

拓展历史认知语言学研究新视野的工具

——美语历史语料库(COHA)的应用*

李华勇

(四川文理学院 外语学院,四川 达州 635006)

摘 要:近 20 年来多年来,200 多年前开创的历史语言学和现代认知语言学逐渐开始融合诞生了历史认知语言学。认知语言学有“基于用法的模式”来研究语言事实背后的理据性的传统,这使历史认知语言学就有与语料库语言学结合的天然优势来观察相关语言现象和语言结构的历史变化过程、追寻其历史变化事实背后的内在规律。文章以具体的实例分别详细说明了 COHA 在对美语词汇、形态、句法、语义到文化方面的作用,这些实例说明 COHA 极大地拓展了历史认知语言学研究的新视野,也为历史认知语言学研究提供了强有力的工具保证。

关键词:COHA;历史语言学;认知语言学;语料库

中图分类号:H0 **文献标志码:**A **文章编号:**1672-0598(2013)06-0115-10

一、历史认知语言学

历史认知语言学(Historical Cognitive Linguistics;简称 HCL)是历史语言学和认知语言相互渗透和相互融合的一门交叉学科;二十世纪九十年代,历史语言学开始关注认知语言学,(通过相关国际会议)两者开始携手合作,开启两者融合之路;2010 年《历史认知语言学》的出版标志着 HCL 的正式诞生(王寅,2012:927)。历史语言学(Historical Linguistics;简称 HL)是研究语言演变历史的学科(徐通锵,1991:1),它以历史文献为基础,以时间为主线来研究语言在不同时间段的状态、结构、渗透和变异过程与规律的语言学分支学科;它作为一门正式的学科约有 200 多年的历史。由于历史文献无法直接被感知,同时其间的证据还有可能缺失或不连贯,因此,历史语言学的研究者

们不得不对这些有限的文献采用相关理论和方法进行“推演和重构”,如徐通锵(1991)提出的“历史比较发、结构分析法”。

随着认知科学的发展,认知语言学(Cognitive Linguistics,简称 CL)近几十年来发展迅速,它不断拓展研究范围和方法,早已走了出了“认知语义学和认知语法”的视野,近年来先后出现了认知语音学、认知词汇学、认知辞书学、认知语篇学、认知语用学、认知诗学和认知社会语言学等。其中,语言研究者们自然也不会忽视 HL 和 CL 的融合:HL 从 1990s 年代开始关注 CL,借用其理论和方法来解释语言变化的一般理论、方法和原则。1990 年和 1996 年在德国召开了由认知语言学家和历史语言学家参与的以“共时与历史:语言历史与认知”和“历史语义学和认知”为主题的国际

* [收稿日期]2013-08-14

[基金项目]四川省教育厅 2013 年资助科研项目(13SB0136)

[作者简介]李华勇(1979—),男,苗族,四川达州人;四川文理学院外国语学院讲师,英语硕士,主要从事认知语言学及英语语料库研究。

研讨会。2007年8月召开的第10届国际认知语言学大会,学者们开始主张在CL理论框架下审视HL的相关问题,积极推动HL和CL的研究相互渗透和融合,主张HCL的研究思路,以挖掘边缘学科的潜在优势。Winters *et al.* (2010)主编的*Historical Cognitive Linguistics* (《历史认知语言学》)的出版,标志着历史语言学(HCL)这门学科正式问世。王演(2012:928)对HCL作了如下定义:它是以认知语言学为理论框架,追寻语言各层面演变的原因和过程,探寻其中的认知规律,分析其间的各种关系,重构扩散和变异的路径,从共时和历史两个维度深入研究语言诸方面的问题。

二、历史认知语言学与语料库的结合

HL重视分析和解释历史文献资料,从中获取有用信息,追源语言变化发展规律。CL也是很重视这一研究方法,并推动其继续向前发展——进一步在理论上构建了“基于用法的理论”(Usage-based Theory,简称UBT),这是后现代语言研究的一条重要指导原则,也使HL获得更坚实的理论依据。融合了HL和CL的HCL必然重视对大量真实语料的分析,对“频率”产生浓厚的兴趣;随着计算机和互联网的发展,各种界面友好和检索功能强大的语料库随之创建起来,为历史认知语言学与语料库的进一步结合提供了坚实的工具基础和技术基础;在语料库中,通过计算机对其中大量语料的自动检索,很容易获得语言中的各类结构(如词、短语、句法结构、语义结构)的频率。以语料库为基础获得的语言结构的这些频率数据,为解释和说明相应语言结构历史变化规律提供了坚持实证据,避免了依靠以语言研者的直觉为基础的內省式推理的缺陷。Taylor (2012)指出,整个语系统都是具有“频率效应”(frequency effects)的——同类结构中,有些结构比其它结构出现的频率高;在同一结构中,有些词语比其它词语使用频率高。语言结构的使用频率对语言变化和发展起着重要作用,高频使用可以促进某些语言结构在个体心智中的固化(entrenchment)和促使其在整个语言使用社团中逐渐规约化(conventionalization),也可以促使其迅速传播,从而引起

