语料库语言学研究专辑

美国当代英语语料库(COCA)
——英语教学与研究的良好平台

汪兴富[1],Mark Davies[2],刘国辉[3]
(1. 重庆大学 外国语学院,重庆 400030; 2. Brigham Young University, USA;
3. 杭州师范大学 外国语学院,浙江杭州 310036)

摘 要: 本文系统介绍美国杨伯翰大学 Mark Davies教授开发的 COCA美国当代英语语料库。该语料库库容量为 3.6亿词汇,涵盖美国 1990年至 2007年间 18年内的各种类型语料,是当今世界上最大的英语平衡语料库。现该语料库免费在线供研究者和学习者使用。

关键词: COCA美国当代英语语料库;平衡语料库
中图分类号: H319.3 **文献标识码**: B **文章编号**: 1001-5795(2008)05-0027-0007

1 引论

COCA——美国当代英语语料库(Corpus of Contemporary American English)(http://www.americancorpus.org/),是由美国 Brigham Young University的 Mark Davies教授开发的高达 3.6亿词汇的美国最新当代英语语料库,是当今世界上最大的英语平衡语料库。与其它语料库不同的是它是免费在线供大家使用,给全世界英语学习者带来了福音,是不可多得的一个英语学习宝库,也是观察美国英语使用和变化的一个绝佳窗口。

COCA美国当代英语语料库于 2008年 2月 20日在互联网上正式推出。在来自全球 25个国家和地区有 140余位专家学者参会的 AACL-2008(American Association for Corpus Linguistics)学术会议上,会议的组织者 Mark Davies教授介绍了自己开发的 COCA美国当代英语语料库,此外在此学术会议上还有部分学者对研究使用这一语料库进行了交流,获得了热烈反响。为让更多中国的英语教师和学习者从大型语料库的知识海洋中受益,现积极向国内读者作一介绍。

COCA美国当代英语语料库具备了一个好语料库的三项最基本条件:规模(Size)、速度(Speed)以及词性标注(Annotation)(Davies, 2005: 301)。COCA美国当代英语语料库规模为超过 3.6亿词汇库容;在线查询速度依据查询内容差异及复杂度、网络状况和联网计算机配置而定,尽管本语料库数据存储达到了 108GB,但服务器共有 8个 2.4GHz的处理芯片同时计算,因此绝大多数查询都能控制在几秒内呈现结果;词性标注使用的是英语语料广泛采用的 CLAWS7软件。最为重要的是 COCA美国当代英语语料库收集的数据是最近 18年(1990年到 2007年)美国境内多个领域的语料,每年约 2000万词汇,而且今后每年至少更新两次。它在这四项上的突出表现目前都是其它语料库不可企及的。

COCA美国当代英语语料库涵盖美国这一时期的口语、小说、流行杂志、报纸和学术期刊五大类型的语料,并且在这五个类型方面基本呈均匀平衡分布,在每五年的时段中也是基本均匀分布的(见 Table 1)。语料库其它相关数据分别见 Table 2和 Table 3。

使用者可以在全库中查询,或限定查询的范围为任何一个大语料类型,也可以限定为任何一个子库,或几个子库的组合,

Table 1 五大类型的语料及五年时段的词数分布

语料类型	口语(SPOK)	小说(FIC)	杂志(MAG)	报纸(NEWS)	学术期刊(ACAD)	1990 - 1994	1995 - 1999	2000 - 2004	2005 - 2007
总词数(百万)	76.6	69.6	78.1	73.4	73.0	103.1	102.3	102.9	62.4

作者简介:汪兴富: 男,副教授,博士生,美国 Brigham Young University访问学者。研究方向:认知语言学和语料库语言学。
Mark Davies:男,美国 Brigham Young University教授。研究方向:语料库语言学。
刘国辉:男,教授,博士。研究方向:认知语言学,语用学和英汉对比研究。

收稿日期: 2008-02-18,修改日期: 2008-03-17——2008-07-30
基金项目:本研究得到中国国家留学基金委赴美国访学的资助及重庆大学人文社科青年教师科研启动项目资助,谨致谢忱!

Table 2 COCA美国当代英语语料库数据

分类	总词汇	文章总数	词型(types)	词目(lemmas)	Hapaxes	句子总数
总词数	370,691,937	147,093	2,297,689	2,436,450	1,455,301	22,136,258

