2

USING LARGE ONLINE CORPORA TO EXAMINE LEXICAL, SEMANTIC, AND CULTURAL VARIATION IN DIFFERENT DIALECTS AND TIME PERIODS

Mark Davies

Introduction

Three of the most interesting types of linguistic variation are variation over time, variation between dialects, and variation at the level of the individual speaker. Variation by individual speakers is typically related to demographic variables such as ethnicity, gender, and socioeconomic factors (income, occupation, education, etc.). Many of the other chapters in this book focus on these demographic variables and how they relate to language variation.

In this chapter, rather than discussing demographic variables and variation, I will focus on corpora that allow us to look at interesting issues related to culture and society, either in terms of change over time or variation between dialects. I will be focusing particularly on several of the BYU family of corpora (http://corpus.byu.edu) that are the most useful for looking at cultural change and variation—COCA (the Corpus of Contemporary American English), COHA (the Corpus of Historical American English), the Google Books corpus, GloWbE (Global Web-based English) (see also Chapters 3 and 4 of this volume), and the NOW corpus (News on the Web).

In many cases, these corpora are 50–100 times as large as comparable corpora of dialectal and historical English, and as a result, these corpora enable us to look at many types of variation that would not be possible to study otherwise (see Davies, 2015). But size is not everything. There are certainly larger corpora of English than those just mentioned, such as the Sketch Engine corpora and the "Corpora from the Web" project (COW). Without the right architecture and interface, however, massive corpora are often just large, undifferentiated "blobs" of data, which provide little insight into variation and change. There are relatively

few corpora that are both large enough and contain have the correct architecture and interface to enable meaningful insight into language variation and change for a wide range of linguistic phenomena.

In this chapter, I will focus primarily on differences in lexis, meaning, and discourse between dialects over time. Many examples of syntactic and morphological phenomena can be found in publications like Davies (2009, 2012, 2014, 2015), as well as in the help files that accompany the corpora online (http://corpus.byu.edu).

Examining Lexical Variation Between Varieties of English (Including Cultural Differences)

With the free two billion-word GloWbE corpus (http://corpus.byu.edu/glowbe; see Davies & Fuchs, 2015), researchers simply enter a word or phrase into the corpus, and they can then see the frequency in each of 20 different countries. For example, consider *banjaxed* 'messed up, screwed up' (Irish English), *hand phone* 'cell phone/mobile phone' (Malaysia and Singapore), and *cope up* (South Asia and other "Outer Circle" varieties) in Figures 2.1–3.

To provide just a few more examples, simple searches in GloWbE show that all the following are more common in [British] than American English: *fortnight, trousers, rained off, on holiday, at university, [be] different to,* and *rather more ADJ*. Examples from other countries include: [Irish] *jackeen*, culchie*, childer,* and *soft day*; [Australia] *bikkies, thongs,* and *rockmelon**; [Malaysia] (+Singapore) *rakyat, makan, [take] ADJ food,* and *lah!;* and [Jamaica] *ackee, bammy, guinep,* and *callaloo*. Users can also see comparisons across groups of countries, for example, [South Asia] *out of station, eve teas*, be elder to,* and *keep in view*; or even [non-"core" countries]:, *thrice, godown, same to the,* and *[discuss] about.*

Such comparisons across varieties of English may seem overly trivial or simplistic, but they would actually be quite difficult or even impossible in all other corpora of World Englishes. The only other "large" corpus of World Englishes is the International Corpus of English (ICE), which is about 15–20 million words in size (new sub-corpora are currently under development). Although that might seem like a large size, it is actually too small for most comparisons of lexis. ICE is only about 1/100th the size of GloWbE, and so—on average—there would be about 1/100th the number of tokens from ICE as there are from GloWbE. In other words, it is likely that none of the examples shown in Figures 2.1–3 would have been possible with ICE. At 1/100th the number of tokens, there would be less than one token total for *banjax** (cf. 77 tokens in GloWbE), less than one token for *hand phone* (cf. 90 in GloWbE), and only about 5–6 tokens of *cope up* (cf. 510 in GloWbE)

Because GloWbE is robust enough to look at lexical variation, we can of course use it to look at words that relate to cultural differences between the dialects as well. For example, Figures 2.4–6 shows the frequency of Buddh* (*Buddhist, Buddhism, Buddha,* etc.), *Quran,* and feminis* [*feminism, feminist(s)*] in the 20 different countries in GloWbE.

SECTION	ALL	US	CA	GB	IE	AU	NZ	IN	LK	PK	BD	SG	MY	PH	HK	ZA	NG	GH	KE	TZ	JM
FREQ	77	1	0	16	59	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
PER MIL	0.04	0.00	0.00	0.04	0.58	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

FIGURE 2.1 GloWbE: *banjax**: Irish English

SECTION	ALL	US	CA	GB	IE	AU	NZ	IN	LK	PK	BD	SG	MY	PH	HK	ZA	NG	GH	KE	TZ	JM
FREQ	90	0	0	4	0	1	0	6	4	2	2	10	54	2	0	1	1	1	0	1	1
PER MIL	0.05	0.00	0.00	0.01	0.00	0.01	0.00	0.06	0.09	0.04	0.05	0.23	1.30	0.05	0.00	0.02	0.02	0.03	0.00	0.03	0.03

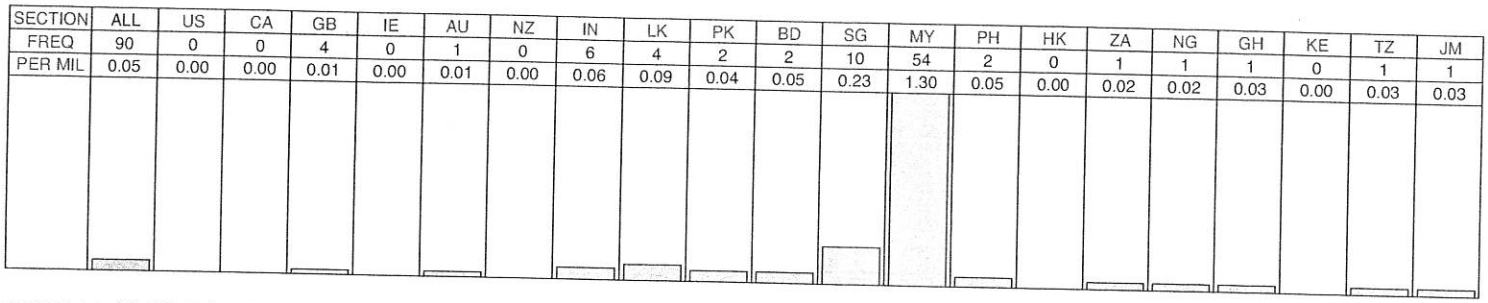
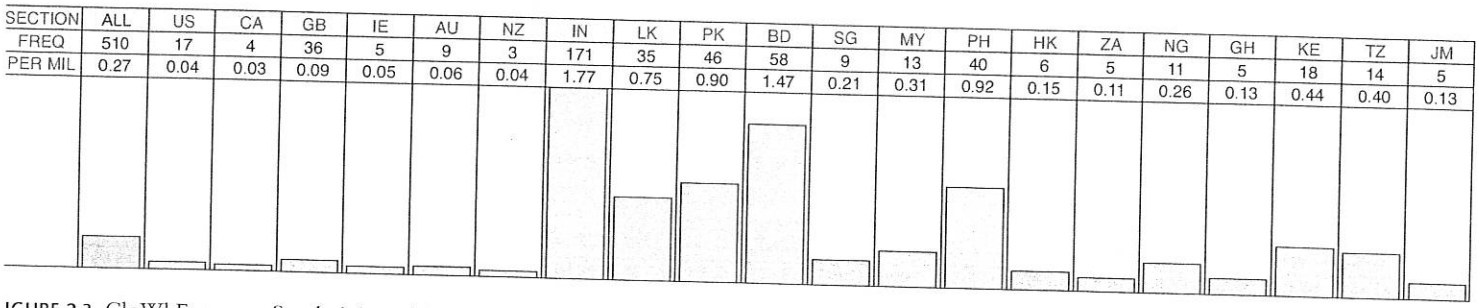


FIGURE 2.2 GloWbE: *hand phone*: Malaysia and Singapore

SECTION	ALL	US	CA	GB	IE	AU	NZ	IN	LK	PK	BD	SG	MY	PH	HK	ZA	NG	GH	KE	TZ	JM
FREQ	510	17	4	36	5	9	3	171	35	46	58	9	13	40	6	5	11	5	18	14	5
PER MIL	0.27	0.04	0.03	0.09	0.05	0.06	0.04	1.77	0.75	0.90	1.47	0.21	0.31	0.92	0.15	0.11	0.26	0.13	0.44	0.40	0.13



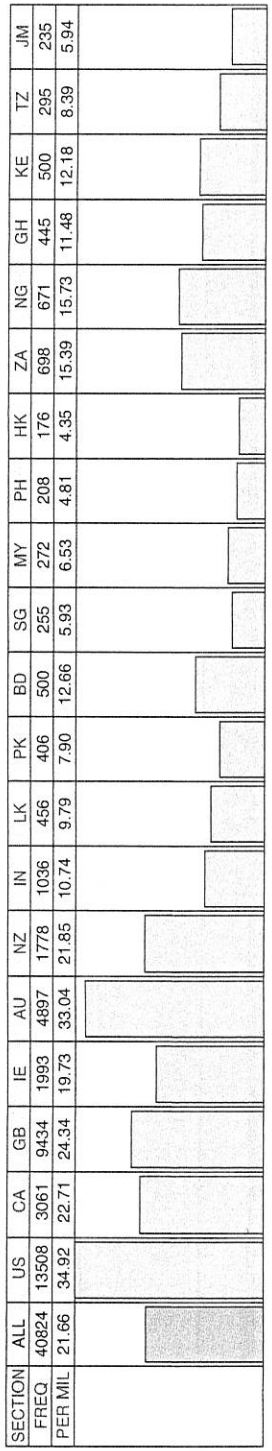
IGURE 2.3 GloWbE: *cope up*: South Asia and Outer Circle

SECTION	ALL	US	CA	GB	IE	AU	NZ	IN	LK	PK	BD	SG	MY	PH	HK	ZA	NG	GH	KE	TZ	JM
FREQ	120303	7870	1395	6193	1326	3287	1581	9001	55912	1840	52.44	5134	7762	1280	10485	390	268	275	326	393	341
PER MIL	63.84	20.35	10.35	15.98	13.12	22.18	19.42	93.34	1,200.26	35.82	132.81	119.47	186.39	29.60	259.21	8.60	6.28	7.09	7.94	11.18	8.62

FIGURE 2.4 GloWbE: *Buddh**

SECTION	ALL	US	CA	GB	IE	AU	NZ	IN	LK	PK	BD	SG	MY	PH	HK	ZA	NG	GH	KE	TZ	JM
FREQ	33332	2067	303	2903	59	472	75	1754	501	19116	2088	132	1664	31	52	238	914	362	110	474	17
PER MIL	17.69	5.34	2.25	7.49	0.58	3.18	0.92	18.19	10.75	372.14	52.88	3.07	39.96	0.72	1.29	5.25	21.43	9.34	2.68	13.48	0.43

FIGURE 2.5 GloWbE: *Quran*



SECTION	ALL	US	CA	GB	IE	AU	NZ	IN	LK	PK	BD	SG	MY	PH	HK	ZA	NG	GH	KE	TZ	JM
FREQ	40824	13508	3061	9434	1993	4897	1778	1036	456	406	500	255	272	208	176	698	671	445	500	295	235
PER MIL	21.66	34.92	22.71	24.34	19.73	33.04	21.85	10.74	9.79	7.90	12.66	5.93	6.53	4.81	4.35	15.39	15.73	11.48	12.18	8.39	5.94

FIGURE 2.6 GloWbE: *feminis**

Perhaps not surprisingly, *Buddh** is the most common in Sri Lanka (the most Buddhist country in the corpus), *Quran* is the most frequent in Pakistan, and *feminis** is the most common in the "Inner Circle" countries (US, Canada, Great Britain, Ireland, Australia, and New Zealand). While these are perhaps trivial examples, they do show that GloWbE can be used to look at cultural differences between varieties of English. Consider also that many of these culturally oriented searches might be much more difficult or impossible with a corpus like ICE, which would have only about 1/100th the number of tokens.