语音、词汇、句法、语义等方面的变化。总之,频率效应是贯穿整个语言历史变化和发展。对语言的历史研究和考查最终都应该落脚在“频率”上来。桂诗春(2004)在《以概率为基础的语言研究》一文中,从语言事实和语言理论的角度对语言学中的理性主义及其內省式研究方法提出了质疑,介绍了作为相反潮流的概率语言学、语料库语言学等最新发展。认知语言学日益重视用“基于用法的模式”来研究语言事实背后的理据性,而语料库里检索到的大量语料事实能客观地呈现出语言结构的“用法模式”,因此语料库是认知语言学和历史认知语言学相关研究的有力工具,越来越多的历史认知语言学家重视语料库在其研究中的作用、促进两者的融合、开启认历史知语言学研究的新视野。

三、COHA 扩展历史认知语言学研究新视野

美语历史语料库(Corpus of Historical American English,简称COHA)于2010年晚期由美国的Mark Davies教授建成并发布在国际互联网上供语言研究者免费使用。COHA是一个具有4亿词每隔10年为一个时刻度,时间跨度为1810s~2000s近200年的篇型平衡(genre-balanced)的历史语料库,它包括了小说、杂志、报纸和非小说四大类篇型,并且做到词类标注详细和准确,保持了每年每种篇型的平衡性和均等性,能尽量反映出近200年以来美语的词汇、形态、句法和语义的历史发展和变化情况(Mark Davies 2012:122)。COHA的建成和发布,为HCL的研究提供了一个功能强大的研究平台,进一步推动了HCL和语料库的渗透与融合,极大地拓展了HCL研究的新视野。

由于COHA对其中的语料进行了详细的词类标注(tagging POS)、独特的相对数据架构设计(architecture of relational databases)和多功能的检索界面(interface),COHA能迅速完成对近200年以来美语中的词汇、形态、句法、语义、语用和语篇等方面的研究。Davies (2012:125)指出对COHA

进行词类标注的工具是 CLAWS tagger,这一标注工具是 BNC,COCA 和其它主流语料库的标注工具,其准确性相当高。Davies (2009:164)指出他自己设计的这几大语料库(COHA, COCA, BYU-BNC 和 Time Magazine Corpus)采用的都是关系数据库,认为采用这一数据库结构的好处是检索速度更快和使语料库库容够更大,对 4 亿词的 COHA 进行任何复杂检索都是在 1-3 秒完成;同时关系数据库还可以使语料库中存在“分子”结构(Davies 2009:166),它能让其它附加特征集(additional feature sets)很容易整合到 COHA 这种语料库中去,进一步拓展其检索功能的多样性。COHA 准确的词类标注和关系数据库架构保证了

其网页检索界面功能多样性和友好性。下面分别举例说明 COHA 在对美语中的词汇、形态、句法、语义和社会文化等方面历史研究方面的独特功能,以展示其在拓展 HCL 研究新视野的优势。

(一) COHA 对词汇历史变化的研究

COHA 的架构方式和检索界面,最基本的功能是能检索到任何一个词或短语由 1810s 至 2000s 近 200 年每隔 10 年的频率变化情况。例如,在 COHA 检索界面的 WORD(S)处输入单词“research”,同时显示方式(DISPLAY)选择图表(CHART),就可以得到单词“research”的历史频率变化情况的图标显示,如图 1。

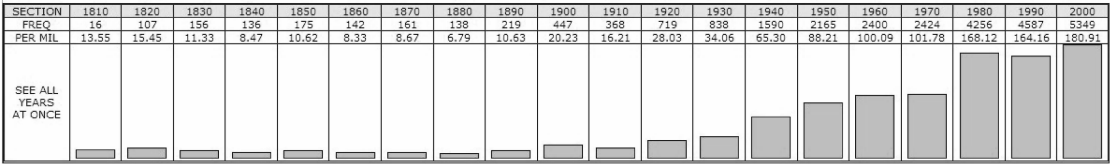


图 1 research 在 1810s—2000s 中每隔 10 年的历史频率变化情况

从图 1,我们可以一目了然地观察到单词 re-search 在 1810s ~ 1910s 频率波动变化不大,而 1910s ~ 2000s 频率变化经历了迅速上升的过程,说明该词的使用日益频繁。

