多语种在线语料库检索平台 BFSU CQPweb 使用简明手册

许家金
中国外语与教育研究中心

1、访问及登录

访问 **http://111.200.194.212/cqp/**（用户名：**test** 和密码：**test**），可点击使用相应的语料库。目前 BFSU CQPweb 平台上已安装英语、汉语、德语、日语、俄语、阿拉伯语、冰岛语等数十个语料库。



图 1：BFSU CQPweb 主界面

2、CQPweb 功能概要

按 McEnery & Hardie（2012）对语料库分析工具的时代划分，CQPweb 属于第四代语料库工具，即在线语料库分析工具。四代工具的突出代表是美国杨百翰（Brigham Young）大学 Mark Davies 教授创建的 BYU 系列语料库检索界面（http://corpus.byu.edu/）。类似的在线语料库检索系统还有 SketchEngine、CWB、BNCweb、Phrase in English 等。而当前主流的语料库工具属于第三代，其中以 WordSmith、AntConc 和 PowerConc 等为代表。

第四代语料库工具，将语料库与分析工具合二为一，越来越受到普通用户的青睐。在线语料库工具通常将语料库文本按特定格式建成索引（index），存储在服务器上。用户检索响应速度要远高于三代软件在本地电脑上的检索速度。其操作也较三代语料库软件简便得多。

四代语料库工具可完成三代语料库几乎所有的功能，其中又以 CQPweb 所能实现的功能最多最全。更重的是，CQPweb 是开源软件。概括说来，CQPweb 可以实现以下功能。

（1）在线生成语料库的词频表（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（frequency），并可以按语体、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）计算语料库中的核心关键词（keywords），等。

