。其中 WordSmith 功能最全，学界认可度最高。其他软件有的是 WordSmith 的（部分）重写，有的只能实现 WordSmith 的少量功能。综合来看，这些软件通常包含词汇索引（concordancing）、词表生成（word list）、主题词计算（keywords）等功能，但在统计和搭配计算等方面，对正则表达式（regular expressions）和 N 元组（N-gram）的支持，易用性和计算效率方面还有待提高。

近些年来，国内外语料库建设蓬勃发展，但语料库分析软件的开发却相对滞后，一方面新工具开发较少，同时，原有语料库分析工具升级缓慢，在核心功能上改进不大。本研究希望结合语料库语言学近年的发展，开发出与之相适应的分析工具。

1.2 语料库分析软件发展概述

语料库是指按一定原则取样获得的大规模电子文本汇集（Sinclair, 1991；Hunston, 2002；Baker, 2006）。语料库规模通常很大，因此需要借助计算机软件来辅助分析。近半个世纪以来，语料库分析工具层出不穷，数量、种类不断增加。

语料库软件包括：词汇索引工具（concordancer）、自动和手工标注工具（词性标注、句法标注、语义标注、语用标注等）、文本整理工具（文本格式转换、文本编

作者简介：许家金：男，博士，副教授。研究方向：话语分析、语料库语言学。
贾云龙：男，硕士。研究方向：语料库语言学、教育技术。

收稿日期：2012-08-18
基金项目：本文的撰写得到国家社科基金项目“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。

码转换)、口语转写工具、统计分析工具等。语料库分析工具中最常用的是索引工具。一般所谓的通用型语料库分析工具即指索引工具。最早的计算机索引工具由 Roberto Busa 于 1951 年开发(McEnery & Hardie, 2012:37)。当时的索引工具只能提供索引行语言实例。后来索引工具的功能得到很大扩展,但名称仍然叫做索引工具。现在的通用型索引工具,通常至少包括生成索引行和词表两大功能。

根据 McEnery & Hardie(2012:37-48)对语料库分析工具的时代划分,我们将相关工具开发情况汇总如表 1。

表 1 四代语料库分析工具

代次	年代	作者	索引工具名称
第一代	1951	Roberto Busa	不详
	1978	A. Reed	CLOC
	1992	Roger Garside	XANADU
第二代	1989	J. Bradley, et al.	TACT
	1991	Higgins	MiniConcordancer
	1993	Mike Scott & Tim Johns	MicroConcord
第三代	1997	Mike Scott	WordSmith Tools V2.0
	2012	Mike Scott	WordSmith Tools V6.0
	2002	Laurence Anthony	AntConc
	早于 1999	Michael Barlow	MonoConc Pro
	早于 2005	Lou Burnard & Tony Dodd	Xaira
第四代	2000	S. Hoffmann, S. Evert & A. Hardie	CQPweb
	2002	Mark Davies	BYU corpora
	2002	Adam Kilgarriff	SketchEngine

其中,第一代和第二代索引工具主要是在 DOS 环境下运行。第一代工具更受硬件限制,运行速度缓慢。第二代索引工具已能初步实现今天索引工具的基本功能,如:索引行的生成、词频表、短语表的生成,甚至是词语搭配的计算(如 TACT)。

第三代语料库工具以 WordSmith 为代表,这些软件主要在 Windows 或其他图形界面操作系统中运行。WordSmith 是商业软件,AntConc 为功能相近的免费替代软件。这两款软件最能代表第三代语料库分析工具,两者都拥有广泛的用户群体。WordSmith 各版本主要功能划分为三大模块,即:词汇索引(Concord)、主题词(KeyWords)、词频表(WordList)。到 5.0 版本(2008 年)以后,WordSmith 增加了框合结构(ConcGram)功能,但从界面功能划分看,主要还是维持三大核心模块。大模块下还有词簇提取(cluster)和词语搭配等子功能模块。WordSmith 的三大模块成为了通用语料库软件开发领域的事实标准。

