CAFLE
外语电化教学
双月刊
2012年 第3期
（总第144期）

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

语料库语言学研究

外语课程与教学研究

外语教育技术理论研究

技术与应用

期刊基本参数：CN31-1036/G4*1981*b*A4*80*zh*P*￥12.00*8000*14*2012-2

CAFLE
Computer-assisted Foreign Language Education

No.3 May. 2012
(General Serial No.144)

Computer-assisted Foreign Language Education in China, a bimonthly journal of theory research and practice, is sponsored by the Ministry of Education P.R.China and Shanghai International Studies University (SHISU), edited by Foreign Language Professional Committee, Chinese Association for Educational Technology, and published by Shanghai Foreign Language Audiovisual Publishing House. Domestic distributors: Local post offices. Overseas distributor: China International Book Trading Corporation (China Publication Centre), P.O. Box 399, Beijing, China. Address all correspondence to the CAFLE editorial board: 550, Dalianxi Rd., Shanghai, China.

Tel:(021)35373318
E-mail:wydhjx204@163.com
http://wydh.qikan.com
http://wydhjx.shisu.edu.cn

CONTENTS

语料库语言学研究

谷歌图书与谷歌图书语料库比较

汪兴富[1],Mark Davies[2]
(1. 重庆大学外国语学院,重庆 400030; 2. Brigham Young University,USA)

摘 要: 本文主要对比谷歌图书界面和语料库专家 Mark Davies 教授依据谷歌图书数据制作的谷歌图书语料库界面,在语料库界面可以查询谷歌图书收录的 1810 – 2009 年间 130 多万本美国英语书籍的 1550 多亿词汇中的信息,有比谷歌图书丰富的查询条件和限定范围,并且利于进行历时比较研究。

关键词: 谷歌图书;语料库;n-grams
中图分类号: H319.3 **文献标识码:** A **文章编号:** 1001-5795(2012)03-0015-0004

1 引论

美国杨伯翰大学(Brigham Young University)语料库专家 Mark Davies 教授的语料库家族又有重量级成果呈现,这就是美国英语谷歌图书语料库(Google Books (American English) Corpus) (http://google-books.byu.edu/)。Davies 教授已相继推出多个大型在线免费英语语料库,如 COCA (4.25 亿词汇,1990 – 2011 美国英语语料),COHA (4 亿词汇,1810 – 2009 美国英语语料),BNC (1 亿词汇,1980s – 1993 英国英语语料,Davies 教授的查询界面),这些语料库均在 http://corpus.byu.edu 有链接),2011 年 5 月又将在线的谷歌图书纳入了他的语料库研究范围。

美国英语谷歌图书语料库包含了谷歌图书收录的 1810 – 2009 年间 130 多万本美国英语书籍共 1550 多亿词汇(其中仅 1980 – 2009 年间的词数就为 620 亿)。如果说亿级词汇库容的语料库已成为现在建库的基本标准,那么美国英语谷歌图书语料库的库容可以用海量来形容了。该语料库依据谷歌图书却并非谷歌的官方产品,而是由 Davies 教授制作丰富的语料库查询界面,大大拓宽了对谷歌图书的用户友好度查询和研究范围,为从事英语语言教学和研究的人员以及其他需要查询谷歌图书资料的人员搭建了一个进行深度研究的平台,借助浩瀚的数据为用户开启了新的研究视窗。