3、CQPweb 使用实例

3.1 标准查询模式

在简单查询模式（Simple query mode）下，可输入单词、短语等进行检索。

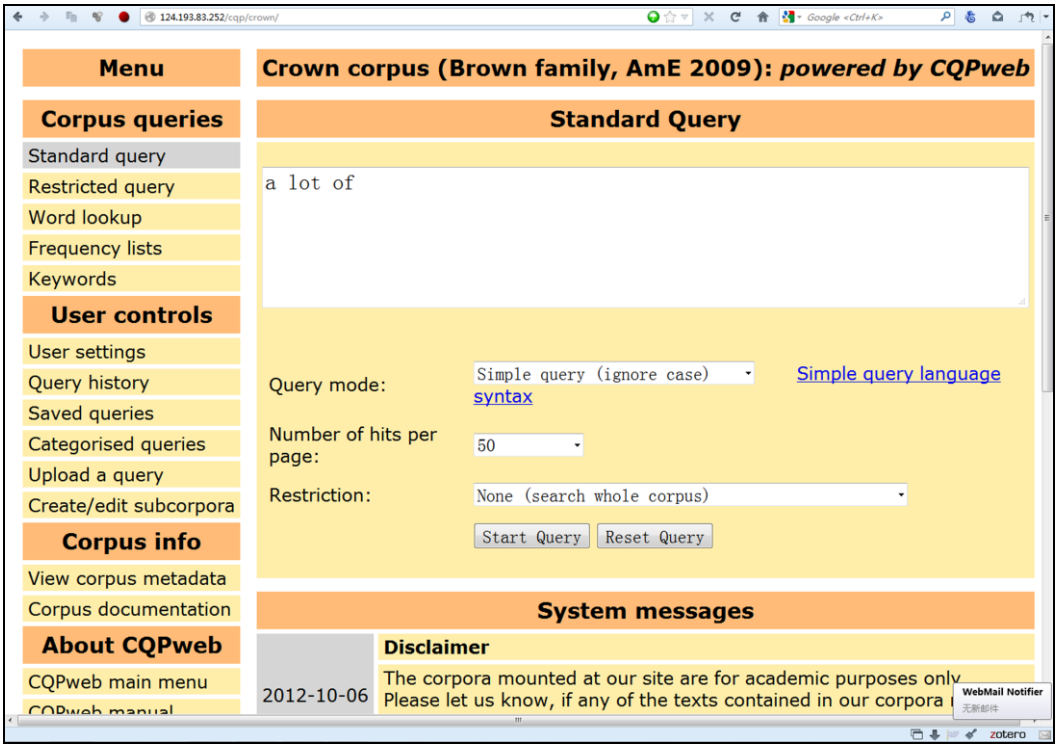


图 2：BFSU CQPweb 语料库查询界面

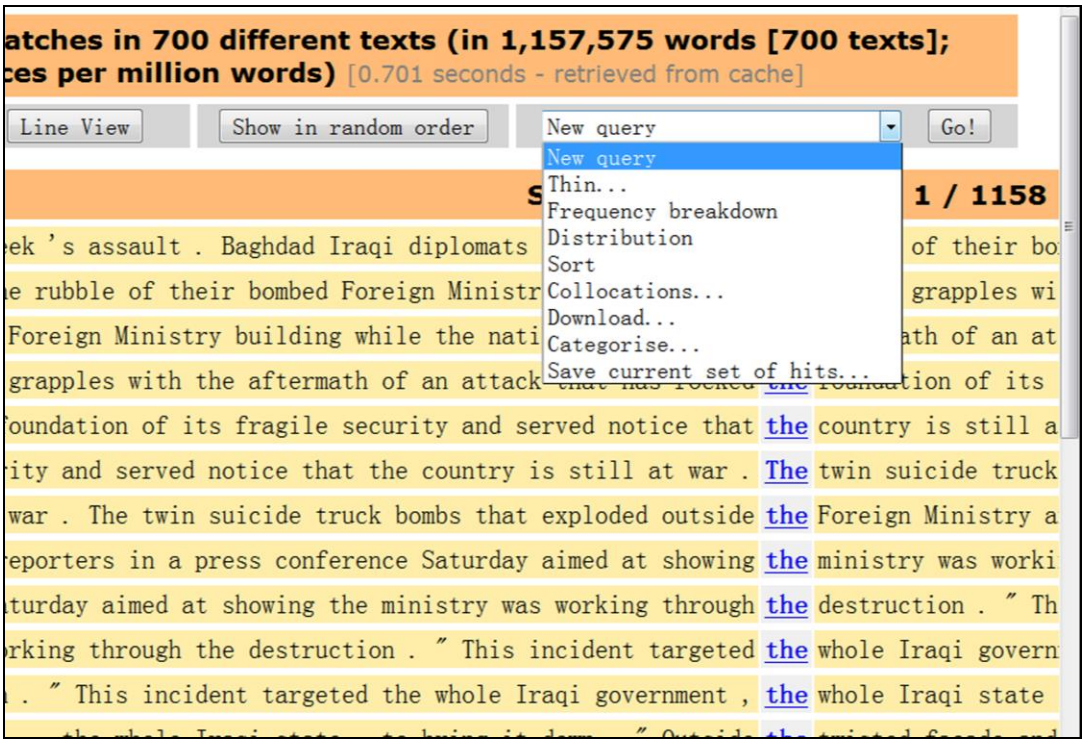


图 3：BFSU CQPweb 查询结果界面

点击查询结果页面右上角下拉菜单，显示 New Query（新查询）时，按 Go 键，即可重新回到语料库检索界面。相当于“返回”按钮。

表 1：查询结果后续操作分项功能表

New query	新查询，返回语料库检索首页
Thin	查询结果随机抽样
Frequency breakdown	频数分解、分解频数
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

Thin（随机取样），比如，可从 5 万行结果中，随机抽取 100 行。

Frequency breakdown（频数分解）表示在进行复杂查询时，对命中的不同词项分别计算频数。比如，查询"am"|"is"|"are"|"was"|"were"|"be"|"being"|"been"时，会按这 7 个词项分别报告命中频数和频率。

No.	Search result	No. of occurrences	Percent
1	is	9713	27.28%
2	was	8955	25.15%
3	be	5697	16%
4	are	4488	12.6%
5	were	2867	8.05%
6	been	2751	7.73%
7	being	886	2.49%
8	am	250	0.7%

图 4：be 动词查询 Frequency breakdown（频数分解）结果示例

Distribution：按语体、年代、章节、学生语言水平、写作题材等分别呈现查询结果

Categories: General information　Show as: Distribution table

Category for crosstabs: No crosstabs　Show distribution　Go!