Table 3 各子语料库(共 42个)词汇总数分布

口语子库(9个)	子库总字数	流行杂志子库(11个)	子库总字数
SPOK: ABC	13,132,214	MAG: Afric-Amer	3,357,201
SPOK: CBS	10,702,932	MAG: Children	2,228,071
SPOK: CNN	18,579,309	MAG: Entertain	3,479,537
SPOK: FOX	3,883,546	MAG: Financial	5,311,157
SPOK: Indep	4,513,182	MAG: Women/Men	5,858,827
SPOK: MSNBC	809,105	MAG: Home/Health	12,963,244
SPOK: NBC	4,009,442	MAG: News/Opin	15,843,590
SPOK: NPR	15,315,470	MAG: Religion	3,009,200
SPOK: PBS	5,623,129	MAG: Sci/Tech	9,881,440
学术期刊子库(9个)	**子库总字数**	MAG: Soc/Arts	6,696,799
ACAD: Education	6,409,519	MAG: Sports	9,440,869
ACAD: Geog/SocSci	12,928,300	**报纸子库(8个)**	**子库总字数**
ACAD: History	10,367,436	NEWS: Editorial	4,063,608
ACAD: Humanities	9,629,214	NEWS: Life	12,883,821
ACAD: Law/PolSci	7,887,439	NEWS: Misc	24,691,477
ACAD: Medicine	4,512,889	NEWS: Money	6,295,632
ACAD: Misc	3,358,303	NEWS: News_Intl	3,731,400
ACAD: Phil/Rel	5,902,797	NEWS: News_Local	5,237,152
ACAD: Sci/Tech	12,087,448	NEWS: News_Natl	5,318,754
		NEWS: Sports	11,162,901
小说子库(5个)	**子库总字数**	FIC: Juvenile	2,794,394
FIC: Gen (Book)	14,266,742	FIC: Movies	9,208,596
FIC: Gen (Jml)	28,894,677	FIC: SciFi/Fant	14,456,959

这对于对比研究非常有用。

2 语料库查询界面

COCA美国当代英语语料库查询网页(见 Figure 1)上部有1个基本信息显示区(A),显示语料库名称(The Corpus of Contemporary American English(COCA))和作者信息(Mark Davies/Brigham Young University)。另有3个查询显示区,分别为左边的"显示及查询条件界定区"(B),右上方的"查询结果数据显示区"(C)和右下方的"例句显示区"(D),整个界面直观明了。Figure 1为查询'laugh[n*]的界面。"显示及查询条件界定区"(B)需点击部分小项才能有如图显示,"例句显示区"(D)的内容为点击(C)区内第一个链接的显示结果。

2.1 显示及查询条件界定区(B)

"显示及查询条件界定区"分为四部分,分别是:显示方式区(DISPALY)、字串查询区(SEARCH STRING)、语料库分类区(SECTION)、以及查询结果排列方式区(CLICK TO SEE OPTIONS)。

2.1.1 显示方式(DISPALY)

显示方式(DISPALY)又分为图表显示(CHART)、列表显示(LIST)和单词比较(COMPARE WORDS)。这三项默认情况下的语料库分类显示(SECTIONS)为NO,查询排列方式(SORT)默认为以频率排列(FREQUENCY),结果将在(C)区展现。'显示方式'区的图表显示(CHART)为返回柱状图,能直观地看到各个子语料库的总的词频,尤其在做五大语料类型比较或各个时段的字词使用比较时最为有用。在图表显示(CHART)被选定后,除输入到字符串查询区(WORD(S))的信息有用外,其它所有的选择项能被选择但不起限定作用。列表显示(LIST)是以列表方式返回所查询的词或字符串,也是本语

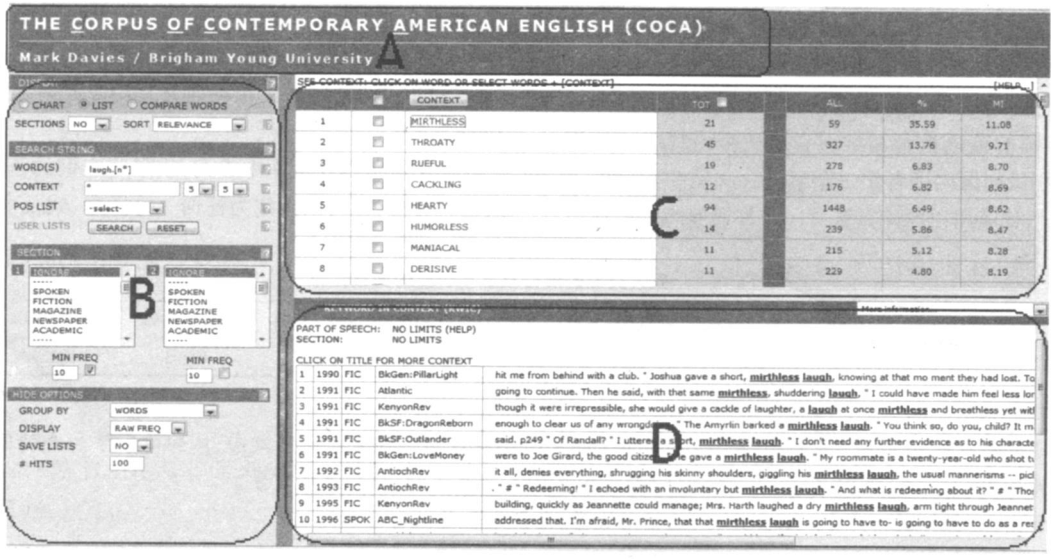
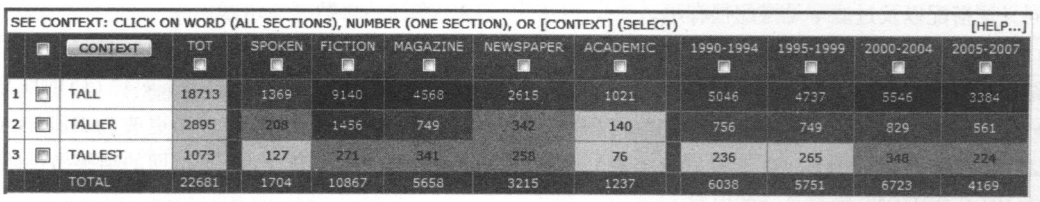