In addition to size, one of the advantages of GloWbE (in terms of lexical variation) is the ease with which one can see variation for hundreds or thousands of words, all at the same time. For example, with a simple search, one can see the frequency of all words ending in **ism* in each of the 20 countries, as is shown in Figure 2.7.

A quick look at this data shows that *autism*, *feminism*, and *atheism* are more frequent in Inner Circle countries and that discussions of *terrorism*, for example, are more frequent in the South Asian countries of Pakistan and Sri Lanka.

Perhaps more useful, however, are searches where we have the corpus find those words that are more common in one variety than in another. For example, the data in Figure 2.8 show those words ending in **ies* that are more common in Australian English (left) than in the other Inner Circle varieties (right).

Note that not all the results are examples of the Australian **ies* "diminutive" (e.g., *swannies*, *telemovies*, *mesenteries*), but the majority are: *vinnies* (wine stores), *firies* (fire fighters), *furphies* (rumors), *dunnies* (toilets), *eskies* (coolers), *bikies* (bikers), *tradies* (tradesmen), *pollies* (politicians), *schoolies* (breaks from school), *streeties* (homeless people), and *tanties* (tantrums).

The ability to quickly and easily carry out such "mass comparisons" of lexis can also be used to examine cultural differences between varieties of English. For example, Figure 2.9 shows those **ism* words that are more common in Great Britain (left) or in South Asia (right).

In Great Britain, people are writing about *Euroscepticism*, *Labourism*, *presenteeism*, *nimbyism* (*nimby* = 'not in my backyard'), *monetarism*, *Thatcherism*, and *Blairism*—with most of these being political in nature. In South Asia, on the other hand, most of the **ism* words are related to religion, such as *Qadianism*, *castism*, *Talibanism*, *Vaisnaism*, *Shivaism*, and *Shiaism*. Thus, there seems to be a real difference in terms of what people in these two regions are writing about on the internet.

Examining Lexical Variation over Time (Including Cultural Change)

In the same way that GloWbE can be used (almost uniquely) to look at lexical and cultural differences between varieties of English, the 400 million word Corpus of Historical American English (COHA; http://corpus.byu.edu/coha; see Davies, 2012) can be used to look at lexical (and cultural) change.

	CONTEXT	ALL	US	CA	GB	IE	AU	NZ	IN	LK	PK	BD	SG	MY	PH	HK	ZA	NG	GH	KE	TZ	JM
1	TOURISM	66201	2859	3177	7376	3290	4234	3871	3563	3716	922	1703	2135	2448	2312	2945	2635	1092	2838	3746	6370	4969
2	CRITICISM	62734	14465	3644	15808	3164	4983	2298	3017	1839	2200	1148	811	1022	813	1123	1450	1314	1036	967	720	912
3	MECHANISM	44334	8850	2552	8021	2291	3576	1793	3274	1736	1105	1178	885	920	705	1634	1063	758	830	1344	1067	752
4	TERRORISM	42198	8783	1909	6845	732	2101	881	2940	5427	5529	1570	317	471	317	417	395	1277	461	1023	544	259
5	JOURNALISM	41459	10280	2878	10441	1591	3953	1090	1693	997	743	927	522	332	613	647	840	784	908	895	863	462
6	CAPITALISM	37319	9466	2266	10261	1942	2835	1549	1356	681	602	871	461	218	367	875	850	516	394	370	817	622
7	RACISM	36539	11535	1894	8545	1859	2987	1051	797	1082	579	332	501	831	198	327	1185	586	675	505	367	703
8	BUDDHISM	21793	1829	309	1434	351	757	390	1790	9061	324	828	845	1200	314	1954	74	66	68	58	86	55
9	AUTISM	20293	7240	1508	5277	1585	2207	260	715	73	56	273	73	98	159	104	65	38	76	36	70	380
10	SOCIALISM	19837	6423	792	4291	1020	1732	733	746	292	284	535	191	114	223	534	412	173	202	156	690	294
11	OPTIMISM	15125	2950	1249	3764	765	988	532	677	265	375	324	346	244	327	303	295	363	379	483	242	254
12	NATIONALISM	14389	1521	880	3052	1020	851	268	1032	1473	886	772	141	184	285	310	368	347	277	212	229	281
13	COMMUNISM	14201	4465	630	3284	632	1249	401	504	190	317	377	202	227	234	330	395	159	118	132	207	148
14	BAPTISM	12367	2696	1506	1314	965	917	814	193	178	82	793	129	88	695	252	284	223	571	166	300	201
15	FEMINISM	12215	4157	886	2928	556	1491	482	248	124	125	95	61	92	77	52	152	255	138	165	84	47
16	ATHEISM	10705	4644	354	1763	404	1565	552	238	70	241	101	37	74	139	66	131	114	16	94	51	51

FIGURE 2.7 GloWbE: **ism* words by country

SEC 1 (Australia): 148,208,169 WORDS

	WORD/PHRASE	TOKENS 1	TOKENS 2	PM 1	PM 2	RATIO
1	SWANNIES	82	1	0.6	0.0	604.0
2	VINNIES	33	2	0.2	0.0	121.5
3	TELEMOVIES	15	1	0.1	0.0	110.5
4	MESENTERIES	13	1	0.1	0.0	95.7
5	KNACKERIES	11	1	0.1	0.0	81.0
6	LOGIES	64	6	0.4	0.0	78.6
7	FIRIES	32	3	0.2	0.0	78.6
8	DUNNIES	21	2	0.1	0.0	77.3
9	FURPHIES	21	2	0.1	0.0	77.3
10	ESKIES	19	2	0.1	0.0	70.0
11	BIKIES	113	13	0.8	0.0	64.0
12	YABBIES	42	5	0.3	0.0	61.9
13	TRADIES	198	34	1.3	0.0	42.9
14	POLLIES	599	104	4.0	0.1	42.4
15	SCHOOLIES	231	41	1.6	0.0	41.5
16	STREETIES	60	0	0.4	0.0	40.5

SEC 2 (United States, Canada, Grea...): 1,091,609,517 WORDS

	WORD/PHRASE	TOKENS 2	TOKENS 1	PM 2	PM 1	RATIO
1	SUBDIRECTORIES	158	1	0.1	0.0	21.5
2	HYPERINTENSITIES	147	1	0.1	0.0	20.0
3	BOBBIES	215	2	0.2	0.0	14.6
4	JEFFERIES	98	1	0.1	0.0	13.3
5	MASTECTOMIES	81	1	0.1	0.0	11.0
6	CRYBABIES	159	2	0.1	0.0	10.8
7	SORORITIES	139	2	0.1	0.0	9.4
8	MAMMIES	68	1	0.1	0.0	9.2
9	HOMOLOGIES	68	1	0.1	0.0	9.2
10	BARONIES	266	4	0.2	0.0	9.0
11	EQUALITIES	568	9	0.5	0.1	8.6
12	TORIES	15127	258	13.9	1.7	8.0
13	SQUADDIES	116	2	0.1	0.0	7.9
14	TONALITIES	55	1	0.1	0.0	7.5
15	STO-RIES	54	1	0.0	0.0	7.3
16	BENNIES	107	2	0.1	0.0	7.3

FIGURE 2.8 GloWbE: **ies* words in Australian and other Inner Circle dialects

SEC 1 (Great Britain): 387,615,074 WORDS

	WORD/PHRASE	TOKENS 1	TOKENS 2	PM 1	PM 2	RATIO
1	EUROSCEPTICISM	137	1	0.4	0.0	82.7
2	LABOURISM	124	1	0.3	0.0	74.8
3	ANTI-FASCISM	86	1	0.2	0.0	51.9
4	PRESENTEEISM	81	1	0.2	0.0	48.9
5	THROMBOEMBOLISM	432	6	1.1	0.0	43.4
6	NIMBYISM	72	1	0.2	0.0	43.4
7	ISOMERISM	86	2	0.2	0.0	25.9
8	MONETARISM	81	2	0.2	0.0	24.4
9	BLAIRISM	93	0	0.2	0.0	24.0
10	THATCHERISM	382	10	1.0	0.0	23.0
11	LOCALISM	765	22	2.0	0.1	21.0
12	ANGLICANISM	259	8	0.7	0.0	19.5
13	BONAPARTISM	63	2	0.2	0.0	19.0
14	MANAGERIALISM	119	4	0.3	0.0	17.9
15	NATURISM	102	4	0.3	0.0	15.4

SEC 2 (India, Sri Lanka, Pakistan,...): 233,866,709 WORDS

	WORD/PHRASE	TOKENS 2	TOKENS 1	PM 2	PM 1	RATIO
1	QADIANISM	100	1	0.4	0.0	165.7
2	NAXALISM	145	2	0.6	0.0	120.2
3	ISMAILISM	68	1	0.3	0.0	112.7
4	SHAIVISM	56	1	0.2	0.0	92.8
5	BRAHMANISM	114	3	0.5	0.0	63.0
6	ETERNALISM	72	2	0.3	0.0	59.7
7	CASTEISM	249	7	1.1	0.0	59.0
8	MARXISM-LENINISM-MAOISM	127	0	0.5	0.0	54.3
9	BUDHISM	107	4	0.5	0.0	44.3
10	JISM	404	16	1.7	0.0	41.8
11	VAISHNAVISM	83	0	0.4	0.0	35.5
12	JAINISM	551	32	2.4	0.1	28.5
13	SAIVISM	65	0	0.3	0.0	27.8
14	COMMUNALISM	733	44	3.1	0.1	27.6
15	SHI'ISM	414	25	1.8	0.1	27.4

FIGURE 2.9 GloWbE: **ism* words in Great Britain and South Asia

At the most basic level, COHA enables us to see the frequency of any word or phrase in each of the 20 decades in the corpus from the 1810s to the 2000s. This is of course much more useful than resources like the Oxford English Dictionary, which can show the first attestation of a word but not the usage frequency over time. Examples of these frequency charts are shown in Figures 2.10–12, where we see words that have been decreasing in frequency since the 1800s (*grievid*), a word that peaked several decades ago and has since decreased (*swell* as an adjective), and words that have been increasing over time (*frustrating* as an adjective).

These are, of course, just a handful of examples of lexical change. Other examples might include: (decrease over time): *bosom, bestow**, *beauteous, fellow, sublime, lad, many a time*, and *of no little*; (an increase and then decrease): *anyhow, mustn't, naughty, as though to, don't know as* (=that), *far-out*, and *lousy*; (increase over time): *a lot of, guys, unleash, sexual, calm down, screw up, freak out*, and *mommy*. And of course users can search for more complex phrases, such as the following (all of which have decreased over time): *so* ADJ *as to* VERB (e.g., *so good as to show me*), *be but* (*they are but the last examples*), *have quite* VERB-ed (*until she had quite finished*), NOUN *be that of* (*her dress was that of a beggar*), or *a most* ADJ NOUN (*a most helpful child*).

The GloWbE data enabled us to see cultural differences between different dialects of English. Similarly, the COHA data can be used to look at cultural changes in the US during the last 200 years. Consider the data from Figures 2.13–15: *steamship* (increase through the early 1900s—when steamships were more common—and then decrease since then), *communis** (peaks in the 1950s), and *teenager* (which may be related to a changing view of adolescents since World War II).

The important point is that until recently, basic searches such as these were simply impossible. In the same way that GloWbE is about 100 times as large as the ICE corpus (the next-largest corpus of English dialects), at 400 million words COHA is about 100 times as large as the Brown family of corpora, which was the largest historical corpus of English until COHA was released in 2010. As with the GloWbE data, we can look at the totals for Figures 2.13–15 and see that with only 1/100th the number of tokens, very few of these searches would be possible. With a 4–5 million word corpus (like the Brown family), there would only be 1–3 tokens per decade—far too small to say much about such lexical change.