让我们再来看看短语[be] going to 在 COHA

中的历史变化情况:在 COHA 检索界面的 WORD(S)处输入短语“[be] going to”,同时显示方式选择图表,得到“[be] going to”的历史频率变化情况图标显示,如图 2。

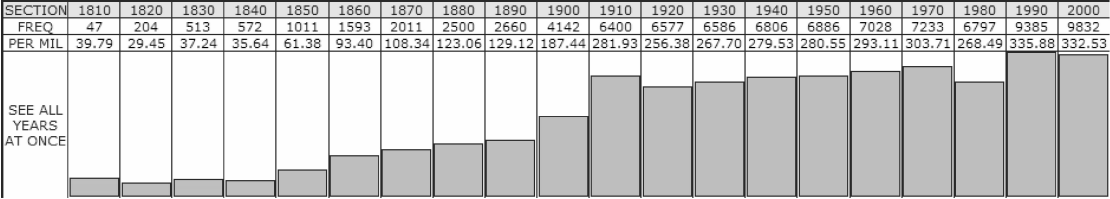


图 2 [be] going to 在 1810s—2000s 中每隔 10 年的历史频率变化情况

图 2 表明动词短语[be] going to 的使用频率经历了由少到多逐渐上升的历史过程,于 1990s 其使用频率达到最高峰。

由于 COHA 独特的架构设计和数据存取方式存储了每一个相匹配被检索字符串的每隔 10 年的频数,这使其能显示所特定词类词或短语在一个特定时间段与另一个特定时间段使用频率高低的排序变化情况,这样我们就很容易跟踪到特定时间段特定词类的词或短语哪些词或短语在这

个时间段出现频率最高或最高低,同时进行相关词或短语在另一特定时间段的比较。例如,我们可以比较 1880s ~ 1910s 期间出现的高频形容词和 1970s ~ 2000s 其间出现的高频有什么变化。只要在 COHA 的检索入口 WORD(S)处输入[j *],在 SECTIONS1 选定时间 1810s ~ 1910s,在 SECTIONS2 处选定时间 1970s ~ 2000s,同时显示方式选列表(LIST),然后运行搜索引擎,就得到结果如表 1 所示。

表 1 1880s ~ 1910s 使用的高频率形容词与 1970s ~ 2000s 使用的高频率形容词比较^①

A	1880s ~ 1910s	Token 1	Token 2	PM 1	PM 2	Ration
1	well	7545	0	88.02	0	8,802.42
2	especial	1063	57	12.4	0.53	23.2
3	good-by	1515	86	17.67	0.81	21.92
4	practicable	933	92	10.88	0.86	12.62
5	disagreeable	1540	162	17.97	1.52	11.83
6	homeword	947	104	11.05	0.98	11.33
7	decided	933	108	10.88	1.01	10.75
8	marvelous	1049	135	12.24	1.27	9.67
9	queer	3 299	498	38.49	4.67	8.24
10	utmost	2 500	432	29.17	4.05	7.2

B	1970s ~ 2000s	Token 2	Token 1	PM 2	PM 1	Ration
1	soviet	13090	43	122.75	0.5	244.69
2	Communist	3 658	20	34.3	0.23	147.01
3	longtime	1 349	8	12.65	0.09	135.54
4	longterm	2 834	20	26.58	0.23	113.9
5	ongoing	1 224	9	11.48	0.1	109.31
6	environmental	4 274	34	40.08	0.4	101.04
7	Palestinian	1 681	14	15.76	0.16	96.51
8	regulatory	1 167	10	10.94	0.12	93.8
9	nuclear	7 025	63	65.88	0.73	89.63
10	minimal	1 085	11	10.17	0.13	79.28

表 1 显示,1880s ~ 1910s 使用的高频率形容词与 1970s ~ 2000s 使用的高频形容是不同的:表 1A 表明 1880s ~ 1910s 使用最多的形容词为 well, especial, practicable 这些都是极常用的形容;表 1B 中的形容词 Soviet, Palestinian, environmental, nuclear 是对这一时间段的国际社会背景、社会思潮和社会文化的一种的反映:苏联剧变、巴勒斯坦问题、日益引起关注的环境问题、核能问题。通过