2 谷歌图书界面与谷歌图书语料库界面

传统谷歌图书界面(http://books.google.com/)容许用户对具体词汇进行查询,也能返回一个历时频率表。用户可以在该界面查询到具体词如 collocation, music,NLP 的频率,但对于具体的频数分布情况却不可见,对于这些词的意义和用法也无法通过查询近邻词汇(nearby words)解决,仅有的数据也不方便复制分析,无法查询哪些词在何时进入语言流行开来并且在特定时段内呈发展趋势,也无法比较特定时段内某些名词被修饰的情况(如无法对比 1920 – 1940 和 1980 – 2000 两个时段内 women 一词主要被哪些形容词修饰的情况,20 世纪和 19 世纪与 food 相邻出现的词有什么不同等)。对曲折后缀的相关词性查询只能逐个进行而无法一次性完成(如查询 end up paying,ends up paying,ended up paying 就需要重复进行三次,而且数据是分散的,也没有不同年代的数据)。在对比查询词语不同用法时也没有数据表可以方便研究者复制分析。用标准的谷歌图书界面找寻资料进行研究只能发现"冰山一角",而借助 Davies 教授的美国英语谷歌图书语料库将开启谷歌图书数据的巨大潜力。

表 1 进行的查询清晰地反馈了以 end 为动词时不

作者简介:汪兴富:男,博士生,副教授。研究方向:认知语言学和语料库语言学。
Mark Davies:男,博士,教授。研究方向:语料库语言学。
基金项目:* 系列研究得到重庆大学语言认知及语言应用研究基地项目资助,谨致谢忱!
收稿日期:2011-05-13

表 1　谷歌图书语料库查询[end] up paying 结果

	WORD (S)	TOTAL 1810 – 2009	1940s	1950s	1960s	1970s	1980s	1990s	2000s
1	end up paying	7963	7	35	141	617	1013	2004	4145
2	ended up paying	2537	3	10	82	203	296	666	1277
3	ends up paying	1109		13	41	134	194	272	455
	TOTAL	11,609	10	58	264	954	1,503	2,942	5,877

同形态的结果,表格数据表明[end]up paying 结构自 20 世纪 40 年代开始至 2009 年的 10 年期时段呈现上升趋势,而在谷歌图书界面是无法做到类似查询的。

美国英语谷歌图书语料库界面有比谷歌图书界面更为先进复杂的查询内容,和 Davies 其他语料库如 COCA 美国当代英语语料库的界面类似(该语料库界面具体介绍请参考汪兴富等,2008),都可以进行语料库式的对词、词组、词条、词性、通配符、同义词(同类概念)和搭配的查询。传统谷歌图书界面不能查询通配符,但在谷歌图书语料库界面可以用通配符查询,如查含 heart 的合成词或带-ism 后缀的所有词语,因而可以比较不同时代的 patriotism, communism, heroism, terrorism, skepticism, racism 的具体发展变化。由于该语料库不是平衡语料库,因而数据也只按 10 年时期方式展现和排列,以体现不同时代图书用词的历时变化。查询结果的显示方式与 Davies 教授其他语料库有所不同,现只提供 List 和 Chart 两种,默认查询时结果以 List 方式显示。SECTIONS 分类部分也只有每 10 年的时段选择,如果选定时段 SECTIONS 去查询内容,则可以进行不同时段的数据对比分析,该语料库界面对于需要做历时对比研究是非常得力的助手。查询到的数据可以复制到 Excel 或 SPSS 软件进行统计分析和计算频率差异。如可以查询在谷歌图书中 1850 – 1900 年间与 1950 – 2000 年间 strong 的同义词使用对比情况,也可以对不同时段内 fast 所修饰的名词进行历时分析。如查询 gay 在 1830s – 1910s 和 1970s – 2000s 时段的使用情况显示为:在第一个时段内该词常与 brilliant, attractive, jolly 和 joking 结伴,而现在主要与 heterosexual, sexes, groups 和 bisexual 等词同现,这种搭配变化自然反映了词语 gay 本身意义的变化。

3　谷歌图书语料库特色

具体来讲,用美国英语谷歌图书语料库界面可以查询出独立词如 grieved, sublime, bosom 的使用频率呈下降趋势;steamship, telegraph 和 swell (作形容词)为先上升后下降趋势;而 teenager, funky 和 guys 的用法呈上升趋势。也可以查询短语,并且短语中的具体位置词可以用词性界定或使用通配符。Davies 教授的美国英语谷歌图书语料库界面可以使用词性 POS 查询,这在谷歌图书界面是无法进行的。如可以查询诸如 beautiful NOUN, ADJ woman, walked ADV-ly, 或 VERB the way(大写字母为词性类别,置于查询框中则按语料库要求进行,如 NOUN 需要输入为[n*])。也可以查询一个词的所有词形形式或同义词。还可以合并查询复杂结构如 synonyms of beautiful + a form of woman(查询框里输入[=beautiful][woman]),或 ADJ + synonym of silliness(输入[j*][=silliness])。