第四代语料库工具主要指基于互联网的语料库网络应用(web application)。这类工具通过浏览器与服务器的交互,将语料库与检索工具融为一体。这些工具基于数据库和索引技术,检索响应时间快,用户体验好,一般用来处理大型语料库,如 BNC。但这些工具的灵活性不够,用户通常无法(或很难)处理本地语料库,同时,因受索引格式和数据量的限制,检索语法一般较为简单,不支持复杂检索。

因此,当前及今后很长一段时间,第三代和第四代语料库分析工具将会并存。从研究者的角度看,第三代语料库软件更能满足个人的实际研究需要。我们所开发的工具即属于第三代语料库工具。本研究主要关注单语语料库,双语和多语语料库的分析处理并不涉及。

1.3 第三代语料库工具的不足

WordSmith 和 AntConc 是目前较有代表性的第三代语料库分析工具,但前者不支持正则表达式,且界面复杂,不易操作;后者较为易用,但功能较少、计算效率较低,处理语料时容易死机或意外退出。

因此,在借鉴 WordSmith 和 AntConc 优点的基础上,本研究开发了 PowerConc 软件,对语料库软件的功能进行了梳理和扩充,并在软件易用性和计算效率上进行了优化。PowerConc 由许家金、梁茂成、贾云龙设计,贾云龙负责程序开发。PowerConc 软件的开发是对此前北外语料库语言学团队开发的 Keywords plus、Collocator、Colligator(许家金、熊文新,2009)等系列软件的整合。

2 基于 R-gram 的 PowerConc 软件的设计与开发

PowerConc 软件使用 Delphi 语言开发,支持 Windows 操作系统。与以往语料库分析软件相比,PowerConc 具有以下特点:

2.1 支持多种格式和不同语言的语料

PowerConc 在底层是基于 ANSI 编码的,除了支持英文,ANSI 编码在不同语言的操作系统中对应不同的字符集,以支持相应的操作系统默认语言,如:在简体中文系统中,ANSI 代表 GB2312 编码;在日文操作系统中,ANSI 代表 Shift-JIS 编码。因此,PowerConc 在不同语言操作系统中,可同时支持英文和该操作系统所对应的默认语言。对简体中文操作系统而言,PowerConc 至少能同时处理英文和中文两种语言的文本。

PowerConc 未来会以独立版本的形式支持 Unicode 编码的文本,以便能处理多种语言。因字符编码和存储长度等原因,Unicode 版软件效率会远低于 ANSI 版。若仅是处理某一种语言(操作系统默认语言)或英语语料,使用 ANSI 编码不失为两全其美的选择。

PowerConc 按语言特征将语料分为两大类:

(1) 无空格的连续文本:这类语料以字符为基本单位,字符间无空格分隔。未分词的汉语和日语等语言构成的语料,都属于这一类。

(2) 以空格分词的文本:这类语料以词为基本单位,词与词之间以空格分隔。英语语料属于这一类,分词后的汉语、日语等语言构成的语料也属于这一类。

PowerConc 按语料加工的程度将语料分为三大类:

生语料:主要指未经加工处理的原始语料,如:

面对严重灾情,广东省各级党组织和广大党员,带领广大群众万众一心抗击冰雪灾害。

Corpus Linguistics is empirical by nature.