Parenthetically, it should be noted that, in addition to showing the frequency by decade for any search, users can also see the frequency in each individual year from 1810–2009. For example, Figure 2.16 shows that the word *reds* is the most frequent in the 1950s. Users can then click on the 1950s heading to see the frequency in each year of the 1950s. In this case, as Figure 2.16 shows, they would see that the frequency was highest in 1953, which again corresponds to changes in American history and society (the McCarthy hearings in the U.S. Senate).

As we have seen, a 4–5 million word corpus is often too small to look at lexical change for anything but the most frequent words. There have been some attempts to use such small corpora to look at changes in lexes, including Leech and Fallon (1992), Baker (2009, 2010, 2011), Baron, Rayson, and Archer (2009), Hofland and

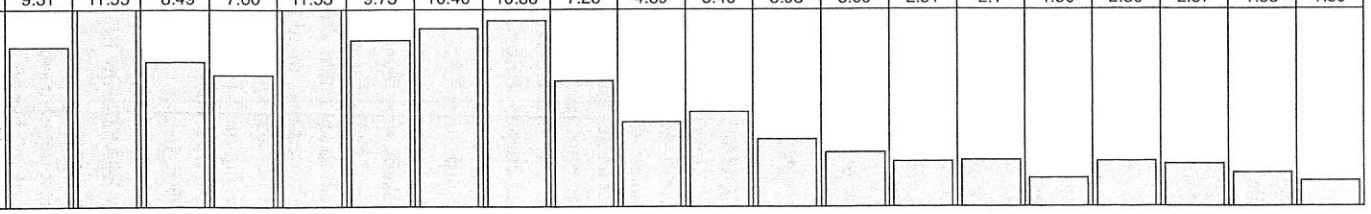
SECTION	1810	1820	1830	1840	1850	1860	1870	1880	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000
FREQ	11	80	117	122	190	166	193	221	150	108	124	102	76	61	64	36	61	60	51	41
PER MIL	9.31	11.55	8.49	7.60	11.53	9.73	10.40	10.88	7.28	4.89	5.46	3.98	3.09	2.51	2.1	1.50	2.56	2.37	1.83	1.39
SEE ALL YEARS AT ONCE																				

FIGURE 2.10 COHA: *grieved*

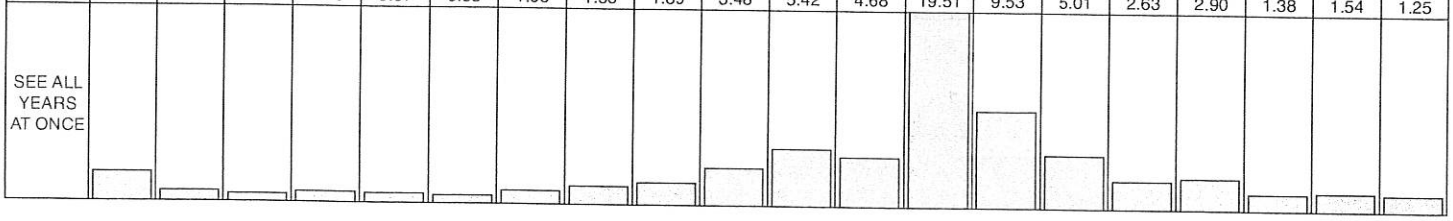
SECTION	1810	1820	1830	1840	1850	1860	1870	1880	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000
FREQ	3	5	8	12	11	9	20	31	39	77	123	120	480	232	123	63	69	35	43	37
PER MIL	2.54	0.72	0.58	0.75	0.67	0.53	1.08	1.53	1.89	3.48	5.42	4.68	19.51	9.53	5.01	2.63	2.90	1.38	1.54	1.25
SEE ALL YEARS AT ONCE																				

FIGURE 2.11 COHA: *swell* (ADJ)

SECTION	1810	1820	1830	1840	1850	1860	1870	1880	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000
FREQ	0	1	0	2	1	1	4	2	1	1	2	3	1	14	75	103	98	155	184	193
PER MIL	0.00	0.14	0.00	0.12	0.06	0.06	0.22	0.10	0.05	0.05	0.09	0.12	0.04	0.57	3.06	4.30	4.12	6.12	6.59	6.53
SEE ALL YEARS AT ONCE																				

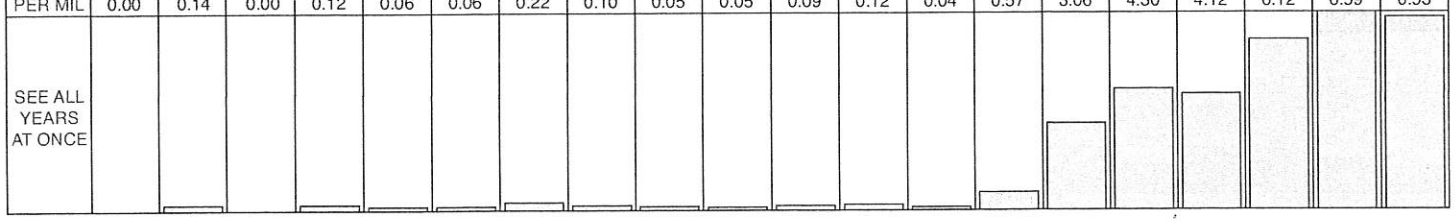


FIGURE 2.12 COHA: *frustrating* (ADJ)

SECTION	1810	1820	1830	1840	1850	1860	1870	1880	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000
FREQ	0	1	0	6	17	56	29	86	191	236	246	219	221	137	96	45	55	22	18	15
PER MIL	0.00	0.00	0.00	0.37	1.03	3.28	1.56	4.23	9.27	10.68	10.84	8.54	8.98	5.63	3.91	1.88	2.31	0.87	0.64	0.51
SEE ALL YEARS AT ONCE																				

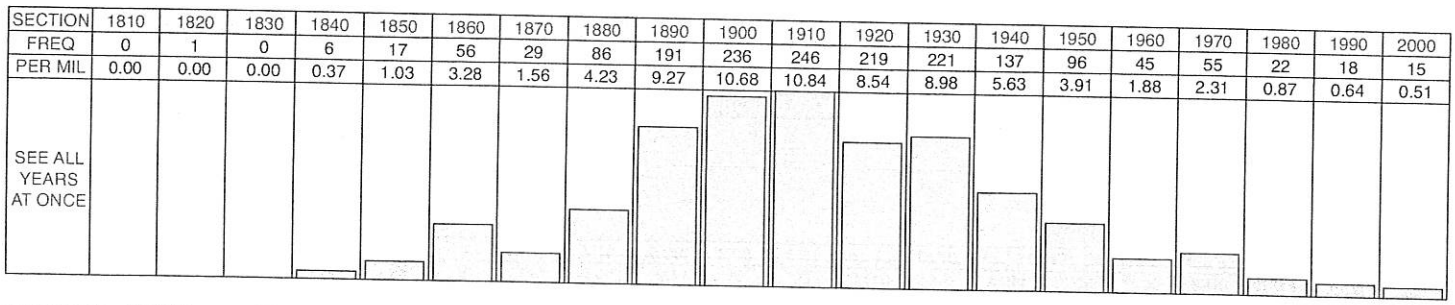


FIGURE 2.13 COHA: *steamship*

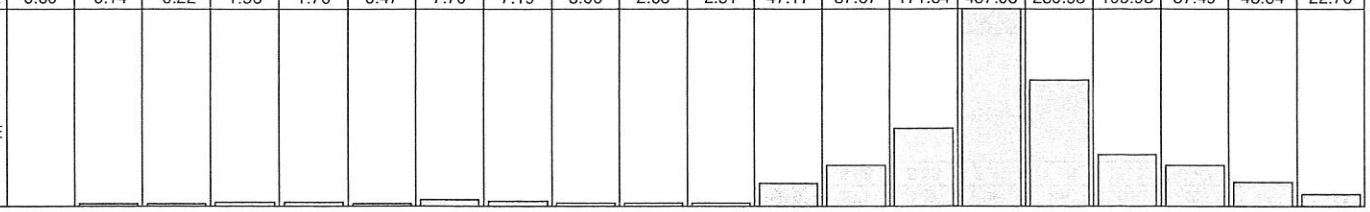
SECTION	1810	1820	1830	1840	1850	1860	1870	1880	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000
FREQ	0	1	3	25	29	8	143	146	63	46	57	1210	2157	4184	10728	6736	2618	2215	1359	673
PER MIL	0.00	0.14	0.22	1.56	1.76	0.47	7.70	7.19	3.06	2.08	2.51	47.17	87.67	171.84	437.08	280.93	109.93	87.49	48.64	22.76
SEE ALL YEARS AT ONCE																				

FIGURE 2.14 COHA: *communis**

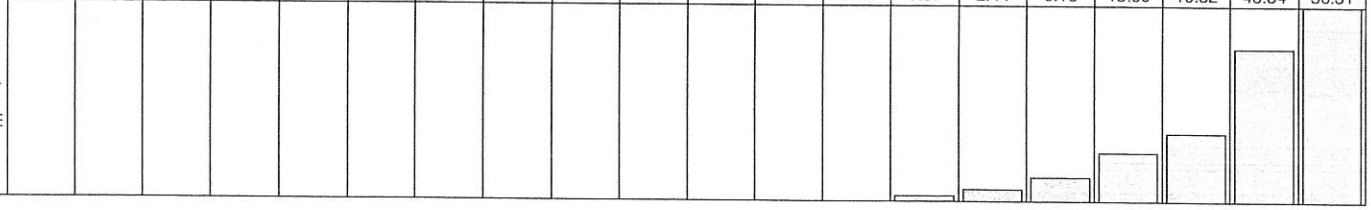
SECTION	1810	1820	1830	1840	1850	1860	1870	1880	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000
FREQ	0	0	0	0	0	0	0	0	0	0	0	0	0	14	60	147	324	489	1225	1665
PER MIL	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.57	2.44	6.13	13.60	19.32	43.84	56.31
SEE ALL YEARS AT ONCE																				

FIGURE 2.15 COHA: *teenager*

SECTION	1810	1820	1830	1840	1850	1860	1870	1880	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000
FREQ	0	0	1	4	3	7	26	14	36	54	65	179	202	191	567	315	125	337	166	204
PER MIL	0.00	0.00	0.07	0.25	0.18	0.41	1.40	0.69	1.75	2.44	2.86	6.98	8.21	7.84	23.10	13.14	5.25	13.31	5.94	6.90
SEE ALL YEARS AT ONCE																				

1950	1951	1952	1953	1954	1955	1956	1957	1958	1959
88	92	61	123	64	22	28	15	25	49
33.26	**36.74**	**24.60**	**52.12**	**27.92**	**8.83**	**10.79**	**6.49**	**9.36**	**20.15**

FIGURE 2.16 COHA: *Reds* (by decade and year)

Johansson (1982), Oakes and Farrow (2007), and Sigley and Holmes (2002). But as one of the most active researchers in this field (Baker, 2011, p. 70) notes:

> Leech and Fallon (1992) point out that the corpora in the Brown family contain only about 50,000 word types in total, which is relatively small for lexical research, and that the majority of words will be too infrequent to give reliable guidance on British and American uses of language.
>
> For that reason, this study focuses only on frequent words in the corpora. It was stipulated that for a word to be of interest to this study, it would need to occur at least 1,000 times when its frequencies in all four corpora were added together. Three hundred eighty words met this criteria, but a number of high frequency words (e.g., *class, miss, black, true,* and *English*) were excluded because they missed the cutoff.

In other words, with a corpus of just a few million words, we are limited to just looking at a handful of very high frequency words, which often makes them of little use when we are looking at lexis to see cases of societal or cultural shifts, as in Figures 2.10–12 and 2.13–15.