① 表 1 中的数据说明如下:例如表 1B 中的第 6 个形容词 environmental,它在 1970s ~ 2000s 期间(Token2)共出现 4274 次,而在 1880s ~ 1910s(Token 1)仅出现 34 次。PM1 和 PM2 分别是 Token1 和 Token2 对应词的每百万词的标准化频率(Normalized frequency per Million)。表中的 Ration 是 PM1 和 PM2 的比值,在计算 Ration 时,如果 token 2 出现的次数为 0,即 PM2 为 0,要对其进行平滑(smooth)处理,以避免分母为 0。Environmental 的 ration 为 101.04 表明它在 1970s ~ 2000s 时间段出现的频率是其在 1880s ~ 1910s 时间段出现频率的 101.04 倍,表 1A 和 B 是以 Ration 的大小进行排序的,这些数据都是系统自动计算生成的。后面的类似表格数据含义与此相同。

检索 COHA 中特定时间段的特定词类也能很好地从语言视角窥视当时社会背景、社会文化或社会思潮,据此能更好认识和解释某些词和短语的历史演化过程的社会背景。

(二) COHA 对词的形态历史变化研究

COHA 也能提供关于词的形态 (morphological) 方面的历史演变情况。我们在 COHA 的检索界面的入口 WORD(S) 处输入 [* cracy], 显示方式处选列表, SECTION1 处的时间选 1810s ~ 2000s, 运行搜索引擎得到结果例如表 2。它就是我们 COHA 中检索过去 200 年以来以 -cracy 结尾的前 10 个词的每隔 10 年一个时间刻度的历史演变情况。由表 2 我们可以看到有些词的使用频率在历史发展中是在增加的: 如 democracy, 在 1810s ~ 1930s; bureaucracy 在 1830s ~ 1990s; 而有些词的使用频率是在减少的: 如 aristocracy 在 1850s ~ 1990s 由最初的 317 次减少到 1990s 的 53 次, 差不多减少了 83.3%; 还可以看到有些词语

比如 technocracy 是随着时代的进步而产生的, 随着社会对某些观念 (科技) 的重视而使用频率逐渐增加。从 COHA 获取的这些关于词的形态方面的信息, 为我们深入窥视、了解和认识美国的社会和文化方面的变化有着重要的意义。

COHA 除了可以观察词的形态历史频率变化情况外, 还可以比较特定时间段词的形态使用情况的变化, 如表 3。观察美语中词的形态方面的这些变化情况可以更好的理解美国文化和社会的总体变化情况。由表 3 中 B 表可以看出: meritocracy, marketocracy, technocracy 这些词只在 1970s ~ 2000s 期间有分布, 是这阶段美国“精英管理的社会”“市场经济”“技术主导”的社会、文化和主流意识形态的反映; 而表 3A 表中: plantocracy, plutocracy, aristocracy 这些词语频繁使用在 1870s ~ 1910s 期间, 很少使用在 1970s ~ 2000s 期间, 集中反映了美国这一时期的“大庄园主阶级”“富豪统治”和“贵族”社会背景和文化状况。

表 2 以 -cracy 结尾的词历史频率变化情况^①

	-cracy word	All	1810	1830	1850	1870	1890	1910	1930	1950	1970	1990
1	democracy	15072	10	332	493	453	643	1098	1280	1046	619	1112
2	aristocracy	3802	10	232	317	235	294	160	186	106	93	53
3	bureaucracy	1772		6	16	19	15	41	90	107	305	256
4	autocracy	452		1	8	16	20	127	42	12	17	11
5	theocracy	205	3	9	24	6	9	10	3	7	9	8
6	plutocracy	200			1	15	35	13	13	9	4	3
7	meritocracy	46								2	6	15
8	technocracy	43						1	14	4	7	4
9	pro-democracy	37							1		1	17
10	gerontocracy	17						1			6	5

(All 中的数据由于是统计 1810s ~ 2000s 近 200 年出现的所有总频率, 而表中显示的是每隔 20 年的频率, 所以后面的频率之和是不等于每个词的总频率的, 因为还有如 1820s, 1840s 这些年份没有显示但是计算在总数里面去了。)

① 表 2 中, 为了节约版面空间, 这里选取 1810s ~ 2000s 中每隔 20 年为一个时间刻度来展现这前 10 个词的历史频率变化情况, 显示的也只是原频率数 (raw frequency)。在 COHA 中的实际检索结果是按 10 年为一个时间刻度显示, 同时不同频率频率而用不同的颜色突显出来, 频率数据可以用原始频数显示也可用标准化 (normalized) 了的频率显示, 检索的结果也不只限这 10 个词, COHA 能检索到所有的以 -cracy 结尾的词并全部显示出来 (我们检索到的 -cracy 共有 192 个)。