Figure 1 查询'laugh[n*]的界面


SEE CONTEXT: CLICK ON WORD (ALL SECTIONS), NUMBER (ONE SECTION), OR [CONTEXT] (SELECT) [HELP...]

	CONTEXT	TOT	SPOKEN	FICTION	MAGAZINE	NEWSPAPER	ACADEMIC	1990-1994	1995-1999	2000-2004	2005-2007
1	TALL	18713	1369	9140	4568	2615	1021	5046	4737	5546	3384
2	TALLER	2895	208	1456	749	342	140	756	749	829	561
3	TALLEST	1073	127	271	341	258	76	236	265	348	224
	TOTAL	22681	1704	10867	5658	3215	1237	6038	5751	6723	4169

Figure 2 查询[tall]的界面

料库的默认显示方式。单词比较(COMPARE WORDS)能让使用者比较两个不同词或短语的搭配情况,当选择比较单词时显示方式区(DISPLAY)中的查询排列方式(SORT)将自动变为RELEVANCE(相关度)对结果进行排列。在SORT(按类型排列)中点击下拉箭头可以有3项选择,分别为FREQUENCY(按频率排列),RELEVANCE(按相关度排列),ALPHABETICAL(按字母序排列)。按相关度排列(RELEVANCE)是非常有用的选项,它能给出所查询的词与哪些词的关系最为紧密,但却过滤了高频搭配的噪音词(主要是那些empty words)。使用按相关度排列(RELEVANCE)功能时依据比较的信息不同而有细微差异。通常情况下,查询结果按所查询的词的互信息(Mutual Information)值的高低排列。互信息(MI)是对随机的两个词相关性的度量。也就是要查询的词和可能性搭配词在所有语料库中的共现搭配比重(百分比),同时也决定了其互信息的值。

在比较两个词语时,按相关度排列(RELEVANCE)功能时会返回一个与最常见词搭配的列表,比如在查询rob和steal的搭配时就可以看到与rob搭配的词语并且与steal搭配的情况,反之亦然。当设置为FREQUENCY(按频率排列)时,返回的结果就只是一个按搭配频率排列额从大到小的一个列表,但这样会包含很多噪音。当设置为ALPHABETICAL(按字母排列)时所返回的查询结果则按字母序排列。

当这一区的分类显示(SECTIONS)被选为YES后,五种语料类型和四个时间段的信息将显示出来,但同时子语料库信息不再展示。Figure 2为选择LIST和SECTIONS选择为YES后查询[tall]的C区显示图,表格颜色的深浅表示单词的频率高低。

2.1.2 字串查询(SEARCH STRING)

字串查询(SEARCH STRING)区包含字符串查询(WORD(S))、上下文限定(CONTEXT)、词性列表(POS LIST)和用户列表(USER LISTS)。

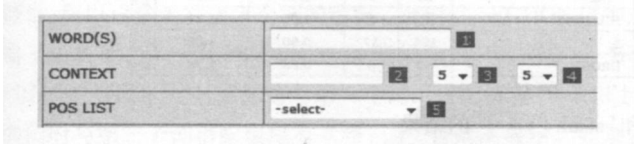


Figure 3 字串查询(SEARCH STRING)区示例

2.1.2.1 字符串查询(WORD(S))

默认情况下字符串查询(WORD(S))处([1])有一个对话框,供使用者输入要查询的字符串。输入的单词不分大小写,而且大小写混合输入也能查找到正确信息。此对话框中最多可以输入连续9个词的字符串,超过9个词则会有报错提示。

2.1.2.2 上下文限定(CONTEXT)