In addition to the issue of size, another problem with some "historical" corpora is that there is not enough "data granularity." For example, with the Brown family of data, there is only data from two different years—1961 and 1991. Current projects are extending the family of corpora back to 1931 and even 1901, but in any case, there is still only data from every 30 years. This means that any changes that take place in between these years are essentially "invisible," and in terms of lexical change, this is often too long of a gap. For example, consider the frequency for *groovy* in COHA (Figure 2.17). (Note that in COHA, we have robust data from not only each decade, but also from each year. For example, there is data for 75,377,000 words for the 30 years from 1955 to 1985—more than 2,400,000 words *each year* for this 30 year period.)

Imagine that we had a corpus that had (like the Brown family of corpora) only two data points. Rather than the years 1961 and 1991 in Brown and FROWN, imagine that our corpus had data from just 1955 and 1985. In this case, it would appear (based on the COHA data from the 1950s and the 1980s) that *groovy* is on the increase. While it has increased slightly in these 30 years (0.12 in the 1950s and 0.36 in the 1980s), we would miss entirely the steep increase in the 1960s and the steep decrease from the 1960s/1970s to the 1980s. Lexical frequency often changes too quickly to be sampled just every 30 or so years, but that is unfortunately the only option with these very small corpora.

As a second example, consider the case of *normalcy* in Figure 2.18.

This word was famously "rescued" from obscurity by President Warren G. Harding in 1920, who (according to purists) mistakenly used it instead of the more "correct" *normality*. The word caught on with a public tired of World War I and other foreign involvements, and Harding went on to win the election. But imagine that we only had two small corpora from 1901 and 1931 (as with the

SECTION	1810	1820	1830	1840	1850	1860	1870	1880	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000
FREQ	0	0	0	0	0	0	0	0	0	0	0	0	0	2	3	44	42	9	14	21
PER MIL	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0	0.00	0.08	0.12	1.84	1.76	0.36	0.50	0.71
SEE ALL YEARS AT ONCE																				

1960	1961	1962	1963	1964	1965	1966	1967	1968	1969
2	0	1	0	0	0	0	10	23	8
0.77	**0.00**	**0.35**	**0.00**	**0.99**	**0.00**	**0.00**	**3.66**	**10.38**	**3.33**

FIGURE 2.17 COHA: *groovy*

SECTION	1810	1820	1830	1840	1850	1860	1870	1880	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000
FREQ	0	0	0	0	0	0	0	0	0	0	0	24	27	29	25	26	27	28	34	45
PER MIL	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.94	1.10	1.19	1.02	1.08	1.13	1.11	1.22	1.52
SEE ALL YEARS AT ONCE																				

FIGURE 2.18 COHA: *normalcy*

planned extensions in the Brown family of corpora). There would obviously be a large increase in frequency between 1901 and 1931, but there would be no way to know whether that predated Harding, whether his campaign caused the increase in usage, or whether it was after his time. Corpora with texts that are spaced decades apart may be adequate for looking at more gradual grammatical change, but they are much more problematic when it comes to lexical change, which can occur quite suddenly.

Recall that with the GloWbE corpus, we could use one simple search to find all words that are more common in one dialect than another. With COHA, we can do something similar—we can find all words that have increased or decreased in frequency between different periods. In other words, we do not need to decide ahead of time what words we will look for—the corpus will find them for us. For example, we can find the frequency of all words ending in **ism* over time (Figure 2.19).

In the corpus, we see a decrease in the use of the words *patriotism* and *despotism*, an increase and then decrease in *communism* and *nationalism*, and an overall increase in *capitalism, optimism*, and *journalism*.

As with GloWbE, we can also compare across specific sections of the corpus. For example, we can find those **ism* words (Figure 2.20) that were more common in the 1860s–1910s (left) or in the 1970s–2000s (right):

People nowadays rarely talk about *pauperism, fetichism, Romanism, bimetallism*, or *heathenism* (at least using those terms), but there has been an increase over time in people talking about *racism, tourism, activism, consumerism*, and *sexism*. Not all these are relevant to changes in American culture and society, but many are.

Using Collocates to Examine Semantic and Cultural Differences Between Dialects

In addition to lexical differences (what topics people are talking about in different varieties of English or over time), we might also want to know *what* they are saying about different topics (see also Berber Sardinha's collocation study using GloWbE in Chapter 4). In other words, the frequency of the words *family, religion*, or *women* might not vary much between dialects or change much over time, but *what* people are saying about these topics may vary quite a bit by country or show interesting changes over time.

But what corpus-based data could provide evidence of differences in meaning and usage between two or more dialects? For example, how would we know that in American English *cupboard* is restricted primarily to storing items in a kitchen or pantry, whereas in British English it can also be used for a storage place in other rooms in the house (cf. American *closet*)? Or, how would we know that *scheme* is typically used in a negative sense in American English but that this is not the case in other varieties of English?

One approach would be to look at concordance lines for the word or phrase in different dialects and to see whether the surrounding context might indicate

CONTEXT	ALL	1810	1820	1830	1840	1850	1860	1870	1880	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000
CRITICISM	13508	25	156	243	341	370	408	689	682	706	1016	1212	1123	860	771	922	974	975	835	659	541
PATRIOTISM	4923	25	147	439	359	332	406	259	306	357	329	481	290	221	179	114	116	155	170	125	113
COMMUNISM	4790				6	14	4	57	101	26	15	34	168	441	496	1450	940	292	279	320	147
MECHANISM	4538		20	71	141	96	106	119	98	151	285	276	381	358	291	376	329	338	267	492	343
SOCIALISM	3540				17	55	10	148	92	181	213	446	269	397	331	279	295	311	304	136	56
ORGANISM	3423		3	2	69	36	82	174	229	179	321	289	374	273	256	343	191	214	136	120	132
JOURNALISM	2631		3	1	57	17	34	87	60	99	108	150	196	131	201	177	182	207	292	283	346
CAPITALISM	2509								2	7	11	65	135	270	247	217	208	259	453	436	199
OPTIMISM	2505		1	12	4		6	22	45	49	104	164	236	225	188	236	218	225	275	194	301
DESPOTISM	2255	23	84	204	293	388	287	199	117	119	90	86	102	54	47	46	19	27	41	22	7
BAPTISM	2106		49	40	217	531	260	131	162	110	100	61	38	44	49	47	54	49	42	43	79
HEROISM	2031	18	61	71	115	169	163	113	179	131	108	170	111	84	90	83	61	69	95	60	80
REALISM	2012		5		1	13	24	49	122	120	122	112	198	147	139	297	165	115	152	157	134
NATIONALISM	1841				1	1	3	4	44	24	15	99	203	172	232	195	264	141	180	167	96
TERRORISM	1817			1	2	3	7	9	2	6	12	19	57	55	50	30	62	221	387	148	746

FIGURE 2.19 COHA: **ism* words by decade

SEC 1 (1860, 1870, 1880, 1890, 190...): 121,332,176 WORDS

	WORD/PHRASE	TOKENS 1	TOKENS 2	PM 1	PM 2	RATIO
1	PAUPERISM	217	1	1.8	0.0	190.7
2	FETICHISM	113	0	0.9	0.0	93.1
3	ROMANISM	62	0	0.5	0.0	51.1
4	BIMETALLISM	62	0	0.5	0.0	51.1
5	HEATHENISM	94	2	0.8	0.0	41.3
6	DEMAGOGISM	41	1	0.3	0.0	36.0
7	PROPAGANDISM	43	0	0.4	0.0	35.4
8	MOHAMMEDANISM	117	3	1.0	0.0	34.3
9	ECCLESIASTICISM	41	0	0.3	0.0	33.8
10	SPIRITISM	38	0	0.3	0.0	31.3
11	INVALIDISM	69	2	0.6	0.0	30.3
12	TRADE-UNIONISM	34	1	0.3	0.0	29.9

SEC 2 (1970, 1980, 1990, 2000): 106,640,094 WORDS

	WORD/PHRASE	TOKENS 2	TOKENS 1	PM 2	PM 1	RATIO
1	RACISM	994	0	9.3	0.0	932.1
2	TOURISM	756	0	7.1	0.0	708.9
3	ACTIVISM	404	1	3.8	0.0	459.7
4	MARXISM	342	2	3.2	0.0	194.6
5	FUNDAMENTALISM	176	0	1.7	0.0	165.0
6	MULTICULTURALISM	156	0	1.5	0.0	146.3
7	COUNTERTERRORISM	134	0	1.3	0.0	125.7
8	DYNAMISM	130	0	1.2	0.0	121.9
9	AUTHORITARIANISM	101	1	0.9	0.0	114.9
10	EXPRESSIONISM	100	1	0.9	0.0	113.8
11	CONSUMERISM	121	0	1.1	0.0	113.5
12	SEXISM	115	0	1.1	0.0	107.8

FIGURE 2.20 COHA: **ism* words, 1860s–1910s vs 1970s–2000s

differences in meaning. For example, Figure 2.21 shows a few of the concordance lines for *cupboard* from the Great Britain section of the GloWbE corpus.

Notice that in these sentences, the *cupboard* is over the *chimney* (#1) or under the *stairs* (#10), that boxes of *photos* (#5) or *stationary* (#7) are stored there, and that it is possible to purchase a *stand-alone cupboard* (#9)—all of which would seem strange in American English.

However, given a large enough corpus, we can use another approach. Rather than looking at all 8,726 tokens of the word *cupboard* in GloWbE, for example, we can simply use the corpus interface to look for all collocates ("nearby words") of *cupboard*. We could then compare the collocates to see which ones occur in one dialect but not another, and which may therefore signal differences in meaning and usage.

For example, Figure 2.22 shows a comparison of the collocates of *cupboard* in 386 million words of American English (left) and 387 million words from British English (right) in GloWbE.

While not all the collocates are relevant, many are. For example, *refrigerator* and *pantry* are more frequent (per million words) in American English, probably because there are more references to *cupboard* in the context of a kitchen. In British English, on the other hand, there are references to *brooms* and *wardrobes*, as well as to *skeletons in the cupboard*, all of which would be used with *closet* in American English.

As another example of semantic differences between dialects, let us consider the collocates of *scheme* in American and British English (Figure 2.23). In American English (left), there are references to *alleged, evil, fraudulent, Ponzi, (get) rich quick*, and *illegal* schemes, whereas in British English (right) the collocates are much more neutral in tone (or even positive: *generous, innovative, competent*, and *qualified*). In corpus linguistic terms, we could say that *scheme* has "negative prosody" in American English (cf. Louw, 1993), whereas this is not the case for British English.

In these three cases, we compared British and American English. This was done for two reasons. First, these are the two varieties with a global reach, and many speakers of other varieties are familiar with them. Second, these are the two largest segments of GloWbE, at about 385 million words each. Such comparisons may still be possible with smaller segments, perhaps even with countries like Tanzania (35 million words), Ghana (39 million words), or Bangladesh (40 million words), which are among the smallest in the corpus. This is especially the case if regional dialects are compared (e.g., Africa = 203 million words or South Asia = 234 million words).