切分语料:分词后的语料,如:

会议 听取 了 有关 阿坝 藏族 羌族 自治州 汶川 县 等 地 灾情 的 汇报 。

标注语料:对语料进行词性赋码后产生的语料,如:

Miss_NNB Green_NP1 came_VVD in_II secrectly_RR

国家_n 体育馆_n 瞬间_nt 成为_v 欢乐_a 的_u 海洋_n

基于以上分类标准,在简体中文操作系统环境下,ANSI 版的 PowerConc 软件共支持五类语料:英文生语料、英文标注语料、中文生语料、中文切分语料、中文标注语料。

PowerConc 目前只支持"词汇_码"格式的标注语料,即:词汇在前,标注码在后,词码间以下划线"_"连接。对于英文生语料,PowerConc 将自动在标点前添加空格,以去除标点符号对检索的干扰。

本文将以英文语料为主,对 PowerConc 的各项功能进行介绍。

2.2 支持 R-gram

基于 N 元组的检索分析是对单词的扩充,它突破了研究单位长度的限制。PowerConc 除了检索及统计功能完全支持 N 元组外,其搭配计算、词表生成和关键词计算也完全支持 N 元组,PowerConc 将单词视作一元组(uni-gram)处理。

正则表达式是语料库分析的利器,它的最大特点之一在于其高度的概括性,使研究者可以按字符种类、字符数量和字符位置三个维度描述字符特征,使研究的范围得到了极大的丰富。

PowerConc 在 N 元组中加入了对正则表达式的支持,使 N 元组得到了扩展。本研究把基于正则表达式的 N 元组称为 R-gram。

在描述能力方面,R-gram 非常强大,它既可以描述具体的语言单位(如单词或短语),也可以对抽象的语言现象(如,动词 + 名词、a + 名词 + of)进行描述;R-gram 不像单词和 N 元组那样受长度的限制。总之,R-gram 继承了正则表达式的优点,具有高度灵活性和抽象度,可以对长度和种类都不确定的语言单位进行描述,词汇和 N 元组都可以看作是 R-gram 的子集。对于赋码语料,R-gram 的优势表现得尤为突出,可以对不同数据类型的语言单位进行描述。

R-gram 除了用于检索外,也可用于搭配计算。PowerConc 支持针对 R-gram 的搭配计算,用户可自定义一个 R-gram 列表,PowerConc 将在指定跨距内计算这些 R-gram 和节点(node term)的搭配强度,这扩大了搭配研究的范围,使研究者可以对更为抽象和复杂的搭配现象进行研究。

正则表达式的缺点是语法抽象复杂,可读性差,不便掌握,对于初学者而言,出错的几率很高。为使 R-gram 便于操作和容易理解,我们设计了一种全面兼容正则表达式的简易语法——Smart Input,以降低用户的检索难度。

2.3 支持 Smart Input 语法

Smart Input 语法主要包含以下内容,软件会将用户输入的 Smart Input 语法自动转换成相应的正则表达式:

(1) "@"(@ 符):@ 放在单词原形前,表示将匹配该单词的所有屈折形式,如:@ be 将匹配 am、are、be、been、being、is、was 和 were 等 8 种形式;

(2) "#"(#号):#放在词性类别码(如:n、v、adj 等)之前,表示将匹配该词性大类对应的各词性码子类,对 CLAWS (C7) 赋码语料,#n 将匹配 NN、NN1、NN2、NP 等 22 类表示名词(n)的词性码;

(3) "*"(星号):对于以空格分割的文本代表任意一个词,如:be * at 将匹配 be good at、be amazed at 等;* 对于无空格分割的汉语文本代表一个字,如:国 * 将匹配"国家"、"国际"等;

(4) 词性码检索:对于词性赋码语料,可直接输入词性码或码串进行检索,如:对 CLAWS(C7)赋码语

料，NN1 将匹配 man_NN1、world_NN1 等；

（5）词形检索：对于词性赋码语料，可直接输入单词或短语进行检索，如：对 CLAWS（C7）赋码语料，look 将匹配 look_NN1、look_VV0、look_VVI 等；

（6）正则表达式检索：用户可直接编写正则表达式进行检索。

以上（1）~（5）可混合使用在同一表达式中，如：@be * #prep 将匹配 is capable of、was involved in、be responsible for 等内容。