上下文限定(CONTEXT)区默认情况下是没有对话框的,将光标放在CONTEXT上单击,会出现一个对话框([2])和两个选择框([3][4]),在此对话框中用户可以输入希望与字符串查询(WORD(S))的字词在一定上下文中出现的字词。两个选择框中左边一个([3])表示CONTEXT中的词在字符串查询(WORD(S))的字词左边出现的最远位置,右边的选择框([4])表示CONTEXT中的词在字符串查询的字词右边出现的最远位置,默认情况下均为前后4个词的距离内。用户可以将两个选择框选择为最远9个词距离。当左边的选择框选择0时表示不考虑CONTEXT中的字词在字符串查询(WORD(S))的左边出现的情况,在右边选择框选择0时表示不考虑其右边出现的情况。

2.1.2.3 词性列表(POS LIST)

词性列表(POS LIST)区([5])默认情况下也是没有对话框,用户将光标放在POS LIST上单击就可以看到一个选择框。点击下拉箭头在各种词性中进行选择,选择出的词性将以简缩形式放在字符串查询区;若上下文限定区被打开则会放在上下文限定区,对所希望查找的字符串进行词性范围限定。词性列表(POS LIST)下拉框中共有39种词性分类,其中包括标点符号。

2.1.2.4 用户列表(USER LISTS)区

在字串查询(SEARCH STRING)区的最后有一个灰色的用户列表(USER LISTS)区。其作用是帮助用户建立自己的查询结果存储列表,方便今后调用。点击C区中的小方框就可以往用户列表添加选定内容。第一次使用时用户需要用一个名称命名自己建立的列表,随后可以参照本语料库网页关于用户列表的说明进行操作。

2.1.3 语料库分类区(SECTION)

语料库分类区(SECTION)有两个内容相同的部分,但其功能是不一样的。此区可以对查询的字符串限定语料类型(Genre)和时段(Year),并且可以明确到查询某一个子语料库,时段也可以查询任何一年的某个字词的使用情况。默认情况下语料库分类区的查询为IGNORE状态,即忽略语料库和时段的分类,在所有语料中查询。此处应注意的技巧是可以在点选了一个语料库后按住计算机键盘上的Ctrl键继续选择多个语料库或时段;另外在CHART显示结果后可在C区中点击SEE ALL SECTIONS也可以看到所查询的字符串在每个子语料库和每年中的细节信息。语料库分类区的两个选择框下均有MIN FREQ一栏,其下有一个数字框和一个勾选框,其意义是当用户勾选了后可以指定所查询字符串的最低出现频率,默认情况下为10。

这一项对查询同义词搭配以及过滤罕见搭配很有用。

2.1.4 查询结果排列方式区(CLICK TO SEE OPTIONS)

点击查询结果排列方式区(CLICK TO SEE OPTIONS)将显示四项内容,分别是: GROUP BY(按词形排列)、DISPLAY(显示方式)、SAVE LISTS(存储结果)和 #HITS(最多显示条数)。本区展开后可点击 HIDE OPTIONS隐藏显示的内容。

2.1.4.1 按词形排列(GROUP BY)

按词形排列(GROUP BY)下有 5项选择: LEMMAS、WORDS、NONE、BOTH WORDS、BOTH LEMMAS,默认状态下为 WORDS。按 LEMMAS排列是指所查询的词以原形词排列,但查询结果中包含动词时态的变化、形容词的级别变化以及名词的单复数等;按 WORDS查询时不考虑词的多重词性;按 NONE查询排列时会显示所查询词作不同词性的搭配情况,此时查出的词性详细标注在 CLAWS网页上有介绍;在使用 CONTEXT(上下文)查询时,同时也选择 BOTH WORDS或 BOTH LEMMAS查询选项,结果能返回所查询词汇的多重搭配结果,尤其是在查询同义词时这一选项更能发挥作用。

2.1.4.2 显示方式(DISPLAY)

此处的显示方式(DISPLAY)下有 4项选择: RAW FREQ(字符串总词频)、PER/MIL(每百万词频)、RAW FREQ + (总词频和每百万词频)、以及 PER/MIL + (每百万词频和总词频)。即表示查询结果在 C区的呈现方式,选择后两项会在每一个框内各出现两项信息。要使这些设置有实质作用,需要将 B区的第一大项 DISPLAY选择为 CHART及 SECTIONS选择为 YES。

2.1.4.3 存储结果区(SAVE LISTS)及最多显示条数(#HITS)

存储结果区(SAVE LISTS)容许用户将查询结果存入自己的列表(需先建立 USER LISTS),供今后调用。最多显示条数(#HITS)的默认值为 100,表示将合乎查询条件的 100项查询结果显示于 C区的情形。用户可以根据查询需要对这一数值进行修改,但最大值为 1000。

2.2 查询结果数据显示区(C)

此区在(B)设定为 LIST显示时将呈现出前 100个查询结果(默认)。下面以查'laugh [n*]'(CONTEXT框内为 *,左右均设定为 5, SORT此时自动更改为 RELEVANCE,勾选 Section 1处的 MIN FREQ为 10)返回的结果做说明。