(三) COHA 对句法历史变化的研究

由于 COHA 中的语料进行了按屈折变化形式进行归类标记(lemmatized)和词类标注,我们可以用它对美语句法变化(syntactic change)进行深入历史研究。例如,我们可以利用 COHA 观察分裂不定式(split infinitive)和非分裂不定式(non-split infinitive)在 1810s ~ 2000s 近 200 年使用频率的历史变化情况,分裂不定式是指“to + ADV - ly + do”,非分裂不定式是指“to + do + ADV - ly”;表 4 只列出了按照检索句法“to * ly. [r *]

[v *]”(如,to seriously take)和“to [v *] * ly. [r *]”(如,to take seriously)输入到 COHA 中的检索入口 WORD(S)处,同时分类和限定(SORTING AND LIMITS)处设定为 1,在 COHA 中检索到的结果在 1940s ~ 2000s 年的各自使用频率变化情况,尽管在 COHA 的检索结果界面有 1810s ~ 1930s 数据,但本表的篇幅不允许将其能完整显示出来,我们只选取了 1840s ~ 2000s 这段时间出现的频数变化结果。

表 3 1870s ~ 1910s 年以-cracy 结尾的词与 1960s ~ 2000s 年以-cracy 结尾词的比较

A	1870s ~ 1910s	Token 1	Token 2	PM 1	PM 2	Ration
1	plantocracy	12	0	0.122	0	11.51
2	plutocracy	116	24	1.11	0.18	6.05
3	autocracy	211	66	2.02	0.51	4.00
4	aristocracy	1145	553	10.98	4.23	2.59
5	theocracy	56	46	0.54	0.35	1.52
6	mocracy	6	6	0.06	0.05	1.25
7	democracy	3328	4560	31.91	34.91	0.91
8	bureaucracy	116	1125	1.11	9.38	0.12

B	1970s ~ 2000s	Token 2	Token 1	PM 2	PM 1	Ration
1	meritocracy	44	0	0.34	0.00	33.69
2	Pro-democracy	36	0	0.28	0.00	27.56
3	technocracy	24	1	0.18	0.01	19.16
4	gerontocracy	15	1	0.11	0.01	11.98
5	bureaucracy	1225	116	9.38	1.11	8.43
6	slavocracy	5	1	0.04	0.01	3.99
7	kleptocracy	5	0	0.04	0.00	3.83
8	marketocarey	5	0	0.04	0.00	3.83

表 4 “to + ADV - ly + do”和“to + do + ADV - ly”在 1940s ~ 2000s 年使用频率变化情况

句法结构	1940	1950	1960	1970	1980	1990	2000
To [v *] * ly. [r *]	3 991	4 071	4 025	3 835	3 901	3 884	3 729
To * ly. [r *][v *]	160	237	371	595	772	1 478	2 058
To * ly. [r *][v *] 占%	4.72	5.82	9.22	15.25	19.79	38.05	55.19

从表 4,我们可以看到分裂不定式结构“to + ADV - ly + do”在 940s ~ 2000s 期间的使用频率

是在大幅增加的,从最初的 4.72% 到 2000s 年 55.19%,增加 11.7 倍;因此,相应的非分裂不定式结构“to + do + ADV - ly”使用频率自然是在大幅减少。在以前的规定语法 (prescriptive grammar) 中,认为分裂式不定式是属于不符合语法表达规范的,是错误的表达。而基于 COHA 检索到的语料结果来看,分裂不定式被日益频繁地使用在美语各类文体的表达中,并有取代非分裂式表达正统地位的趋势;显然规定语法中关于分裂不定式的这一条说明该过时了,因为其与美语日常真实使用分裂不定式之事实严重不符。

表 5 get-passive 和 be-passive 在 COHA 中使用频率历史变化情况

句法形式	1820	1840	1860	1880	1900	1920	1940	1960	1980	2000
be-passive	73 174	155 292	150 060	176 133	189 344	211 826	193 277	174 384	164 839	151 890
get-passive	98	370	614	912	1 178	1 542	2 366	2 665	2 960	4 937
get-passive%	0.13	0.24	0.41	0.52	0.62	0.73	1.22	1.53	1.80	3.25

(由于篇幅限制,本表是按每隔 20 年一个刻度来显示,在 COHA 中是按 10 年一个刻度来显示的,历史变化过程表现得更精细化)

从表 5 我们可以看到 get-passive 的使用频率随着时间的发展是在逐渐的缓慢增加,由最初的不到 0.13% 发展到 2000s 年的 3.25%,显示这一表达已经在美语中某些文体中大量使用。