综上所述，Smart Input 语法的引入使 R-gram 变成了一种可读性高、描述能力强的表述体系。表 2 列出了一些 R-gram 应用实例。

表 2 Smart Input 语法

类型	R-gram	表达式含义
词性（部分）	#n	匹配名词
	#v	匹配动词
	#adj	匹配形容词
词汇原型	@be	匹配 be 的不同屈折形式
	@go	匹配 go 的不同屈折形式
POS 码	NN1	匹配 POS 码为 NN1 的语言单位，如：part_NN1
	JJ	匹配 POS 码为 JJ 的语言单位，如：local_JJ
单词	books	匹配词汇部分为 books 的语言单位，如：books_NN2、books_VVZ 等
	can	匹配词汇部分为 can 的语言单位，如：can_VM、can_NN1 等
正则表达式	所有正则表达式	略
叠加表达式	JJ NN1	匹配 increasing_JJ number_NN1、integral_JJ part_NN1 等
	@be @good	匹配 are_VBR best_RRT、be_VBI better_RRR 等
	#v #adv	匹配 are_VBR still_RR、be_VBI more_RGR 等
混合表达式	@be #adj #prep	匹配 be_VBI capable_JJ of_IO、being_VBG involved_JJ in_II 等
	NN1 @do not #v	匹配 agriculture_NN1 does_VDZ not_XX enable_VVI 等等
	do * #prep	匹配 do_VD0 agree_VVI to_II、do_VD0 attend_VVI with_IW 等

2.4 支持基于文件或文件分组的分布统计

语料库研究经常会涉及对比分析，如按时间、语域、性别、话语者类型（母语还是学习者）、学习者阶段等维度进行对比。分布数据可以直观地显示出语料之间的差异，对于对比性研究具有非常重要的意义。

PowerConc 支持两种分布：

（1）*基于语料库文件的分布*：以语料库文件进行分布统计，如图 1。

（2）*基于语料库文件分组的分布*：按用户指定的分组条件，将语料库文件分成若干个组，然后以文件分组为单位进行统计。分组条件由针对文件名的正则表达式构成。比如，在配置文件中用正则表达式“0[1-9]”将相关文件归为一组，组名为“AmE”；文件名满足

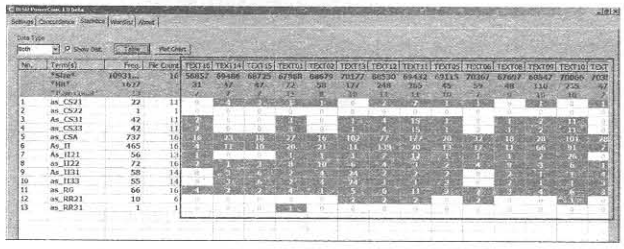

图 1 按文件进行分布统计

正则表达式“1[0-9]”的语料库文件归为另一种，组名为“BrE”。基于正则表达式的分组方式，具有高度的灵活性，同一个语料库文件可以出现在不同的分组之中。图 2 中，各语料库文件已按指定分组原则，被归入了“AmE”和“BrE”两组中。

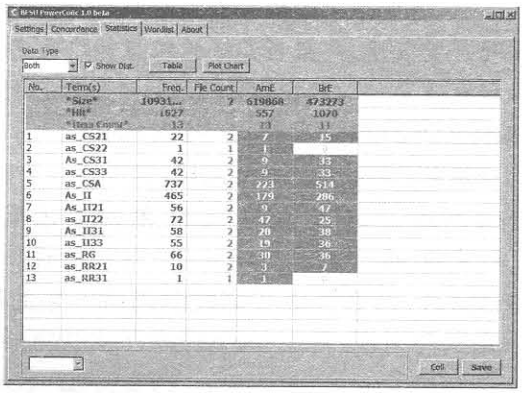

图 2 按用户指定分组条件呈现统计结果

2.5 功能结构清晰

应用软件不仅要实现特定的功能，还需要根据用户习惯对这些功能合理规划，以提高易用性。鉴于此，PowerConc 在设计过程中，对语料库分析和研究常用的各种功能进行了梳理和分类，淘汰了不常用的功能，加入了多项以往语料库软件没有的功能，同时归并到统一的框架下，使所有功能一目了然，当用户熟悉了一个模块后，可以很快掌握其他模块，有效降低软件的学习