Table 4 查'laugh [n*]返回的前 5项信息

	CONTEXT	TOT	ALL	%	MI
1	MIRTHLESS	21	59	35.59	7.68
2	THROATY	45	327	13.76	6.73
3	RUEFUL	19	278	6.83	6.03
4	CACKLING	12	176	6.82	6.03
5	HEARTY	94	1448	6.49	5.98

此表的含义就是查找在 laugh作名词时其上下文左右各 5个词语内与其搭配最紧密的词的情况,由于形容词通常与名词搭配,因此显示结果的前 5项均为形容词。TOT表示 laugh作名词时与上下文各 5个词距离内所查到的词(如 mirthless)总数; ALL表示所查到的词(如 mirthless)在所有语料库中的总词数; %即表示 TOT与 ALL的商; 互信息 MI(Mutual Information)的值表示 mirthless与 laugh(作名词)之间的搭配紧密度,互信息越高说明二者联合使用的频率越高。点击 CONTEXT中的词语或 TOT中的数字,细节信息将在 Figure 1中的 D区(例句显示区)展现。也可以选择点击 CONTEXT前一栏中的选择方框,然后点击 CONTEXT,则在 D区显示的就是已被选择了的词语上下文信息,没有被选择的则不在 D区显示。

查询结果数据显示区(C)在(B)设定为 CHART显示时将呈现出如 Figure 5一样的柱状图。这时选择子语料库没有实际意义。如果在 B区的 DISPLAY选择为 COMPARE WORDS并查询,在 C区将有两个对比图表。以下为在 COMPARE WORDS的两个框中分别输入'rob. [v*]和'steal. [v*]',在 CONTEXT

SEE CONTEXT: CLICK ON WORD (ALL SECTIONS) OR NUMBER (SPECIFIED SECTION) [HELP...]

WORD 1 (W1): ROB (50.0%)

	WORD	W1	W2	PERC-1	SCORE
1	BANKS	112	5	0.96	1.91
2	BANK	199	17	0.92	1.84
3	STORE	66	10	0.87	1.74
4	GUNPOINT	47	11	0.81	1.62
5	VICTIMS	14	5	0.74	1.47
6	SLEEP	11	4	0.73	1.47
7	PLACE	14	7	0.67	1.33
8	RAPE	10	5	0.67	1.33
9	HOUSE	37	19	0.66	1.32
10	GROCERY	11	7	0.61	1.22

WORD 2 (W2): STEAL (50.0%)

	WORD	W2	W1	PERC-2	SCORE
1	CARS	155	4	0.97	1.95
2	MONEY	415	12	0.97	1.94
3	CAR	258	10	0.96	1.93
4	FOOD	129	5	0.96	1.93
5	TRUCK	47	4	0.92	1.84
6	FATHER	38	4	0.90	1.81
7	TIME	66	7	0.90	1.81
8	MILLIONS	47	5	0.90	1.81
9	$	156	17	0.90	1.80
10	DRUGS	43	6	0.88	1.76