对句法感兴趣的用户同样可以在该界面查询特定句式结构,如[start] to VERB(输入[start]to[v*]),VERB one's way PREP (输入[vv*][ap*]way[i*],得到结果如 force his way into 等)。对于曾经研究过助动词的研究者可以再在该千亿级大型语料库中对比 must VERB, should VERB, ought to VERB, has to VERB 或 need to VERB 的用法,相信会有新的斩获。而且这些查询均是返回整类查询结果,在谷歌图书中则只能逐项查询,但动词类是开放词类,无法进行穷尽式的查询。在美国英语谷歌图书语料库查询这些结果只需几秒钟就可以呈现,但标准谷歌图书进行相同查询则可能历时数小时、几天甚至几周都无法完成。

近邻词汇(nearby words)与搭配(collocates) 严格来讲是不同的,语料库中查询的是一个词的近邻词语而并非全部是 Firth 所定义的"由词之结伴可知其词"的有意义词项组合。近邻词汇比搭配包含的范围更广,包括了词之间没有修饰关系但在一定词距内的共现成分,为了方便描述本文仍然通称为"搭配"。偶然的词语近距离同现也许只是特例,但高频的同现则可能说明尽管同现词间无显著结构修饰关系,但它们之间一定有某种语义关系。近邻词汇搭配的查询(需要将 COLLOCATES 一栏点击显现出输入框才可以进行)极大地突破了谷歌图书的限制,查询不再局限于谷歌图书的实际字符串查询,可以使用词性类别,如进行相邻的动名搭配、形名搭配和名名搭配等查询;也可以跨距查询搭配,但由于谷歌图书 n 元组中 n 的最大值为 5,所以只支持左或右跨距为 4 词内的查询,并且查询搭配时输入 WORD(S)框的核心词目前只能是单个词汇,两词以上的词串暂时不能进行该类处理。如查询

entropy 的左侧搭配动词可以得到如 calculate, find, see, compute 和 alter 等动词,点击所查询的字串结果或点击每 10 年时期显示的频数即可链接到相应的谷歌图书网站界面,并且以所查询的字符串在谷歌图书搜索并显示图书预览,用户不用再次在谷歌图书界面输入想查询的内容。

在谷歌图书语料库查询词汇 a 与 b 搭配时(b 输入的是表类别的词性),实际查询将进行两个步骤:第一步由语料库查出所有 b 与 a 共现的词汇(单个词汇),第二步将在用户点击具体显现的 b 词项后展现所定义词距结果(即 n 元组)。

在谷歌图书语料库和 COCA 语料库查询词汇词性时可以使用更精准的比例限定标记{xx},表示按该词在语料库中的词性比例进行查询,xx 的值为 01 - 99,默认值为 50。如输入[v*]{90} like that 表示查询该形式中第一个词(即[v*]{90})为动词词性且为动词时比例是 90% 的情形,得到的结果有 sound /work /act /face /places /dress like that,因为该结构第一个词还可能为名词,如 a strange sound like that, do work like that, with a face like that。

谷歌图书语料库与 COCA 语料库相比也有不同的特色展现,如在 COCA 查询 language proficiency 有 134 次,在谷歌图书语料库查询同时段的数据为 26545 次,差异自然有库容的因素(同时段 COCA 收词 4.25 亿,谷歌图书语料库 468.9 亿),但主要是语料库收录的材料不同所致,如谷歌图书谈 language proficiency 的专著就有多本。