表 3 PowerConc 各模块功能描述

	Ⅰ 检索相关分析（微观分析）			Ⅱ 词表相关分析（宏观分析）	
	1. 词汇索引	2. 结果统计	3. 搭配计算	1. 词表生成	2. 主题词计算
1. 参数设置	√	√	√	√	√
2. 数据计算	√	√	√	√	√
3. 分布统计	×	√	×	√	×
4. 结果显示	√	√	√	√	√
5. 结果排序	√	√	√	√	√
6. 结果过滤	√	×	√	√	√
7. 结果抽样	√	×	×	×	×
8. 结果保存	√	√	√	√	√

成本和操作难度。

PowerConc 将语料库分析归为微观和宏观两大类,词汇索引主要是对某个或某类语言现象进行的有针对性的分析,属微观分析;词表是对语料库整体情况的考察,属宏观分析。

PowerConc 共五个模块,检索相关的功能包括三个子模块,词表相关的分析包含两个子模块:

(Ⅰ)检索相关功能(微观分析)

(1)词汇索引(Concordance):根据输入内容进行检索并返回索引行;

(2)结果统计(Statistics):对检索命中的内容进行统计;

(3)搭配/类联接计算(Collocation & Colligation):根据索引行计算检索结果的搭配信息。

(Ⅱ)词表相关功能(宏观分析)

(1)词表生成(N-gram list):根据指定的数据类型和长度生成词表;

(2)主题性计算(Keyness):根据参考词表计算主题词,或主题短语。

每个模块的操作又可细分为参数设置、数据计算、分布统计、结果显示、结果排序、结果过滤、结果抽样、结果保存八个类别。

3 PowerConc、WordSmith、AntConc 的对比

PowerConc 不是对 WordSmith 等软件的重写或复制,PowerConc 有其独立的设计原则。

继承发扬:吸收以往语料库软件中那些广为接受的功能(如:词汇索引、词表等),在充分考虑用户体验的基础上,删繁就简,对功能的设计、展示和实现方式进行优化,以降低学习成本和操作难度,并将它们纳入 PowerConc 的统一设计框架中,使软件设计具有高度一致性;其次,对这些功能进行扩展,丰富已有功能;再次,对算法进行优化,提高程序处理数据的能力和效率。

将 PowerConc 与 AntConc 和 WordSmith 进行对比,可以归纳出以下几方面差异:

学习成本和操作难度:三款软件中 PowerConc 的学习成本和操作难度最低,具有一般软件操作的常识和语料库的基本知识的研究者,都可以很快上手。WordSmith 的学习成本和操作难度最高,它的功能非常繁杂,即使经验丰富的研究者也很难快速掌握它的全部功能,WordSmith 6.0(最新版)的说明书已达 415 页。AntConc 学习难度居中,但配置复杂,不便操作。

功能划分和界面布局:PowerConc 完全以研究者的视角来进行功能划分和界面布局,全部功能的设计遵守统一规范,用户可举一反三。WordSmith 的功能规划和界面设计不合理,不符合常规软件的设计原理,若不借助说明书,仅靠界面本身提供的信息,很难进行操作。AntConc 在 WordSmith 基础上进行了一定的优化,但在局部设计上,缺乏连贯性。

核心功能:PowerConc 和 AntConc 对 WordSmith 的功能进行了取舍,分别实现了 WordSmith 的词汇索引、词表、关键词计算三大核心功能。AntConc 几乎是对 WordSmith 三大功能的简化和重写,相对 WordSmith 而言没有本质的变化。PowerConc 以 R-gram 为基础重新设计,使语料库软件的功能得到了扩展。