Figure 4 查询'rob. [v*]'和'steal. [v*]'的结果

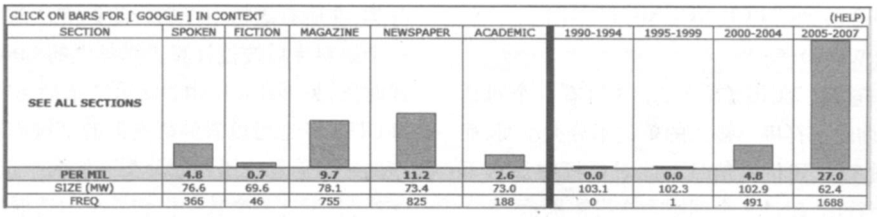


Figure 5 查 Google 在五大语料中以及四个时间段内的分布结果

Table 5 字符串查询(WORD(S))处输入‘输入词示例’内容的说明

输入词示例	作用	说明与技巧
jumbo或 soft landing	查具体的词或短语	也可以输入长字符串(9词以下)
borrow/lend	简单对比两个词的使用频率	/表示‘或者’;在 LIST中上下排列结果;选择 SHOW SECTIONS时结果更直观
fairly *	查与 fairly搭配的情况	*是通配符,此处代指任一个词;注意:fairly与 *之间有空格
un * ly	查以 un开头以 ly结尾的词	查到单词如 unlikely, unusually;此处 *代指任意数量的字母
[slip]. [v*]	查 slip作动词的情况	查到 slipped、slip、slipping和 slips列表
* heart *	查含 heart的词	含 heart本身单复数及用其它含 heart的派生词及合成词
r? n *	查以 r与 n之间为一个字母的词并且后面跟若干字母的词	查到 run, running, ran等;?代指一个字符。技巧:只知道一个词长度和首尾字母便可以查询到这个词,对纵横字谜的单词查询很有帮助
[sing]	查 sing的任何形式	查到 sing, singing, sang等,但不包括 song
[=publish]	查 publish的同义词	=表示同义关系,结果为 publish, circulate, announce等。说明:查同义词是 COCA语料库的一大特色
[=publish]	查 publish的同义词并且不限词形	结果有 announce, circulated, publishes等
[=knock] the door	查与 the door搭配的与 knock同义的动词	有 slam, hit, crack, pound, bash等。技巧:对选择最佳搭配词很有效
thick [nn*]	查 thick与名词的搭配情况	各种词性的简缩式可以通过在 POS LIST选择后查看到
un * ed [j*]	查以 un开头 ed结尾的形容词	查到 united, unexpected, unprecedented, unidentified,等
dis *. [v? d]	查以 dis开头并且词形为过去式的情况	查到如 discovered, disappeared, discussed等
dis * [v? d]	查第一个词以 dis开头,下一个词为过去式结构	(注意与上面的区别)查到 district had, disease had等
* ly. [j*]	查以 ly结尾的形容词	仅查 ly结尾的词作形容词的使用情况
* ly. [r*] * ly. [j*]	查以 ly结尾的副词修饰以 ly结尾的形容词	结果有 highly unlikely, environmentally friendly, potentially deadly等
[xx] * without	查否定词 not/n't +任一词 +without的情形	由于否定词后通常接动词,因此通配符 *此处也可以换成动词[v*]
it's [j*] that	查 is被缩写为 's情况的句式结构	's在本语料库中可以被视为一个词单独查询,其它缩写形式也是用类似方法查询
it is [v*] that或 we [vv*] that	查句式结构	选择 CHART显示可以看出第一个是学术结构,是口语的 8.5倍;第二个结构口语中最常用
to [v*] or not to [v*]	查类似名句‘to be or not to be’式的结构	[v*]处的前 5个动词分别是:be、do、buy、tell和 engage
[vv*] * into [v? g]	查动词接任一词再接 into V-ing结构	查到含致使意义的 fool you into thinking, talked him into going, trick people into thinking等
[=clean]. [v*] the [n*]	查作动词的 clean及其同义词的不同词形与被 the修饰的名词搭配的情况	查到 wiped the sweat, mopping the floor等

注:并非所有的同义词在不同上下文中均能成为同义词,因此查询结果可能会不相关。

框中输入‘[nn*]’,上下文的两个框分别为 0和 3的查询结果。以上条件的含义是比较 rob和 steal作动词时其右边 3个词内的名词情况,由于名词通常有冠词、物主代词、或形容词等修饰,所以设定为右边 3个词的情况。此时在 SECTION区均忽略子语料库,第一个 MIN FREQ的值设定为 10,第二个 MIN FREQ的值设定为 4,意思是与第一个词的搭配应有 10条以上,但同时与第二个词搭配也应有 4条以上,否则不显示。如果希望加大两个词的对比度,可以将与第二个词的搭配设置得更低甚至为 0。

查询结果表明(Figure 4) rob与 banks上述条件搭配有 112次(W1),而 steal与 banks搭配只有 5次(W2),所以 W1占所有搭配的 0.96(PERC-1), PERC-1与 rob在 rob和 steal中的比例(50%)的商为 1.91(SCORE)。这是用语料库数据证明许多文章论述的“‘抢’可以隐去被抢物,‘偷’不可隐去被偷物”的论断。

2.3 例句显示区(D)

在查询结果数据显示区(C)中点击数据链接后,所查询字符串的上下文将在 D区以列表显示,每一个上下文显示约 29个单词。例句列表前面有 4栏信息,分别是例句序号、年代、语料库大类和子语料库及细节。点击其中任何一项,将在本区呈现这一例句的更详细内容,包括本例出现的年代、出版信息、文章标题、作者信息和语料库分类,以及更长的上下文。由于版权原因,这时所显示的字词上下文目前在所查询词语的左右最多各显示 120词左右,但已经足够使用者从上下文中找到所查询字词的信息。本区使用的是 KWIC关键词搜索显示软件。

3 具体运用实例

运用语料库查询字词在不同语料中的分布以及各个时段

的分布是 COCA美国当代英语语料库最常见的用法。如用户想查询‘Google’一词在 COCA语料库中的频率以了解这个词的各项分布状况,可以不选择其它查询条件界定,直接在字符串查询(WORD(S))处输入 Google点击 SEARCH按钮或按键盘上的回车键,结果将用简洁的 LIST列表返回(共 2180次)。重新选定 DISPLAY(显示方式)中的 CHART(柱状图表显示)再搜索,返回的结果将会让用户直观地看到 Google一词在五大语料中以及四个时间段内的分布(Figure 5)。返回的数据共有 4行:第一行为柱状图,是依据第二行的每百万词频率(PER MIL)计算显示,第三行为各语料库库容,第四行为各语料库中所查询词语的总词数。这是 Google在自己的 Google网页上 google不出的数据。