4 谷歌图书语料库查询原理

整个语料库是基于谷歌图书提供的 n-grams(即 n 个词序列)制作的,查询美国英语谷歌图书语料库时实际是在查询这些 n 元组,而不是在直接查询谷歌图书。但由于这些频率表与谷歌图书间建立了链接,点击查询结果或频率数后也就能在 Google 图书中看到具体的呈现情形。谷歌图书语料库是将每个区别性的 n 元组整合为 10 年期的数据存储在关系数据库中供用户查询。

只有那些在 1550 亿词汇中出现 40 次以上的词串才会进入 Google Books 的 n 元组;谷歌图书 n 元组本身不是按词性 POS 标注的,但 Davies 教授对谷歌图书语料库中的 n 元组数据专门做了词性标注。由于 1-gram 只有一个词汇,当词具有多个词性又无上下文参照时无法准确判定其词性,如 light。1-gram 相当于一个词频表,即给出所有词出现的频数。2 元组及以上的词串方便标注,尤其是在有冠词或其他限定词修饰时能适当排歧(disambiguated),较易判定其名词词性(如 the light),或不定式符号 to 及助动词有助于判定动词(如 I want to light the fire)。如在谷歌图书查询 20 世纪 50 年代"into revealing"有 382 项,其中有 maneuvered them into revealing 和 trap the husband into revealing,但在谷歌图书语料库单独查询这两例却没有返回任何词条,原因就是因为这样的 n 元组数据没有达到 40 次的阈值限度,没有进入谷歌图书 n 元组集合。另外长词串的 4 元组或 5 元组的查询不如短词串表现出色,查询 2 个词距内的结果会比查询 4 个词距内的结果多得多。

40 次的阈值有其独特的意义,用户至少不会将这类词串看作排版错误(typos)或异常情形(anomalies)。谷歌图书语料库比 4 亿词汇的 COHA 语料库大得多,词条的总频率数(tokens)也要多许多,但词串类别(types)与 COHA 相比只是接近甚至更少,如表 2 所示。

表 2 谷歌图书与 COHA 部分数据对比

construction	examples	Google Books		COHA	
		tokens	types	tokens	types
[j*] groan	heavy/hollow/ muffled groan	56783	183	869	274
sultry [nn*]	sultry heat/ weather/voice	82,419	224	687	236
walked *ly. [r*]	walked quickly/ slowly/briskly	446,414	398	5,162	376
started to [vv*]	started to run /walk/notice	868,371	907	10,566	1282

谷歌 n 元组是大小写敏感的(case sensitive),如查询 mutual information 将返回 4 类结果:mutual information (7099), Mutual Information (628), Mutual information (501),和 MUTUAL INFORMATION (66)(括号中数字为该形式的频数)。Davies 教授争取在今后添加选项框把该类无意义区别的信息集中呈现。Google 图书的 n 元组也包括了标点符号,但与书的对应不是很好,如"idea; nevertheless"中间的标点符号在谷歌图书中显示为,或. 或;,甚至没有。也就是说查询在 n 元组中可见,但在谷歌图书中不一定正确显现。这对于需要将标点符号纳入查询条件的使用者来说是个问题。后续工作将努力解决这一难题。

有一点需要注意的就是用户可能会发现 Davies 教授的谷歌图书语料库频率查询数据与谷歌图书网站数据不一致(mismatch)。然而该语料库数据与谷歌 n 元组数据是完全一致的。原因在于该语料库是美国英语图书数据集,而谷歌图书呈现的图书摘录是所有谷歌图书(包括英国英语图书等),因而在谷歌图书网站中

总会有比谷歌图书语料库 n 元频率更多的数据。依据谷歌图书语料库查询的数据是正确的,谷歌图书摘录数据只是没有给定限定范围而显示了更多数据而已。

5 后续工作

离散查询现在是语料库的弱项,也就是在该语料库界面下无法方便查询几个独立的关键词信息,如查询什么书论述了 NLP, mutual information, antonyms 和 prototype,在谷歌图书中查询则相对方便一些。谷歌图书语料库查询的功能暂时还没有 Davies 教授其他几个语料库丰富,并且在进行搭配查询时 WORD(S)框只能输入 1 个单词,包括通配符。Davies 教授争取尽快增加谷歌图书语料库查询功能,变成和其他语料库一样丰富的有序关系查询。