Based on classification: Category

Category [↓]	Words in category	Hits in category	Dispersion (no. files with 1+ hits)	Frequency [↓] per million words in category
A: News Reportage	101707	12	9 out of 122	117.99
B: Editorial	61409	6	5 out of 74	97.71
C: Reviews (theatre, books, music, dance)	39987	21	14 out of 41	525.17
D: Religion	39298	52	8 out of 21	1323.22
E: Skills and Hobbies	80799	18	10 out of 49	222.78
F: Popular Lore	112052	64	17 out of 56	571.16
G: Belles Lettres, Biography, Memoirs, etc.	175864	142	47 out of 86	807.44
H: Miscellaneous non-fiction	67580	0	0 out of 30	0
J: Learned (academic writing)	173780	14	7 out of 94	80.56
K: General Fiction	69177	45	24 out of 29	650.51
L: Mystery and Detective Fiction	57938	37	15 out of 24	638.61
M: Science Fiction	14421	9	3 out of 6	624.09
N: Adventure and Western Fiction	69781	19	12 out of 30	272.28
P: Romance and Love Story	72100	82	25 out of 29	1137.31
R: Humor	21682	11	6 out of 9	507.33

图 5：Crown 语料库中"lov.*"的分布情况（Distribution table）

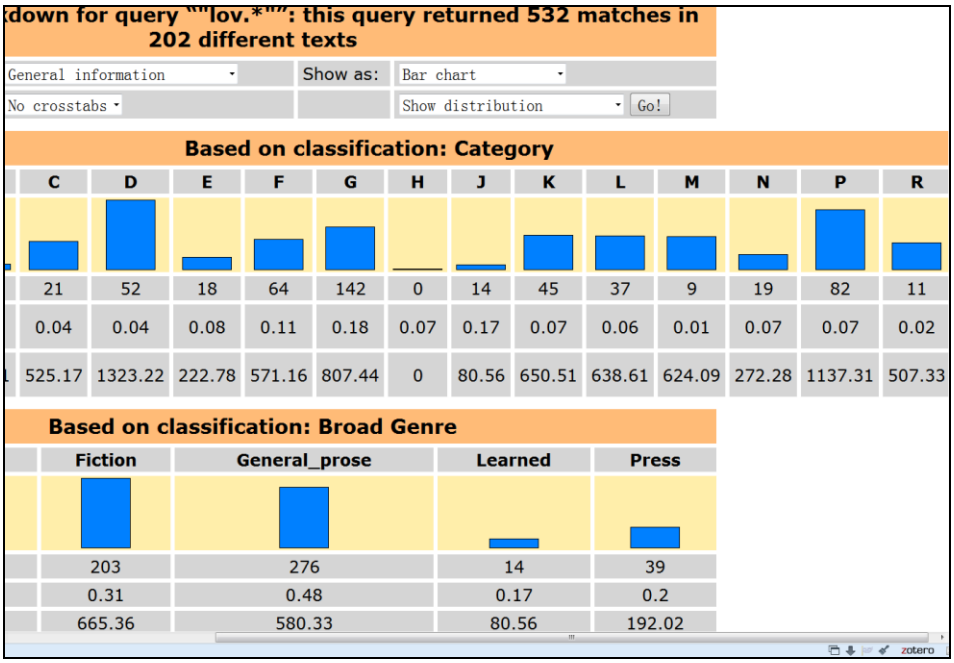


图 6：Crown 语料库中"lov.*"的分布情况（Bar chart）

Collocations：计算特定词语在语料库中的典型搭配

Collocation controls

Collocation based on:	Word form	Statistic:	Log-likelihood
Collocation window *from*:	3 to the Left	Collocation window *to*:	3 to the Right
Freq(node, collocate) at least:	5	Freq(collocate) at least:	5
Filter results by:	specific collocate:	and/or tag: (none)	Submit changed parameters Go!