点击‘SEE ALL SECTIONS’又可以看到一个列表,将 Google一词在各个子语料库中的数据按每百万词数的比率呈现,同时用户还可以看到 Google在各子语料库中的总词数和各个子语料库词总数。继续下拉列表还可以看到 Google一词在每年中的数据,这样查询比用语料库分类区(SECTION)的列表简单得多。

3.1 简单查询实例

以上查询可以根据用户需要组合进行,如想知道 terrorism的搭配情况可以输入‘* terrorism’就可以查到 terrorism的前面的一个词的情况,如果限定为‘[v*] terrorism’就可查询任何动词的任何词形与 terrorism的搭配。但动词短语也会与名词搭配,所以可以用动词加任一词的情况查询,就输入‘[v*] * terrorism’。如果用户记得 terrorism的前面为包含动词共三个词,那么只需要再加一个通配符 *即可查找到。

用户在用词性限定所查询内容时可以在 POS LIST下拉框中看到 COCA美国当代英语语料库所使用的 39种词性分类,既可以点击相应词类输入,也可以直接在对话框中输入其简缩形式。使用者也可以参考网页 http://ucrel.lancs.ac.uk/claws7tags.html,这是 COCA美国当代英语语料库所用的标注软件 CLAWS7的一个更细致的标注标准,它对英语字词进行了137个分类。用户参照这一分类时也需要将词性缩写放在方括号[]内对查询内容进行限定。

3.2 组合查询实例

在 Figure 3中 WORD(S)对话框内([1])输入‘chip.[n*]’,在 CONTEXT对话框([2])中输入‘[n*]’,在([3][4])处各选择默认的数字 5,然后点击 SEARCH按钮。此查找的结果为查询在 chip前或后 5个词距离内的名词,返回的前 5个最高词频的词为 chocolate, computer, chip, cookies和 Intel。更多组合查询实例参看 Table 6,用户在实际使用中可以组合出无穷的查询搭配。

3.3 技巧运用

词的搭配情况可以用 COCA美国当代英语语料库很好地获取。如学习者只知道 toilet一词却不知道该选择什么样的动词,就可以输入‘[v*] * toilet’查询,得出的高频搭配动词为 use和 flush。词语的对比情况记住一定要选择[SORT]为 RELEVANCE才有实际意义。

若查‘distinguish [v*]’(设定 CHART显示)可以看出 distinguish在学术期刊中的使用频率是口语的 6倍,因而应当尽可能在口语中少用。若在 WORD(S)处输入‘[j*] action’,在 SECTION 1中选择 FIC,在 SECTION 2中选择 ACAD,可以比较出小说和学术文章中使用形容词修饰 action的显著差异,尤其是前 10个搭配。因此这样一个出众的语料库对于英语教师和学习者认识美国英语、比较词语在口语、小说和学术文章等语料中的使用状况、了解词汇的搭配和发展等均是不可多得的优秀资源(Davies, 2008)。

在许多查询条件后面都有一个“?”号,用户可以点击并在例句显示区看到相关英文解释。对各板块信息不明白时可以点击其中的[HELP...]链接。尤其应充分利用的是 D区右上方的 More information…,其下拉箭头会将使用者带入各个板块的详细介绍。如果发现查找结果与自己希望显示的方式不一样时请点击 RESET后重新输入查找内容,或者每次查询新内容时都养成先点击 RESET的习惯。

4 评述

4.1 优点

许多语言学习者对普通网络搜索引擎有偏好,那是无法获取免费语料库的无奈之举。但网络信息作为语料库的问题就是用户无法限定要查询字词的词性,无法作词与词的对比,无法限定要查找的语料类型,也无法确切地查找某一时段的字词使用信息,更无法限定要查找的字词间的距离,也就没有办法

Table 6 组合查询列表

WORD(S) [1]	CONTEXT [2]	[3]/[4]	SORT	Function
study [n*]	[j*]	1/0	RELEVANCE	查与 study作名词搭配的任一形容词,结果有 present, recent等
with.	[v*]	1/0	FREQUENCY	查动词与 with搭配并位于句子结尾的情况。注:标点符号在本语料库中可以独立查询
[=beautiful]	[=flower]	5/5	BOTH WORDS	查 beautiful的同义词与 flower的同义词搭配的情况
small little	[nn*]	0/3	RELEVANCE	各查 small和 little后面 3个词内的名词使用对比情况
ground [n*] floor [n*]	[j*]	3/0	RELEVANCE	各查 ground和 floor作名词时前面 3个词内的形容词使用对比情况
statesman politician	[j*]	2/0	FREQUENCY	查 statesman和 politician前 2个词为形容词的情形,按频率排列
de*. [vvi*]	SECTION 1 = ACAD	SECTION 2 = FIC		查以 de开头的动词不定式在学术期刊和小说中的情况
[=smart]	SECTION 1 = NEWS	SECTION 2 = FIC		查 smart的同义词在报纸和小说中的使用情况