该界面现在只是一个先期版本,并且只包含美国英语图书数据。Davies 教授接下来会增加更多特色查询框。尤其希望像在 COHA 中进行语义导向的查询,如查询表达" briefly touch someone" 概念(输入[[= stroke]].[v*] [ap*] [n*]),就会显示 stroking her hair, rubbed his chin, patted her shoulder 等结构。谷歌图书语料库界面证明谷歌图书数据可以整合到现有的 Davies 教授的各语料库界面中。目前的一些基本功能还会有陆续添加以丰富界面和查询内容,如频数分布按年而不只是按 10 年期显示、两个词的多个搭配词之间的比较、两词以上字串的搭配词查询、可存储用户查询的词表及结果、按词目词归类以及兼容不同的操作系统和浏览器等功能。由于受限于谷歌图书数据的 n-grams 表示,10 – 20 词的真实长词串不能查询。但后续的更新将能让用户查询任一词的最常见词串。并且作者也希望在获得相关研究经费后增加英国英语、西班牙语、德语和法语的类似图书语料库,这些语料库库容都至少在 500 亿词汇以上,它们对于已有资源和检索将是非常了不起的补充。借助这些语料库人们将能够进行更深入的语言研究和更精准的行业资料查询,而这样的研究用其他语料库工具目前都是不容易实现的。

谷歌图书对自己汇总图书的工程给了出版社和作者一个美妙的解释:"我们尊重您为您的图书所付出的大量创造性劳动。这就是我们想尽可能地让人们容易找到图书的原因。"(We respect the tremendous creative effort you put into your books. That's why we want to make it as easy as possible for people to find them.)但谷歌图书这句话似乎只是让读者快速找到了书,并没有帮读者高效阅读和查找书中信息。我们的认识是谷歌图书语料库在谷歌图书的注解后加入了"and use them effectively",让读者真正高效地使用图书。

谷歌图书本身就已是一个不可思议的工具,使研究者在资源文献方面获益匪浅。所有图书包含的词汇总量是其他任何已有英语语料库的几百倍甚至上千倍。但其许多作用和潜力都被 Google Books 界面有意或无意地禁锢起来了。借助 Davies 教授的美国英语谷歌图书语料库界面将能让英语研究者和学习者更上层楼,让专业的资料查询更简单友好并且高效,能让使用者发挥无限潜能去大胆挖掘谷歌图书这一巨型语料库宝藏。 □

参 考 文 献

[1] Davies, Mark. (2011) Google Books (American English) Corpus (155 billion words, 1810-2009)[WE/OL]. http://googlebooks.byu.edu/.

[2] Davies, Mark. The advantage of using relational databases for large corpora: speed, advanced queries, and unlimited annotation [J]. *International Journal of Corpus Linguistics*, 2005(10).

[3] 汪兴富,Mark Davies, 刘国辉. 美国当代英语语料库(COCA)——英语教学与研究的良好平台[J]. 外语电化教学,2008(5).

Comparisons between Google Books and Google Books Corpus

Wang Xing-fu[1], *Mark Davies*[2]

(1. College of Foreign Languages, Chongqing University, Chongqing 400030, China;
2. Brigham Young University, Utah 84602, USA)

Abstract: The paper is to compare the interface of Google Books and the interface of Google Books (American English) Corpus created by Professor Davies. The Google Books Corpus allows users to search more than 155 billion words in more than 1.3 million books of American English from 1810 – 2009 (including 62 billion words from 1980 – 2009). And this interface allows users to search the Google Books data in many ways that are much more advanced than what is possible with the simple Google Books interface. One of the outstanding features of the interface is to make a diachronic study of words.

Key words: Google Books; Corpus; N-grams