(四) COHA 对语义历史变化的研究

通过语料库观察词的语义历史变化趋势的传统方法通过是检索出该词的所有形符 (token) 或随机抽样 (randomized) 形符,然后研究其用法模式以考察其语意变化,耗时费力 (Davies, 2012: 142)。但是如果用架构合理的语料库,就能简明和迅度实现对词语语义历史变化情况进行观察。语料库语言学中的一个核心观念是“我们可以通过一个词的伴随词知其语义”(Firth, 1957: 179)。如果我们检索找到一个词的所有搭配 (collocates) 的历史变化情况,这些搭配就能表明其语义的历史变化情况。比如在 COHA 中,我们检索与 queer 一词搭配使用词的形容词历史变化情况——在 COHA 的检索入口 WORD(S) 处输入单词 queer, 在 COLLOCATES 处输入 [j *], 设定前后 4 个字符距, SECTION1 时间选 1870s ~ 1910s, SECTION2

我们还可以得用 COHA 进行更复杂句法结构历史研究。比如说,我们可以很容易利用 COHA 的相关检索规则,观察 be-passive 和 get-passive 这两类被动语态的使用频率在 1810s ~ 2000s 年历史变化情况:在 COHA 的检索入口 WORD(S) 处输入 “[be] [vvn]” 就可以检索到 be + V-ed 形式的被动语态,即 be-passive;在检索入口处的 WORD(S) 处输入 “[get] [vvn]” 就可以检索到 get + V-ed 形式的所有被动结构,即 get-passive, 显式方式选里面的“图表”显示,检索结果如表 5 所示。

时间选 1960s ~ 2000s, 然后运行引擎。我们发现,在 1870s ~ 1910s 期间与 queer 搭配的形容词主要有 little, young, awful, old, 再观察这些词的所在索引行 (concordance lines), 发现 queer 主要的语义是“strange (奇怪的)”;而在 1960 s ~ 2000 s 期间与 queer 搭配的形容词主要有 cultural, straight, gay, 再观察这些词所在索引行,发现 queer 主要语义是“gay (同性恋)”的含义,其语义发生了明显的变化。

Davies (2012: 143) 指出直接比较一个词在两个特定时间段的搭配词能够为其语义和用法的历史演变提供更清晰的证据。例如,我们检索 web 这个词在 COHA 中的 1870s ~ 1910s 期间与之搭配的形容和 web 在 COHA 中的 1970s ~ 2000s 期间与之搭配的形容词,结果如表 6 所示。

从表 6A 我们可以看到:在 1870s ~ 1910s 与 web 搭配最多最密切的是 bright, shinning, golden 等,结合考察这些词所在索引行,可知这一时期 web 的主要语义是“网状之物”;而在 1970s ~ 2000s 与 web 搭配最多的最密切的是 wide, availa-

ble, personal, corporate 等,这些搭配词在 1870s ~ 1910s 几乎没有出现过,结合考察这些词所在的索引行,可知这一时期 web 的浮现(emergent)出了新语义“网络(因特网)”。

除了上面介绍通过搭配来考察词的语义历史演变外,COHA 还提供了具有 30000 多词条的同义词集合(synsets)子库,通过这一工具能考察到词的整个语义历史演变情况。在 COHA 的检索

入口 WORD(S)处输入“[= WORD]”,就可以看到与之相配的同义词每隔 10 年的频率分布变化情况;比如输入“[= beautiful]”,就可以找到与 beautiful 同义的所有形容词,并按 10 年一个刻度显示出来,同时可显示这些同义词的原始频数和标准的频率(每百万词出现的次数),同时还用不同的颜色来标示频率的高低,使检索者一目了然,如图 3。

表 6 1870s ~ 1910s 与 web 搭配的形容词和 1970s ~ 2000s 与 web 搭配的形容词比较

A	1870s ~ 1910s	Token 1	Token 2	PM 1	PM 2	Ration
1	bright	5	0	0.05	0.00	4.79
2	shining	5	0	0.05	0.00	4.79
3	golden	11	4	0.11	0.03	3.44
4	thin	5	2	0.05	0.02	3.13
5	long	5	2	0.05	0.02	3.13
6	human	7	4	0.07	0.03	2.19
7	fine	11	14	0.11	0.11	0.98
8	great	10	14	0.10	0.11	0.89

B	1970s ~ 2000s	Token2	Token 1	PM 2	PM 1	Ration
1	wide	147	2	1.13	0.02	58.68
2	available	21	1	0.16	0.01	16.77
3	personal	18	0	0.14	0.00	13.78
4	free	18	0	0.14	0.00	13.78
5	corporate	15	0	0.11	0.00	11.48
6	popular	15	0	0.11	0.00	11.48
7	official	14	0	0.11	0.00	10.72
8	complex	26	2	0.20	0.02	10.38