确定字词互信息的值。而其它专业点的英语语料库又需要较贵的注册费或软件购买费用,让普通使用者望尘莫及。COCA美国当代英语语料库是一个大型在线并免费供大家使用的语料库,为英语研究者和英语学习者共享美国英语资源提供了一个良好平台。它能比较 lemmas,如输入[borrow]和[lend]查询是网络搜索引擎做不到的。COCA美国当代英语语料库的优点是很突出的,用户只要充分利用 COCA美国当代英语语料库,以及 Mark Davies教授进一步用与 COCA美国当代英语语料库相同界面处理过的其它语料库如 BNC和 Time Corpus(参看 http://corpus.byu.edu/)等,并适当结合 Google或者 Yahoo搜索引擎,可以获得自己所需要的大量有价值的研究数据。

COCA美国当代英语语料库相比其它语料库(或软件)有许多优势: R❶需要有一定的编程能力; Wordcruncher的入门较复杂; WordSmith查询处理速度太慢。对于想了解 COCA语料库的体系结构等内容的,可以参看 Mark Davies教授的最新论著(Davies 2007; Davies 2008)。COCA美国当代英语语料库的界面主要是为语言学家和语言学习者了解单词、短语以及句子结构的频率及进行相关信息比较而设计,而词频数据几乎就是用事实说话的一个体现。充分挖掘 COCA美国当代英语语料库这一宝藏,将会给我们研究和学习最新地道的美国英语带来革命性变化。

4.2 不足之处

COCA美国当代英语语料库的界面总体来讲比较具有亲和力(user-friendly)。但光标放在左边"查询条件界定区"的 SEARCH STRING中的 CONTEXT和 POS LIST以及 CLICK TO SEE OPTIONS时不能显示出超链接状态,很容易让使用者误以为没有信息展示。这一问题在今后会很好地解决。

COCA美国当代英语语料库子语料库分类没有 BNC那样细致(BNC有 70个子类),只是说明了语料来源。D区 Concordance的排列不如在 WordSmith等中整齐直观,主要是因为在 B区若已限定排列方式(SORT),则相应在 C区列表中已显示了查询的字符串最常用搭配,而不需要用户在 Concordance的排列中自行查找统计最常见搭配。

有些词的归类似乎还有疑问,如 one在 one another中就分别被冠以 pron. PERS(人称代词)和 pron. NDF(不定代词)可以被查询到,分别为 13248次和 44次。当然这更多地取决于词性标注软件 CLAWS7的准确性。COCA美国当代英语语料库中大小写不区分看似会给查询专有名词带来困难,但用户可以限定所查询词的词性如 brown.[np*]和 brown.[j*]来区分。总的情况就是一瑕不掩万瑜,充分利用其优势的一面,会给英语教师和学习者带来无穷惊喜和收获。□

参 考 文 献

[1] Davies, Mark. The 360 million word Corpus of Contemporary American English (1990 - 2007) [A]. Unpublished manuscript Paper presented at American Association for Corpus Linguistics, 2008.

[2] Davies, Mark. (forthcoming) Relational databases as a robust architecture for the analysis of word frequency [A]. In AHRC ICT Methods Network: Expert Seminar on Linguistics: Word Frequency and Keyword Extraction [C], ed. Dawn Archer. Ashgate.

[3] Davies, Mark. Semantically-based queries with a joint BNC/WordNet database [A]. In Corpus Linguistics twenty-five years on, ed. Roberta Facchinetti. Amsterdam: Rodopi, 2007: 149 - 167.

[4] 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.

A Good Platform for English Teachers and Learners: the Corpus of Contemporary American English (COCA)

WANG Xing-fu[1], Mark Davies[2], LIU Guo-hui[3]

(1. College of Foreign Languages, Chongqing University, Chongqing 400030, China;
2. Brigham Young University, Provo, Utah, USA 84602;
3. College of Foreign Languages, Hangzhou Normal University, Hangzhou, Zhejiang 310036, China)

Abstract: The article is an introduction to the Corpus of Contemporary American English (COCA) that is created by Prof. Mark Davies. This corpus contains 360 million words of the materials published in America from 1990 to 2007, and it is the largest balanced English corpus. The Corpus of Contemporary American English is free for researchers and English learners to use online.

Key words: Corpus of Contemporary American English; Balanced English Corpus

❶ R是一个统计计算项目的代称,也代表其程序。是由 University of California, Santa Barbara大学开发的。详见 www.r-project.org/.