The examples with *cupboard* and *scheme* show that we can use collocates to compare the meaning of a word in different varieties of English. But we can extend this line of reasoning and see that we can also use collocates to find out what is being said about the same topic in different dialects. For example, Figure 2.24 shows the most frequent adjectival collocates of *belief* in South Asia (left) and the six Inner Circle countries (right).

had been deliverd and that the child was hid in the	cupboard	overhead the chimney in the room over the kitchen . That she
on . Priscilla Fisher , another pottery hand , saw the	cupboard	revolve , heard a bang &; on looking saw the deceased fall
all those late nights when he said he was out airing	cupboard	shopping . Gullible old Helen Cupboard . # Method I 've used
! I will also try using this to keep our kitchen	cupboards	shut . I know you can buy baby safety products to do
boxes of photos tucked away in the loft or a forgotten	cupboard	somewhere . But with CEWE PHOTOBOOK memories come back to life
real buzz kill . # Safe storage and trust # Big	cupboards	that can be locked with your own padlock are great . So
burial ground , or is there a part of the stationery	cupboard	that has always been a degree or two colder than the rest
I thought I 'd gather together those products lurking in my	cupboard	that I either would n't repurchase or which just did n't live
? And to illustrate the point she indicates a Victorian display	cupboard	that she has recently bought and painted , and which is now
3 year olds are very interested in what 's in the	cupboard	under the stairs that the funny people with shiny tools are

FIGURE 2.21 GloWbE: concordance lines for *cupboard*

SEC 1 (United States): 386,809,355 WORDS

	WORD/PHRASE	TOKENS 1	TOKENS 2	PM 1	PM 2	RATIO
1	REFRIGERATOR	9	1	0.0	0.0	9.0
2	CLOSETS	12	2	0.0	0.0	6.0
3	ACTIVITY	6	1	0.0	0.0	6.0
4	CLOSET	9	2	0.0	0.0	4.5
5	PANTRY	8	3	0.0	0.0	2.7
6	PLATE	5	2	0.0	0.0	2.5
7	ITEMS	9	6	0.0	0.0	1.5
8	MOTHER	6	4	0.0	0.0	1.5
9	STUFF	6	6	0.0	0.0	1.0
10	WAY	5	5	0.0	0.0	1.0
11	GLASS	6	7	0.0	0.0	0.9
12	YEAR	6	8	0.0	0.0	0.8
13	YEARS	6	10	0.0	0.0	0.6
14	BATHROOM	7	15	0.0	0.0	0.5

SEC 2 (Great Britain): 387,615,074 WORDS

	WORD/PHRASE	TOKENS 2	TOKENS 1	PM 2	PM 1	RATIO
1	AIRING	131	3	0.3	0.0	43.6
2	DAYS	15	1	0.0	0.0	15.0
3	SIDE	14	1	0.0	0.0	14.0
4	STORAGE	41	3	0.1	0.0	13.6
5	BROOM	53	4	0.1	0.0	13.2
6	SKELETONS	39	3	0.1	0.0	13.0
7	DUST	13	1	0.0	0.0	13.0
8	SKELETON	13	1	0.0	0.0	13.0
9	WARDROBES	13	1	0.0	0.0	13.0
10	FUME	25	2	0.1	0.0	12.5
11	STORE	72	7	0.2	0.0	10.3
12	BACK	92	10	0.2	0.0	9.2
13	CEREAL	9	1	0.0	0.0	9.0
14	WARDROBE	9	1	0.0	0.0	9.0

FIGURE 2.22 GloWbE: collocates of *cupboard* in the US and Great Britain

SEC 1 (United States): 386,809,355 WORDS

	WORD/PHRASE	TOKENS 1	TOKENS 2	PM 1	PM 2	RATIO
1	BLOCKING	42	1	0.1	0.0	42.1
2	URI	80	6	0.2	0.0	13.4
3	OFFENSIVE	61	6	0.2	0.0	10.2
4	DEFENSIVE	89	13	0.2	0.0	6.9
5	SOCIALIST	20	3	0.1	0.0	6.7
6	ALLEGED	26	5	0.1	0.0	5.2
7	EVIL	48	10	0.1	0.0	4.8
8	FRAUDULENT	62	18	0.2	0.0	3.5
9	NEFARIOUS	27	9	0.1	0.0	3.0
10	PONZI	617	255	1.6	0.7	2.4
11	FEDERAL	30	13	0.1	0.0	2.3
12	REGULATORY	50	22	0.1	0.1	2.3
13	AMERICAN	22	13	0.1	0.0	1.7
14	ELABORATE	71	43	0.2	0.1	1.7
15	RICH	86	55	0.2	0.1	1.6
16	QUICK	57	38	0.1	0.1	1.5
17	ILLEGAL	53	36	0.1	0.1	1.5

SEC 2 (Great Britain): 387,615,074 WORDS

	WORD/PHRASE	TOKENS 2	TOKENS 1	PM 2	PM 1	RATIO
1	APPROVED	92	1	0.2	0.0	91.8
2	OCCUPATIONAL	88	1	0.2	0.0	87.8
3	MENTORING	53	1	0.1	0.0	52.9
4	FLAT	36	1	0.1	0.0	35.9
5	ELIGIBLE	31	1	0.1	0.0	30.9
6	OVERSEAS	31	1	0.1	0.0	30.9
7	DEFINED	127	5	0.3	0.0	25.3
8	GENEROUS	50	2	0.1	0.0	24.9
9	LABOUR	25	1	0.1	0.0	24.9
10	TAX-AVOIDANCE	25	1	0.1	0.0	24.9
11	SCOTTISH	24	1	0.1	0.0	24.0
12	INNOVATIVE	70	3	0.2	0.0	23.3
13	AUTOMATIC	23	1	0.1	0.0	23.0
14	COMPETENT	23	1	0.1	0.0	23.0
15	QUALIFIED	22	1	0.1	0.0	22.0
16	JOINT	21	1	0.1	0.0	21.0
17	VULNERABLE	21	1	0.1	0.0	21.0

FIGURE 2.23 GloWbE: collocates of *scheme* in the US and Great Britain

SEC 1 (India, Sri Lanka, Pakistan,...): 233,866,709 WORDS

	WORD/PHRASE	TOKENS 1	TOKENS 2	PM 1	PM 2	RATIO
1	CHIEF BELIEF	10	1	0.0	0.0	53.0
2	*HINDU BELIEFS*	47	6	0.2	0.0	41.5
3	SECTARIAN BELIEFS	13	2	0.1	0.0	34.5
4	CORRUPT BELIEFS	12	2	0.1	0.0	31.8
5	AGE-OLD BELIEF	14	5	0.1	0.0	14.8
6	BLIND BELIEFS	11	4	0.0	0.0	14.6
7	POLYTHEISTIC BELIEFS	18	7	0.1	0.0	13.6
8	ESSENTIAL BELIEFS	15	6	0.1	0.0	13.3
9	HINDU BELIEF	42	22	0.2	0.0	10.1
10	WRONG BELIEF	43	25	0.2	0.0	9.1
11	WRONG BELIEFS	51	31	0.2	0.0	8.7
12	ISLAMIC BELIEF	113	79	0.5	0.1	7.6
13	HERETICAL BELIEFS	10	7	0.0	0.0	7.6
14	BUDDHIST BELIEF	27	19	0.1	0.0	7.5
15	*MONOTHEISTIC BELIEF*	15	12	0.1	0.0	6.6
16	ISLAMIC BELIEFS	111	94	0.5	0.1	6.3
17	MUSLIM BELIEF	39	39	0.2	0.0	5.3

SEC 2 (United States, Canada, Grea...): 1,239,817,686 WORDS

	WORD/PHRASE	TOKENS 2	TOKENS 1	PM 2	PM 1	RATIO
1	SILLY BELIEFS	186	1	0.2	0.0	35.1
2	*THEISTIC BELIEF*	77	1	0.1	0.0	14.5
3	CONTRADICTORY BELIEFS	53	1	0.0	0.0	10.0
4	LIBERAL BELIEFS	42	1	0.0	0.0	7.9
5	APPARENT BELIEF	79	2	0.1	0.0	7.5
6	DEEPEST BELIEFS	39	1	0.0	0.0	7.4
7	SILLY BELIEF	34	1	0.0	0.0	6.4
8	SIMPLE BELIEF	58	2	0.0	0.0	5.5
9	POSITIVE BELIEF	58	2	0.0	0.0	5.5
10	CATHOLIC BELIEF	83	3	0.1	0.0	5.2
11	DIFFERING BELIEFS	55	2	0.0	0.0	5.2
12	DEEP BELIEFS	27	1	0.0	0.0	5.1
13	ECONOMIC BELIEFS	27	1	0.0	0.0	5.1
14	CONFIDENT BELIEF	25	1	0.0	0.0	4.7
15	CAUSAL BELIEFS	24	1	0.0	0.0	4.5
16	NON-RELIGIOUS BELIEFS	24	1	0.0	0.0	4.5
17	PRE-EXISTING BELIEFS	24	1	0.0	0.0	4.5

FIGURE 2.24 GloWbE: collocates of *belief* in South Asia and Inner Circle countries

Notice the use of *Hindu*, *Muslim*, *Islamic*, *polytheistic*, *monotheistic*, *sectarian*, and *heretical* in South Asia (all of which are probably related to religion), compared to *liberal*, *deepest*, *positive*, *economic*, *confident*, *causal*, and *non-religious* in the Inner Circle countries (more secular).

Another example of the ability to gain cultural insight from a comparison of collocates are the adjectival collocates of the lemma *marriage* in the Outer Circle countries (left) and the Inner Circle countries (right) in Figure 2.25.

In the Outer Circle countries, there is concern about *inter-caste*, *fixed*, and *forceful* marriages, as well as *permanent* vs *temporary* marriages (perhaps as a husband is forced to look for work outside of his home country). In the Inner Circle countries, on the other hand, people are apparently more concerned with the "hot button" topic of same-sex marriage, with adjectives like *opposite-sex* and *same-sex*, and related words like *anti-gay*, *supporting*, and *preserving* (i.e., traditional heterosexual marriage), as well as *pro-abortion* and *unborn*—apparently referring to "conservatives" and "liberals," in the context of their views on same-sex marriages.

To take one final example, consider the collocates of *wife* (Figure 2.26) in the Outer Circle (left) and the Inner Circle varieties (right).

The Outer Circle countries include countries where polygamy is legal (such as Pakistan and African countries), as well as countries like India and Sri Lanka where the culture is arguably more traditional than in Inner Circle countries like the US, Great Britain, and Australia. Note that in the Outer Circle countries there is reference to *chaste* wife and *obedient* wife (followed further down the list by *good* wife and *virtuous* wife), which would probably sound quite sexist in Inner Circle countries. Due to the existence of polygamy in some of the Outer Circle countries, we also find reference to *existing* wife, *senior* wife, and *legal* wife. We also find *temporary* wife and *permanent* wife, which probably refer to the practice of very temporary "marriages" to a woman for the purpose of sexual relations, followed by an equally quick "divorce."

In summary, it is quite interesting to see how much insight we can gain about cultural and societal practices in a particular country through a very quick and simple search of collocates of a given word such as *belief*, *marriage*, or *wife*. But of course, this assumes that we have a corpus that is large and robust enough to compare these collocates. If we had a corpus that is only 1/100th the size of GloWbE, we might only have 1/100th the number of tokens for a particular collocate, which would make searches like those shown in Figures 2.24–26 impossible. In addition, we need a corpus architecture and interface that is designed to compare data in different sections of the corpus, as we have with GloWbE.