There are 1,348 different words in your collocation database for ""lov.*"". (Your query ""lov.*"" returned 532 matches in 202 different texts) [3.212 seconds]

No.	Word	Total no. in whole corpus	Expected collocate frequency	Observed collocate frequency	In no. of texts	Log-likelihood value
1	I	7,888	21.751	89	50	118.314
2	you	5,074	13.992	58	27	77.93
3	God	376	1.037	16	10	58.315
4	ones	148	0.408	9	9	39.03

图 7：Crown 语料库中"lov.*"的典型搭配词（以 log likelihood value 排序）

3.2 限定条件查询

限定条件查询，指在任务一开始，即选定一个或多个限制条件（如，语体、年代、章节、写作题材等）进行查询。限定条件的有无、多寡，源自语料库文本的元信息（metadata）。因此，在创建语料库时，应尽可能详细记录语料文本产生的社会语言学信息。丰富的社会语言学信息，可以大大丰富研究的层面和深度。这样的元信息可以存储在文本的头部，也可以在文本之外单独存储。

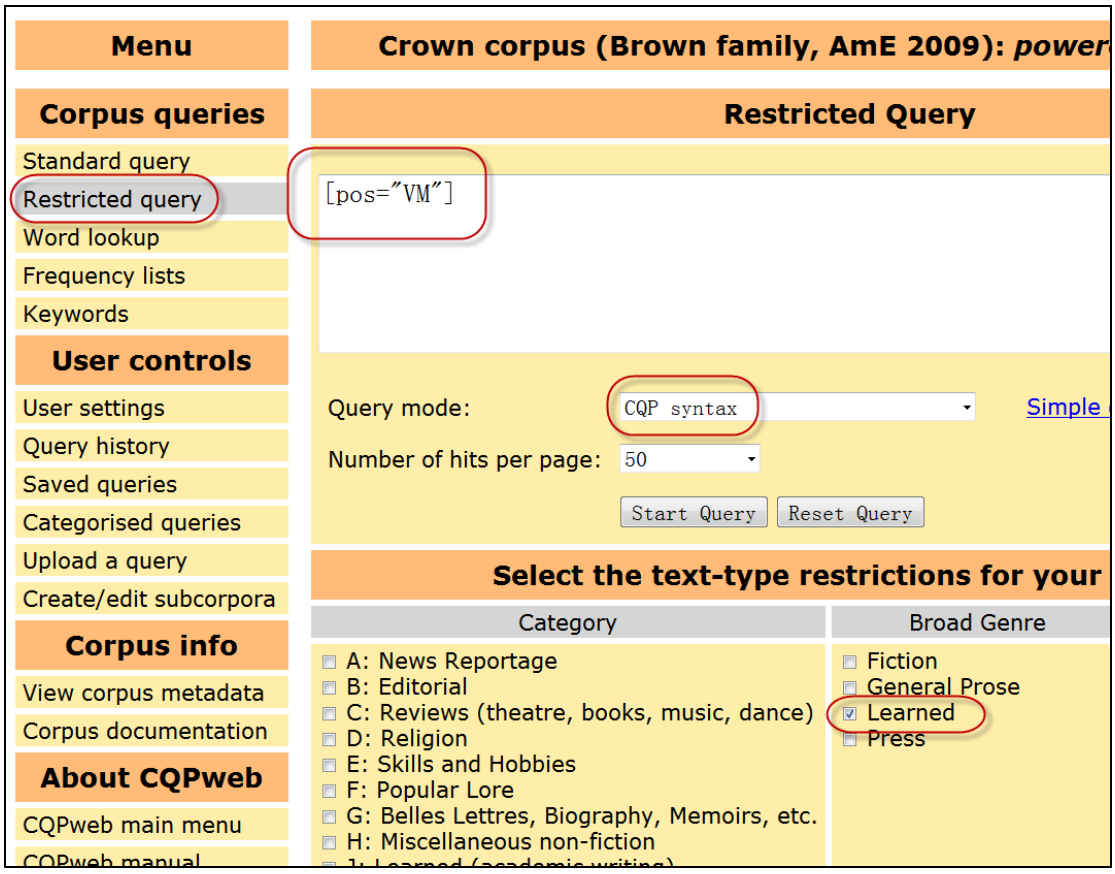


图 8：限定在 Crown 语料库的学术语体中查询情态动词

3.3 生成词频表

Frequency list: Word frequencies in entire "Crown corpus (Brown family, AmE 2009)"

|< << >> New Frequency List Go!

No.	Word	Frequency
1	,	59,062
2	the	57,885
3	.	50,435
4	and	28,173
5	of	27,655
6	to	27,348
7	a	23,370
8	"	19,799
9	in	19,217
10	that	12,359

图 9： Crown 语料库的词频表

3.4 生成主题词表

比如以《红楼梦》与 Crown 语料库进行对比，可能得到《红楼梦》的主题性词汇。

2	Lady	3,233	51	+	5515.02
3	you	10,461	5,074	+	4868.43
4	"	23,700	19,799	+	3920.7
5	she	8,699	4,411	+	3818.53
6	her	8,716	4,664	+	3548.78
7	)	28	3,169	-	2926.37
8	(	26	3,136	-	2909.02
9	--	125	3,548	-	2718.51
10	the	26,726	57,885	-	2337.63
11	of	10,816	27,655	-	2218.99
12	!	2,553	631	+	2195.69