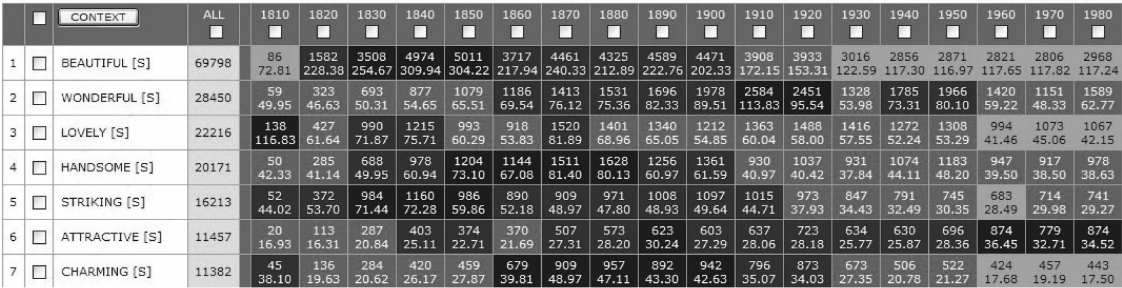


图 3 在 COHA 中输入“[= beautiful]”检索到的 beautiful 同义词截图(部分)

图 3 可让 COHA 的检索者,看到与“beautiful”同义的整个同义语历史变化情况——这些词中,一些词的使用频率随着时间的发展在特定时间增加了,而一些词的使用频率随着时间的发展

减少了。从图 3 我们可以看到:随着时间的发展 beautiful 在 1820s ~ 1980s 年使用频率总体上是呈逐渐减少的趋势,由最初 1820s 的频率每百万词 228.38 次减少到 1980s 的 117.24 次。而作为 beautiful 的同义词 attractive 确是随着时间的发展,其使用频率呈逐渐增加的趋势,由 1810s 的 16.93 次每百万词增加到 1900s 年的 27.29 次,再增加到 1980 年的 34.52 次,可见,美语中人们越来越更喜欢用 attractive 这个 beautiful 的同义词。图 3 上的数据信息,对观察不同的词是怎样为各自的“语义空间竞争”的过程非常有用 (Davies, 2012:143)。

(五) COHA 对社会、文化和历史变化的历史研究

COHA 能提供让我们深入观察词语语义历史演变情况的相关检索。如果把 COHA 这一检索特征扩展到超越纯语言为核心的检索 (linguistically orientated retrieves), 那以就可以通过检索相关词语语义的变化情况来窥视、了解和认识美国的社会、文化和历史变迁情况。比如,我们可以在 COHA 的检索入口 WORD(S) 处输入 sex 一词,然后在 COLLOCATES 输入 [j *], 把 SECTION1 的时间限定在 1810s ~ 1870s, SECTION2 的时间限定在 1940s ~ 2000s; 这样就能检索到这两个时间段与 sex 搭配的形容词历史变化情况,如表 7。

表 7 1810s ~ 1870s 期间与 sex 搭配的形容和 1840s ~ 2000s 期间与 sex 搭配的形容比较

A	1810s ~ 1870s	Token 1	Token 2	PM 1	PM 2	Ration
1	softer	32	1	0.36	0.01	63.81
2	sterner	26	1	0.29	0.01	51.85
3	loverly	12	1	0.13	0.01	23.93
4	gentler	47	4	0.52	0.02	23.43
5	nobler	6	1	0.07	0.01	11.96
6	dear	6	1	0.07	0.01	11.96
7	fond	5	1	0.06	0.01	9.97
8	helpless	5	1	0.06	0.01	9.97
B	1960s ~ 2000s	Token 2	Token 1	PM 2	PM 1	Ration
1	oral	114	0	0.87	0.00	87.28
2	implied	62	0	0.47	0.00	47.47
3	premarital	58	0	0.44	0.00	44.40
4	casual	47	0	0.36	0.00	35.98
5	sexual	44	0	0.34	0.00	33.69
6	explicit	35	0	0.27	0.00	26.80
7	safe	77	2	0.59	0.02	26.53
8	unprotected	34	1	0.26	0.01	23.43

从表 7 可见:在 1810s ~ 1870s 与 sex 搭配最多最常见的形容是: softer, sterner, nobler, gentle, 这些强调“道德”“保守”和“神圣”语义的词语,这反映出当时美国社会的主流意识对“sex”(性)的态度是保守、严格受道德约束和神圣的。而到 1960s ~ 2000s 时期,特别上世纪 60 年代,“性解