Using Collocates to Examine Semantic and Cultural Change Over Time

In the same way that we can use collocates to look at semantic and cultural differences between contemporary varieties of English, we can also look at changes in collocates to investigate semantic and cultural changes over time. Of course we

SEC 1 (India, Sri Lanka, Pakistan,...): 644,753,594 WORDS

	WORD/PHRASE	TOKENS 1	TOKENS 2	PM 1	PM 2	RATIO
1	INTER-CASTE	91	2	0.1	0.0	87.5
2	FIXED	45	1	0.1	0.0	86.5
3	PHILIPPINE	35	1	0.1	0.0	67.3
4	FORCEFUL	26	1	0.0	0.0	50.0
5	NIGERIAN	51	2	0.1	0.0	49.0
6	CUSTOMARY	359	23	0.6	0.0	30.0
7	FIXED-TIME	164	0	0.3	0.0	25.4
8	HALAL	22	2	0.0	0.0	21.2
9	HINDU	271	26	0.4	0.0	20.0
10	PERMANENT	427	54	0.7	0.0	15.2
11	TEMPORARY	678	103	1.1	0.1	12.7
12	BLISSFUL	72	12	0.1	0.0	11.5
13	IRREPARABLE	23	4	0.0	0.0	11.1
14	ISLAMIC	322	65	0.5	0.1	9.5
15	AFRICAN	116	25	0.2	0.0	8.9

SEC 2 (United States, Canada, Grea...): 1,239,817,686 WORDS

	WORD/PHRASE	TOKENS 2	TOKENS 1	PM 2	PM 1	RATIO
1	IRISH	121	1	0.1	0.0	62.9
2	OPPOSITE-SEX	102	1	0.1	0.0	53.0
3	AUSTRALIAN	136	3	0.1	0.0	23.6
4	PRO	78	2	0.1	0.0	20.3
5	CONSISTENT	30	1	0.0	0.0	15.6
6	SAME-GENDER	79	3	0.1	0.0	13.7
7	IDENTICAL	48	2	0.0	0.0	12.5
8	NARROW	24	1	0.0	0.0	12.5
9	FACTO	47	2	0.0	0.0	12.2
10	PRO-ABORTION	23	1	0.0	0.0	12.0
11	UNBORN	20	1	0.0	0.0	10.4
12	ANTI-GAY	175	9	0.1	0.0	10.1
13	SUPPORTING	109	7	0.1	0.0	8.1
14	PRESERVING	28	2	0.0	0.0	7.3
15	INFERTILE	25	2	0.0	0.0	6.5

FIGURE 2.25 GloWbE: collocates of *marriage* in Outer and Inner Circle countries

SEC 1 (India, Sri Lanka, Pakistan,...): 644,753,594 WORDS

	WORD/PHRASE	TOKENS 1	TOKENS 2	PM 1	PM 2	RATIO
1	EXISTING WIFE	25	1	0.0	0.0	48.1
2	CHASTE WIFE	21	1	0.0	0.0	40.4
3	PAKISTANI WIFE	23	3	0.0	0.0	14.7
4	SENIOR WIFE	21	3	0.0	0.0	13.5
5	TEMPORARY WIFE	27	4	0.0	0.0	13.0
6	OBEDIENT WIVES	23	6	0.0	0.0	7.4
7	PERMANENT WIFE	45	0	0.1	0.0	7.0
8	MUSLIM WIFE	94	26	0.1	0.0	7.0
9	AFRICAN WIFE	20	7	0.0	0.0	5.5
10	DIVORCED WIFE	41	15	0.1	0.0	5.3
11	LEGAL WIFE	72	27	0.1	0.0	5.1

SEC 2 (United States, Canada, Grea...): 1,239,817,686 WORDS

	WORD/PHRASE	TOKENS 2	TOKENS 1	PM 2	PM 1	RATIO
1	PLURAL WIVES	35	1	0.0	0.0	18.2
2	DESERTED WIFE	68	3	0.1	0.0	11.8
3	GLAMOROUS WIFE	20	1	0.0	0.0	10.4
4	MILITARY WIVES	172	11	0.1	0.0	8.1
5	MILITARY WIFE	111	14	0.1	0.0	4.1
6	DESERTED WIVES	22	3	0.0	0.0	3.8
7	PLURAL WIFE	20	3	0.0	0.0	3.5
8	DYING WIFE	31	6	0.0	0.0	2.7
9	ILL WIFE	29	6	0.0	0.0	2.5
10	DISABLED WIFE	23	5	0.0	0.0	2.4
11	MERRY WIVES	50	11	0.0	0.0	2.4

FIGURE 2.26 GloWbE: collocates of *wife* in Outer and Inner Circle countries

could simply look at concordance lines with a given word in different historical periods, such as *gay* used in context in the 1870s (Figure 2.27) and the 2000s (Figure 2.28) in COHA. But there are more than 15,000 tokens of *gay* in COHA, and so if we wanted to examine all the concordance lines, this would potentially be quite time-consuming.

Because COHA is so large, however, we might simply examine the collocates over time. As Figure 2.29 shows, we can see the collocates by decade. This shows quite clearly that in the 1800s the meaning of *gay* was "happy" or "cheerful" as with the collocates *bright, flowers, laugh, colors,* and *spirits*. In the 2000s, however, we find collocates like *lesbian(s), rights,* and *marriage*.

To make this even more clear, we can simply search for the collocates in two competing time periods, such as the 1850s–1910s (left) and the 1970s–2000s (right) in Figure 2.30.

Note the collocates *spirits, heart, voices, attire,* and *song* in the 1850s–1910s, and the collocates *rights, lesbian(s), marriage, bar* (verb), *activists,* and *straight* in the last 30–40 years.

Another example of comparing collocates is Figure 2.31, which shows the adjectival collocates preceding *women* in the 1830s–1890s (left) and the 1960s–2000s (right) and how women were represented and portrayed in the two periods.

Note the emphasis in the 1800s on the "moral" or "vulnerable" qualities of women, with collocates such as *true, unfortunate, helpless, wretched, pure, noblest, devoted, cultivated, refined,* and *abandoned*, or the mention of *strong-minded* or *clever* women, as though this was not the norm (which is again quite sexist, according to current norms). In the late 1900s, on the other hand, the collocates of *women* are somewhat more prosaic (*middle-class, adult, local*), and they also relate to topics that might have been somewhat more "taboo" in the 1800s (e.g., *pregnant, battered, menopausal,* and *divorce*).

In Figure 2.31, we searched for just the exact string "ADJ *women*," but in Figure 2.32 we look for collocates of *women*—up to four words to the left and four to the right (and of course we could search up to ten left and ten right, using the corpus interface).

This time we compare the 1930s–1950s and the 1960s–1980s, two very different historical periods in terms of how women were viewed by society. In the 1930s–1950s, note the emphasis on appearance (e.g., *wear* ["*women's wear*"], *fabrics, hips*) or women entering the workforce in World War II (e.g., *factories, coast, wartime*). In the 1960s–1980s, on the other hand, there are references to the feminist movement and other related social movements (e.g., *liberation, minorities, abortion, AIDS, activists*).

Consider one final example. Figure 2.33 shows adjectival collocates of *religion* in the 1800s (left) and the 1970s–2000s (right).

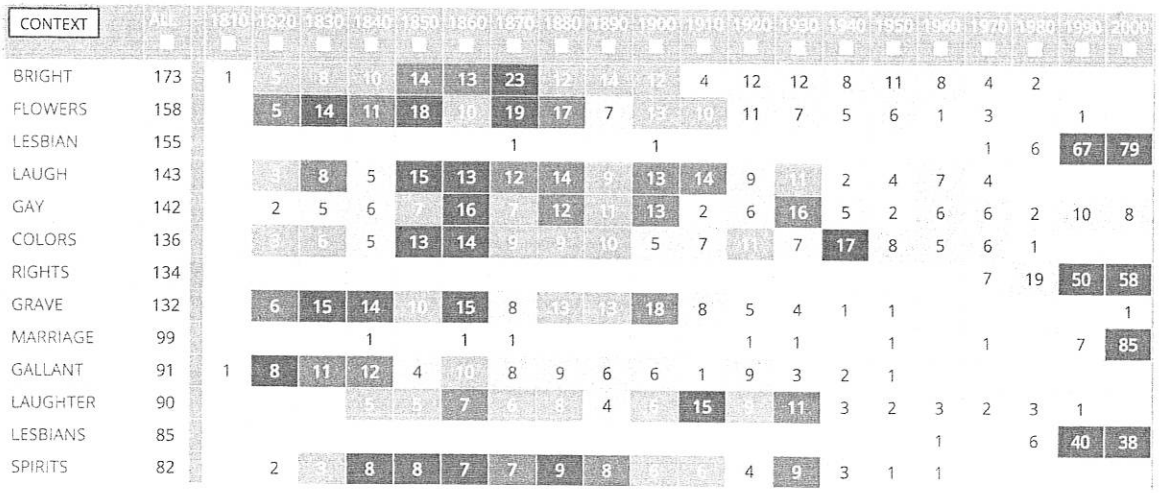
Note that in the 1800s, religion was viewed more as a personal, emotional phenomenon (*beautiful, blessed, sublime, practical*) and that there was also more of an emphasis on the "truth value" of religion (or a particular religion): *absolute, pure, essential,* and *undefiled*. In the 1970s–2000s, on the other hand, religion is more

51	1871	FIC	HoosierSchoolmaster	A	B	C	into a cell , where there was a man with a	gay	red	plume	in	his hat and a strip of red flannel about
52	1875	MAG	Galaxy	A	B	C	women and children on the left . A few have brought	gay	rugs	or	blankets	to kneel on ; but the most kneel humbly
53	1871	FIC	NobleWoman	A	B	C	got it out of a novel ! " Elsie had a	gay	scarf	wound	about	her neck , and began complaining of the warmth
54	1871	MAG	Harpers	A	B	C	uneasily up and down , while others , *veary of the	gay	scene	in	which	they had no share , sought the billiard-rooms
55	1874	FIC	IsabelLeicester	A	B	C	I could have wept , and would gladly have exchanged that	gay	scene	,	for	the quiet of my own room . But this
56	1872	FIC	EdnaBrowningThe	A	B	C	Roy was not present , and how flushed and excited and	gay	she	became	the	moment he appeared , and she raised a warning
57	1873	FIC	WorkAStoryExperience	A	B	C	, where from her mother 's arms she soon regarded the	gay	sight	with	such	sprightly satisfaction that she seemed a little
58	1874	FIC	TerribleSecret	A	B	C	mountains on every hand . The words of the girl 's	gay	song	come	over	the water : " The time I 've lost
59	1871	FIC	NobleWoman	A	B	C	sick . She heard Elsie 's voice ringing out in a	gay	song	,	she	went mechanically on with her dressing , listening to
60	1879	FIC	WiredLove A	A	B	C	, saying nothing , she missed continually the sympathy , the	gay	talk	,	the	companionship that had made the

FIGURE 2.27 COHA: collocates of *gay*, 1870s

36	2005	NF	OurEndangeredValues	A	B	C	predictor of party affiliation -- more important than	gay	marriage	,	homosexuality	, or abortion . // It is encouraging
37	2001	NEWS	AP	A	B	C	If Vermont 's civil union law has helped galvanize opposition to	gay	marriage	,	it	also has inspired many same-sex couples . Amo
38	2005	MAG	Time	A	B	C	, and she 's citing Ohio 's new constitutional ban on	gay	marriage	.	Since	they were never legally married under Ohio
39	2001	NF	SocialPsych	A	B	C	men and masculine women are more likely to be seen as	gay	men	and	lesbians	, respectively , should not be confused wit
40	2004	NF	StonewallRiots	A	B	C	With the onset of Prohibition , artists , intellectuals , and	gay	men	and	lesbians	began to socialize more and more in tearc
41	2002	MAG	PsychToday	A	B	C	. Silence is unacceptable . Hosexuality Some 80 percent of	gay	men	and	women	have experienced verbal or physical harass
42	2001	FIC	SweetSuccess	A	B	C	, but he has his limitations . As a rule ,	gay	men	do	n't	have children that often . Second , he 's
43	2002	MAG	WashMonth	A	B	C	many ways they promote it . It is clear that many	gay	men	find	anonymous	sex appealing , and that as long as the

FIGURE 2.28 COHA: collocates of *gay*, 2000s



CONTEXT	ALL	1810	1820	1830	1840	1850	1860	1870	1880	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000
BRIGHT	173	1	5	8	10	14	13	23	12	14	12	4	12	12	8	11	8	4	2		
FLOWERS	158		5	14	11	18	10	19	17	7	13	10	11	7	5	6	1	3		1	
LESBIAN	155							1			1							1	6	67	79
LAUGH	143			8	5	15	13	12	14	9	13	14	9	11	2	4	7	4			
GAY	142		2	5	6	7	16	7	12	11	13	2	6	16	5	2	6	6	2	10	8
COLORS	136		3	6	5	13	14	9	9	10	5	7	11	7	17	8	5	6	1		
RIGHTS	134																	7	19	50	58
GRAVE	132		6	15	14	10	15	8	13	13	18	8	5	4	1	1					1
MARRIAGE	99				1		1	1					1	1		1		1		7	85
GALLANT	91	1	8	11	12	4	10	8	9	6	6	1	9	3	2	1					
LAUGHTER	90				5	5	7	6	8	4	8	15	8	11	3	2	3	2	3	1	
LESBIANS	85																1		6	40	38
SPIRITS	82		2	3	8	8	7	7	9	8	6	6	4	9	3	1	1				