13	?	5,011	2,698	+	2022.11
14	old	2,030	416	+	1949.6
15	n't	4,901	3,361	+	1304.19
16	'll	1,594	485	+	1183.5
17	in	8,097	19,217	-	1175.07
18	master	818	59	+	1142.55

4、BFSU CQPweb 多语种语料库建设思路

本族语平衡语料库：1 百万词次以上

特定语体语域专门用途语料库：比如文学作品、新闻报导、法律文本、网络文本等

学习者语料库：学习者作文、翻译练习

翻译文本及平行语料库

附录 1：CQPweb 平台中英文术语对照表（表 2）

(word) token	词次
(word) type	词种
Collocation	词语搭配
Corpus documentation	语料库说明文档
Corpus metadata	语料库元信息
CQP syntax	CQP 复杂检索语法
Distribution	分布（按语体等分类条件分别呈现结果）
Frequency	频数、频率
Frequency breakdown	频数分解、分解频数
Frequency list	词频表、词表
Frequency list	词频表
Keywords	主题词
Log likelihood (ratio)	对数似然率（典型词语搭配的统计方法）
Maximum window span	最大跨距（计算搭配时中心词和左右语境词之间的距离）
No. of occurrences	出现次数
Node	检索词、中心词、节点词
Number of hits per page	查询结果每页显示的行数
Query	查询、检索
Restricted query	限定条件查询
returned 35,607 matches in 743 different texts	直译：在 743 个不同文本中返回 35,607 个匹配项 意译：在 743 个文本中查到 35,607 例子
Show hits sorted by node	查询结果按中心词排序
Simple query (ignore case)	简单查询（不区分大小写）
Word lookup	词语相关查询

附录 2：复杂检索举例（查询时，选择 CQP Syntax）

单词检索："lov.*"、".*ment"、"am"|"is"|"are"|"was"|"were"|"be"|"being"|"been"

词码混合检索："good" [pos="N.*"]、"happy" [pos="N.*"]、[pos="J.*"] "love"、[pos="J.*"] [pos="J.*"] [pos="N.*"]、"something" [pos="J.*"]

近义词批量检索："little"|"small"、/color[]

北外语料库语言学团队网站：http://corpus.bfsu.edu.cn

使用北外 CQPweb 平台，可引用：

许家金、吴良平，2014，基于网络的第四代语料库分析工具 CQPweb 及应用实例，《外语电化教学》（5）：10-15，56。