放”在美国风行一时,美国青年男女大都喜欢同居而不结婚,并不断变换同居伙伴。性解放:在性行为上完全抛弃传统道德观念约束的主张和实践。这从表 7B 部分检索到的与 sex 搭配的形容主要有:oral(口交的), premarital(婚前的), casual(随意的), unprotected(无保护的), 这些与 sex 搭

配的形容词是对美国这一“性解放”时代社会中的国民对性的态度在语言上的反映,通过这些形容我们可以更深入窥视和认识美国的“性解放”这一社会背景中对性的态度和相关文化。可见,我们可以利用 COHA 进行与此类似的对相关词语的检索来了解和认识美国社会、文化和历史方面的历史变化,就这方面而言,COHA 这大大地拓展了 HCL 的研究范围和提高了获取相关研究语料的检索速度。

四、结束语

近 20 多年来,200 多年前开创的历史语言学和现代认知语言学逐渐融合诞生了历史认知语言学。历史认知语言学吸收了历史语言学和认知语言学的优势,不但分别推动这两个语言学分支学科的发展,而且使历史认知语言成为目前研究的一个热点。认知语言学日益重视用“基于用法的模式”来研究语言事实背后的理据性,因此认知语言学就有与语料库语言学结合的天然优势,历史认知语言学自然可以而且需要用语料库里的相关语料来观察相关语言现象的历史变化过程和追溯历史变化事实背后的理据性。COHA 的建成为历史认知语言学研究免费提供了 1810s ~ 2000s 长达 200 多年,库容达 4 亿词的相关历史语料检索,COHA 具有独特的语料库架构模式和数据存储模式,极好的篇型平衡性,准确的词类标注,这些保证了 COHA 能实现对美语从其词汇、形态、句法、语义到文化等方面的历史研究。在文中 3.1 ~ 3.5 部分,我们以具体的实例分别详细说明了 COHA 在对美语词汇、形态、句法、语义到文化等

方面的作用。从这些实例使我们看到 COHA 极大地拓展了历史认知语言学研究的新视野和为历史认知语言学研究提供了强有力的工具保证。

[参考文献]

- [1] Davies, M. The 385 + million word Corpus of Contemporary American English (1990 – 2008 +) [J]. *International Journal of Corpus Linguistics* 14: 2, 2008: 159-190.
- [2] Davies, M. Expanding horizons in historical linguistics with the 400-million word Corpus of.
- [3] Historical American English [J]. *Corpora*, 2012, 7 (2):121-157.
- [4] Firth, J. R. *Papers in Linguistics 1934-1951* [C]. Oxford: Oxford University Press, 1957.
- [5] Taylor, J. The Mental Corpus: how Language is represented in the Mind [M]. Oxford: OUP, 2012.
- [6] Winters, M., H. Tissari & K. Allan (eds.). *Historical Cognitive Linguistics* [C]. Berlin: Mouton de Gruyter, 2010.
- [7] 桂诗春. 以概率为基础的语言研究[J]. 外语教学与研究, 2004(1):3-8.
- [8] 徐通锵. 历史语言学[M]. 北京:北京商务印书馆, 1991.
- [9] 王寅. 认知语言学和历史语言学的最新发展——历史认知语言学[J]. 外语教学与研究. 2012(6):925-934.
- [10] 唐健禾. 翻译行为的限制性与创造性的文化透视[J]. 四川理工学报(社会科学版), 2012(3):60-64.

(责任编辑:朱德东)

Tools for Expanding the Horizon of Historical Cognitive Linguistics

——Application of Corpus of Historical American English

LI Hua-yong

(School of Foreign Language, Sichuan University of Arts and Science, Sichuan Dazhou 635000, China)

Abstract: In the recent over 20 years, the Historical Cognitive Linguistics is gradually developing by blending Historical Linguistics, which has more than 200 years history, with Modern Cognitive Linguistics. Cognitive Linguistics has the tradition using the “usage-based model” to study the motivation in the linguistic facts, which makes Historical Cognitive Linguistics have the natural advantage of blending with Corpus Linguistic to observe related language phenomena and the historical change process of linguistic structures, and to search for the law of diachronic change of linguistic facts. This paper specifies the functions of COHA in studying American words, morphology, syntax, semantic and culture, etc, which illustrates the COHA expanding the horizon of Historical Cognitive Linguistics, and COHA also provides a powerful tool for Historical Cognitive Linguistic research.

Key words: COHA; Historical Cognitive Linguistics; Cognitive Linguistics; Corpus