FIGURE 2.29 COHA: collocates of *gay* by decade

SEC 1 (1850, 1860, 1870, 1880, 189...): 137,803,825 WORDS

	WORD/PHRASE	TOKENS 1	TOKENS 2	PM 1	PM 2	RATIO
1	GRAVE	85	1	0.6	0.0	65.8
2	LADY	61	0	0.4	0.0	44.3
3	SPIRITS	51	0	0.4	0.0	37.0
4	HEART	44	1	0.3	0.0	34.0
5	VOICES	44	1	0.3	0.0	34.0
6	BRILLIANT	45	0	0.3	0.0	32.7
7	GLAD	45	0	0.3	0.0	32.7
8	GALLANT	44	0	0.3	0.0	31.9
9	PORTUGUESE	44	0	0.3	0.0	31.9
10	ATTIRE	40	1	0.3	0.0	31.0
11	THRONG	39	0	0.3	0.0	28.3
12	LOOKED	37	0	0.3	0.0	26.8
13	SONG	37	0	0.3	0.0	26.8

SEC 2 (1970, 1980, 1990, 2000): 106,640,094 WORDS

	WORD/PHRASE	TOKENS 2	TOKENS 1	PM 2	PM 1	RATIO
1	RIGHTS	134	0	1.3	0.0	125.7
2	LESBIAN	153	2	1.4	0.0	98.9
3	LESBIANS	84	0	0.8	0.0	78.8
4	MARRIAGE	93	2	0.9	0.0	60.1
5	BAR	40	1	0.4	0.0	51.7
6	ACTIVISTS	32	1	0.3	0.0	41.4
7	STRAIGHT	61	2	0.6	0.0	39.4
8	OPENLY	30	1	0.3	0.0	38.8
9	ABORTION	40	0	0.4	0.0	37.5
10	COUPLES	28	1	0.3	0.0	36.2
11	COMMUNITY	81	3	0.8	0.0	34.9
12	MALE	24	1	0.2	0.0	31.0
13	BISEXUAL	31	0	0.3	0.0	29.1

FIGURE 2.30 COHA: collocates of *gay*, 1850s–1910s vs 1970s–2000s

SEC 1 (1850, 1860, 1870, 1880, 1890): 137,803,825 WORDS

	WORD/PHRASE	TOKENS 1	TOKENS 2	PM 1	PM 2	RATIO
1	STRONG-MINDED WOMEN	24	1	0.2	0.0	18.6
2	CLEVER WOMEN	22	0	0.2	0.0	16.0
3	TRUE WOMEN	18	1	0.1	0.0	13.9
4	NOBLE WOMEN	33	2	0.2	0.0	12.8
5	UNFORTUNATE WOMEN	16	1	0.1	0.0	12.4
6	ELDER WOMEN	15	0	0.1	0.0	10.9
7	HELPLESS WOMEN	55	4	0.4	0.0	10.6
8	WRETCHED WOMEN	14	0	0.1	0.0	10.2
9	TURKISH WOMEN	13	1	0.1	0.0	10.1
10	FAIR WOMEN	63	5	0.5	0.0	9.8
11	PURE WOMEN	13	0	0.1	0.0	9.4
12	NOBLEST WOMEN	12	0	0.1	0.0	8.7
13	DEVOTED WOMEN	12	0	0.1	0.0	8.7
14	CULTIVATED WOMEN	12	0	0.1	0.0	8.7
15	REFINED WOMEN	12	0	0.1	0.0	8.7
16	ABANDONED WOMEN	12	0	0.1	0.0	8.7

SEC 2 (1970, 1980, 1990, 2000s): 106,640,094 WORDS

	WORD/PHRASE	TOKENS 2	TOKENS 1	PM 2	PM 1	RATIO
1	PREGNANT WOMEN	233	4	2.2	0.0	75.3
2	BATTERED WOMEN	70	0	0.7	0.0	65.6
3	AFRICAN-AMERICAN WOMEN	61	0	0.6	0.0	57.2
4	BLACK WOMEN	487	15	4.6	0.1	42.0
5	DIVORCED WOMEN	25	1	0.2	0.0	32.3
6	SOVIET WOMEN	32	0	0.3	0.0	30.0
7	MUSLIM WOMEN	23	1	0.2	0.0	29.7
8	MIDDLE-CLASS WOMEN	23	1	0.2	0.0	29.7
9	NATIONAL WOMEN	68	3	0.6	0.0	29.3
10	MENOPAUSAL WOMEN	22	1	0.2	0.0	28.4
11	CATHOLIC WOMEN	21	1	0.2	0.0	27.1
12	ABUSED WOMEN	19	1	0.2	0.0	24.6
13	AFGHAN WOMEN	26	0	0.2	0.0	24.4
14	ADULT WOMEN	18	1	0.2	0.0	23.3
15	LOCAL WOMEN	22	0	0.2	0.0	20.6
16	LOW-INCOME WOMEN	21	0	0.2	0.0	19.7

FIGURE 2.31 COHA: ADJ collocates of *women*, 1850s–1910s vs 1970s–2000s

SEC 1 (1930, 1940, 1950): 73,495,401 WORDS

	WORD/PHRASE	TOKENS 1	TOKENS 2	PM 1	PM 2	RATIO
1	WEAR	31	1	0.4	0.0	30.8
2	MISSES	15	0	0.2	0.0	20.4
3	FABRICS	13	0	0.2	0.0	17.7
4	GOLFERS	12	0	0.2	0.0	16.3
5	STREAM	10	0	0.1	0.0	13.6
6	FACTORIES	13	1	0.2	0.0	12.9
7	COAST	9	0	0.1	0.0	12.2
8	WARTIME	9	0	0.1	0.0	12.2
9	FOLKS	22	2	0.3	0.0	10.9
10	CASTE	8	0	0.1	0.0	10.9
11	DEPUTIES	8	0	0.1	0.0	10.9
12	HIPS	8	0	0.1	0.0	10.9
13	LEISURE	7	0	0.1	0.0	9.5
14	LIE	7	0	0.1	0.0	9.5
15	WARDROBE	7	0	0.1	0.0	9.5

SEC 2 (1960, 1970, 1980): 73,108,401 WORDS

	WORD/PHRASE	TOKENS 2	TOKENS 1	PM 2	PM 1	RATIO
1	BLACKS	80	0	1.1	0.0	109.4
2	LIBERATION	102	1	1.4	0.0	102.5
3	LIB	72	0	1.0	0.0	98.5
4	MINORITIES	61	1	0.8	0.0	61.3
5	ISSUES	33	0	0.5	0.0	45.1
6	PERCENT	145	4	2.0	0.1	36.4
7	ABORTIONS	36	1	0.5	0.0	36.2
8	CAUCUS	25	0	0.3	0.0	34.2
9	RISK	19	0	0.3	0.0	26.0
10	AIDS	16	0	0.2	0.0	21.9
11	LIBERATIONISTS	14	0	0.2	0.0	19.1
12	RESEARCH	18	1	0.2	0.0	18.1
13	ACTIVISTS	12	0	0.2	0.0	16.4
14	PRESSURE	12	0	0.2	0.0	16.4
15	STUDIES	32	2	0.4	0.0	16.1

FIGURE 2.32 COHA: NOUN collocates of *women*, 1930s–1950s vs 1960s–1980s

SEC 1 (1820, 1830, 1840, 1850, 186...): 130,936,953 WORDS

	WORD/PHRASE	TOKENS 1	TOKENS 2	PM 1	PM 2	RATIO
1	ABSOLUTE	45	1	0.3	0.0	36.6
2	MERE	37	1	0.3	0.0	30.1
3	PURE	113	4	0.9	0.0	23.0
4	BEAUTIFUL	28	0	0.2	0.0	21.4
5	EVANGELICAL	48	2	0.4	0.0	19.5
6	ESSENTIAL	45	2	0.3	0.0	18.3
7	HOLY	112	5	0.9	0.0	18.2
8	GENERAL	42	2	0.3	0.0	17.1
9	BLESSED	22	0	0.2	0.0	16.8
10	UNDEFILED	22	0	0.2	0.0	16.8
11	DESTITUTE	21	0	0.2	0.0	16.0
12	PRACTICAL	78	4	0.6	0.0	15.9
13	OUTWARD	17	1	0.1	0.0	13.8
14	SUBLIME	17	1	0.1	0.0	13.8

SEC 2 (1970, 1980, 1990, 2000): 106,640,094 WORDS

	WORD/PHRASE	TOKENS 2	TOKENS 1	PM 2	PM 1	RATIO
1	ORGANIZED	54	1	0.5	0.0	66.3
2	HINDU	13	1	0.1	0.0	16.0
3	OLD-TIME	17	0	0.2	0.0	15.9
4	SIKH	15	0	0.1	0.0	14.1
5	CIVIC	15	0	0.1	0.0	14.1
6	ETHNIC	12	0	0.1	0.0	11.3
7	CULTURAL	12	0	0.1	0.0	11.3
8	EASTERN	16	2	0.2	0.0	9.8
9	TRADITIONAL	16	2	0.2	0.0	9.8
10	CONTEMPORARY	10	0	0.1	0.0	9.4
11	MEDIEVAL	10	0	0.1	0.0	9.4
12	MAJOR	9	0	0.1	0.0	8.4
13	BIG	8	0	0.1	0.0	7.5
14	ISLAMIC	7	0	0.1	0.0	6.6

FIGURE 2.33 COHA: ADJ collocates of *religion*, 1800s vs 1970s–2000s

pluralistic (*Hindu, Sikh, Eastern, Islamic*); there is discussion of *old-time* (presumably Christian) religion; and religion is viewed through an academic, objective lens (*organized, civic, ethnic, major, contemporary*).

As we have seen, a simple search of collocates in GloWbE provided interesting insight into societal and cultural differences between countries (*belief, family, wife*). In COHA, we find that collocates provide great insight into societal and cultural changes, whether it is *gays, women*, or *religion*. Again, however, we see the importance of corpus size. At 400 million words, COHA enables us to compare collocates in ways that would be quite impossible with a tiny 4–5 million word corpus, where there might be only 1/100th the number of tokens. In addition, we again see the crucial importance of a corpus architecture and interface that is expressly designed to allow users to quickly and easily compare collocates across different sections of the corpus.

Why Not Use Google Books (and Culturomics)?

Soon after COHA was released in late 2010, Google released the searchable "Google Books Ngrams" and the "Ngrams Viewer" (https://books.google.com/ngrams), which allows users to search through incredibly large datasets of historical English (and other languages as well). For example, the American English n-grams are based on more than 155 billion words in millions of books, which makes the Google Books dataset more than 400 times as large as COHA. At the same time that Google Books n-grams were released, the "Culturomics" project was announced (Michel et al., 2011), which showed how the Google Books data could shed light on many different cultural shifts from the last 200 years.

The Google Books project is a great resource for the study of historical English. But, when we look at the frequency of specific words and phrases over time, Google Books often provides data that is fairly similar to what was already available from COHA. For example, Figures 2.34–35 (for the word *steamboat*) and Figures 2.36–37 (for *teenager*) show that the data from COHA and Google Books is nearly the same.

Davies and Chapman (2016) give a much more detailed comparison of the lexis in COHA and Google Books, and it shows that this similarity holds for thousands of words selected at random. In other words, to the degree that Google Books (and "Culturomics") shows interesting cultural shifts over time, much of this is already available in COHA.