功能创新:WordSmith 不支持正则表达式,这使它的功能受到了极大的限制。WordSmith 历史较久,版本众多,但每一个新版本的变化并不大,通常只是加入了个别新功能或修正个别小错误,而核心功能几乎没有多少改变。AntConc 加入了对正则表达式的支持,但相对 WordSmith 而言,其功能并无创新。PowerConc 最大的创新是对 R-gram 和 Smart Input 的支持,它使语料库研究的范围得到了扩展。同时,PowerConc 对数据分布统计也具有很好的实现。

算法效率:因为功能设计的差距,三个软件很难直接对算法效率进行对比。整体而言,AntConc 的算法效率最低,对数据量的大小较敏感,容易死机或意外退出。WordSmith 的效率一般,算法没有进行优化,个别计算要耗费大量的时间。PowerConc 在算法优化上做了大量尝试,最大程度上避免了数据拷贝带来的资源浪费,同时一些模块使用了缓存方式以避免信息的重复计算,使计算效率大大提升。

可扩展性:PowerConc 基于面向对象的方法开发,核心功能被封装在不同的类(Class)中,实现了界面和功能的分离。这使 PowerConc 具有非常好的扩展性,一方面,可以不断对现有功能进行升级和维护,也可以加入新功能;另一方面,可以利用这些核心功能类,开发出衍生产品。这些优势是 AntConc 和 WordSmith 等软件不具备的。WordSmith 的升级和维护几乎是用打补丁的方式进行的,这使 WordSmith 的安装包越来越大,操作越来越复杂。WordSmith 6.0 的安装文件有 54MB,安装后有 108 个文件,而 AntConc 和 PowerConc 都是绿色软件,无需安装,AntConc 3.2.4(最新版)的大小是 4.4MB,PowerConc 只有 1.5MB。

相信 PowerConc 的开发将有效促进语料库语言学

研究的开展。

感谢梁茂成教授对本文提出的宝贵的修改意见。 □

参 考 文 献

[1] Baker, P., Hardie, A. & McEnery, T. (eds.). *A Glossary of Corpus Linguistics*[M]. Edinburgh: Edinburgh University Press, 2006.

[2] Hunston, S. *Corpora in Applied Linguistics*[M]. Cambridge: Cambridge University Press, 2002.

[3] McEnery, T. & Hardie, A. *Corpus Linguistics: Method, Theory and Practice*[M]. Cambridge: Cambridge University Press, 2012.

[4] Sinclair, J. *Corpus, Concordance, Collocation*[M]. Oxford: Oxford University Press, 1991.

[5] 许家金,熊文新. 基于学习者英语语料的类联接研究:概念、方法及例析[J]. 外语电化教学,2009(3).

The Design and Development of the R-gram Based Corpus Analysis Tool 'PowerConc'

XU Jia-jin, JIA Yun-long

(National Research Center for Foreign Language Education, Beijing Foreign Studies University, Beijing 100089, China)

Abstract: This paper describes the innovative corpus tool PowerConc developed by the authors themselves. In its implementation, such key functionalities of corpus tools as concordancing, wordlist generation and keyword analysis, were redesigned, expanded and optimized. What underlies the whole design of PowerConc is the inventive synergy of regular expressions and N-gram, namely, R-gram proposed in this study. The R-gram feature allows for flexibility in concordancing, exhaustive listing of linguistic units, and key terms of varying length, and more likely than not, enables analyses of linguistic structures with uncertain words or categories. To minimize the inconvenience of operation, Smart Input has been introduced to facilitate easy search with enhanced returned hits. The user-friendly PowerConc is object-oriented software. The key functions have been packaged in different classes as a dynamic link library (*.dll) file, independent from the user interface, which warrants easy maintenance and expandability. It is hoped that PowerConc will be conducive to corpus-based research in its own way.

Key words: Corpus Tools; PowerConc; R-gram; Corpus