What distinguishes COHA from Google Books Ngrams, however, is the range of queries that each corpus allows. Google Books is basically limited to just showing the frequency for individual words and phrases. In other words, a user needs to know exactly what word to look for before doing a search. But Google Books cannot generate a list of all words that are more common in one period (e.g., **ism* word or nouns in the late 1900s compared to the 1800s), as can be done with COHA. In addition, it does not allow users to compare the collocates of a word in different periods (e.g., collocates of *gay, women*, or *religion*),

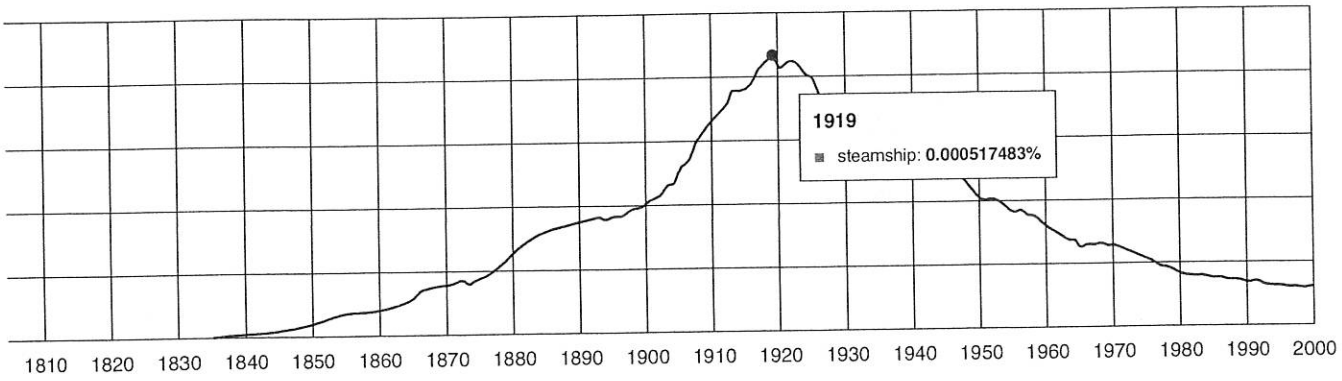


FIGURE 2.34 Google Books: *steamship*

SECTION	1810	1820	1830	1840	1850	1860	1870	1880	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000
FREQ	0	0	0	6	17	56	29	86	191	236	246	219	221	137	96	45	55	22	18	15
PER MIL	0.00	0.00	0.00	0.37	1.03	3.28	1.56	4.23	9.27	10.68	10.84	8.54	8.98	5.63	3.91	1.88	2.31	0.87	0.64	0.51
SEE ALL YEARS AT ONCE																				

FIGURE 2.35 COHA: *steamship*

1810	1820	1830	1840	1850	1860	1870	1880	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000
378	655	1,437	1,938	2,953	2,353	2,844	4,408	5,632	7,520	10,087	7,089	5,795	6,167	8,104	13,192	14,011	15,511	19,816	26,882
1	5	0	0	0	0	0	2	1	3	1	13	20	39	1,321	8,422	21,496	50,645	97,783	151,757
0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.16	0.64	153	3.27	4.93	5.65

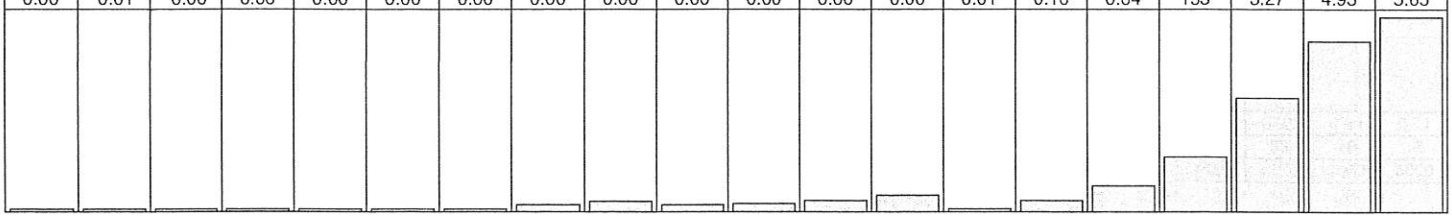


FIGURE 2.36 Google Books: *teenager*

SECTION	1810	1820	1830	1840	1850	1860	1870	1880	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000
FREQ	0	0	0	0	0	0	0	0	0	0	0	0	0	9	44	107	188	336	761	996
PER MIL	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.37	1.79	4.46	7.89	12.27	27.24	33.69
SEE ALL YEARS AT ONCE																				

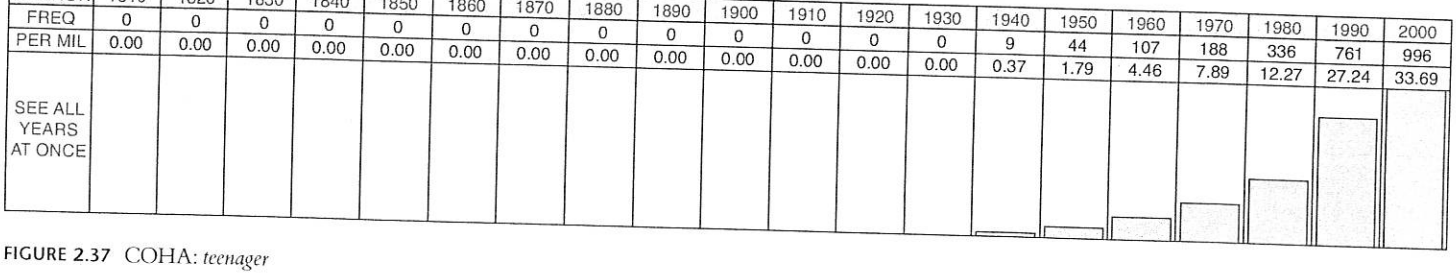


FIGURE 2.37 COHA: *teenager*

as is possible with COHA. In other words, the standard Google Books interface does not allow the types of searches that are the most useful for looking at cultural shifts over time.

Fortunately, however, the Google Books team released the entire set of n-grams when they released the standard Google Books interface, and this allows others to download and then use this "raw data" in other architectures and interfaces. In 2011, we used this data to create the BYU Google Books corpus (googlebooks.byu.edu), which has an interface very similar to what is available for COHA (see Davies, 2014). With this interface, users can see the frequency of all matching words in each decade of the corpus. For example, Figure 2.38 shows **ism* words during the last 200 years in Google Books. In addition, as with COHA, users can compare all these words in different historical periods. For example, Figure 2.39 compares **ism* words in the 1860s–1910s (left) and the 1970s–2000s (right).

As with COHA, users can also compare the collocates across the 200 years of the corpus, as with the collocates of *gay* shown in Figure 2.40.

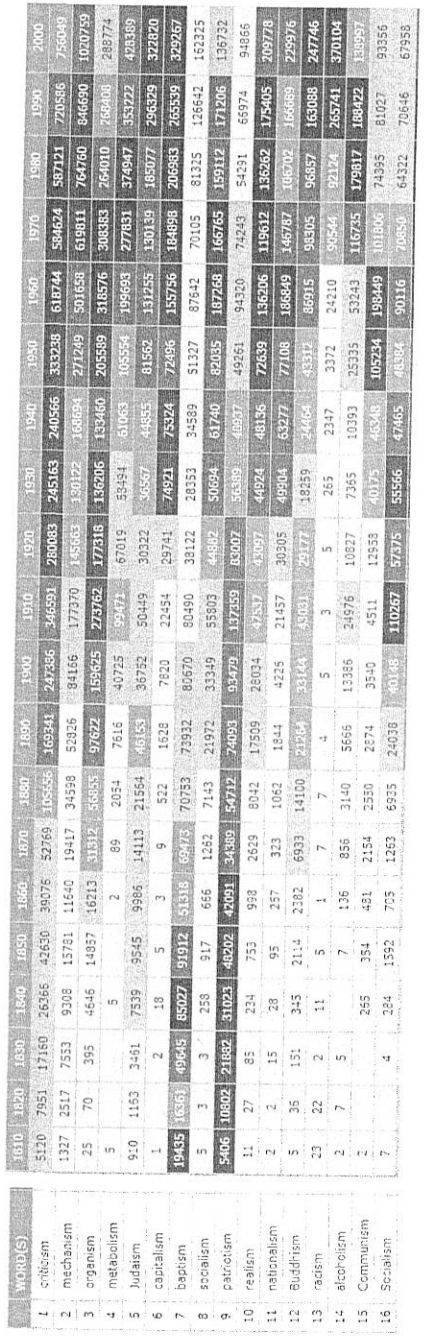
And as with COHA, users can compare the collocates of a word in two different historical periods. For example, Figure 2.41 shows the collocates of *gay* in the 1810s–1890s (left) and the 1980s–2000s (right), and Figure 2.42 shows the collocates of *women* in the 1850s–1910s vs the 1970s–2000s (right). In both cases, the data from the BYU Google Books corpus agrees quite well with the data from COHA.

In summary, the BYU Google Books corpus allows researchers to use collocates to see cultural shifts for a wide range of topics, such as those in Table 2.1.

And again, none of this is possible with the standard Google Books interface, since it does not allow users to compare collocates over time. In other words, researchers who are interested in "Culturomics" (cf. Michel et al., 2011) would likely find the BYU Google Books interface to be of much more value than the simplistic charts to see the frequency of individual words, which is the only thing that is possible with the standard Google Books interface.

TABLE 2.1 Google Books (BYU): changing collocates over time

	Older period	*More recent period*
women	1930s–1950s: *ridiculous, plump, loveliest, restless, agreeable*	1960s–1980s: *battered, militant, college-educated, liberated*
art	1830s–1910s: *noble, classic, Grecian*	1960s–2000: *abstract, Asian, African, commercial*
fast	1850s–1910s: *mail, train, horses, steamers*	1960s–2000s: *food, track, lane, buck*
music	1850s–1910s: *delightful, exquisite, sweeter, tender*	1970s–2000s: *Western, black, electronic, recorded*
food	1850s–1910s: *spiritual, insufficient, unwholesome, mental*	1970s–2000s: *fast, Chinese, Mexican, organic*



	WORD(S)	1810	1820	1830	1840	1850	1860	1870	1880	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000
1	criticism	5120	7951	17160	26366	42630	39076	52769	105566	169341	247386	346591	280083	245163	240566	333238	618744	584624	587121	720586	756049
2	mechanism	1327	2517	7553	9308	15781	11640	19417	34598	52826	84156	177370	145663	138122	168694	271249	501658	619811	764760	846690	1020759
3	organism	25	70	395	4646	14857	16213	31819	56855	97622	159625	273762	177316	136206	133460	205589	318576	308383	264010	768410	288774
4	metabolism	5					2	89	2054	7616	40725	99471	67019	58494	61063	105594	199693	277831	374947	353222	428389
5	Judaism	910	1163	3461	7539	9545	9986	14113	21564	46153	36752	50449	30322	46567	44855	81562	131255	130139	185077	296329	322820
6	capitalism	1		2	18	5	3	9	522	1628	7620	22494	25741	74921	75324	72496	155756	184898	206983	265539	329267
7	baptism	19435	1636	49645	85027	91912	51318	69475	70753	73932	80670	80490	38122	28353	34589	51327	87642	70105	81325	126642	162325
8	socialism	5	3	3	258	917	666	1262	7143	21972	33349	55803	44882	50694	61740	82035	187268	165765	159112	171206	136732
9	patriotism	5406	10802	21882	31023	48202	47091	34389	54712	74093	93479	117359	83007	56389	48907	49261	94320	74243	54291	65974	94866
10	realism	11	27	85	234	753	938	2629	8042	17509	28034	42537	43997	44924	48136	72639	136206	119612	135262	175405	209778
11	nationalism	2	2	15	28	95	257	323	1062	1844	4225	21457	30305	49904	63277	77108	186849	146787	106702	166089	229976
12	Buddhism	5	36	151	345	2114	2382	6933	14100	22364	33144	50601	29177	18259	24464	43312	86915	98305	96857	163088	247746
13	racism	23	22	2	11	5	1	7	7	4	5	3	5	265	2347	3372	24210	90544	92124	265741	370104
14	alcoholism	2	7	5		7	136	856	3140	5666	13386	24976	10827	7365	10393	25335	53243	116735	179817	184422	138937
15	Communism	2			265	354	481	2154	2550	2674	3540	4511	12538	40175	46548	105234	198449	101800	74395	81027	93356
16	Socialism	7		4	284	1592	755	1263	6935	24038	40148	110267	57375	55566	47465	41304	90116	70859	64322	70646	67958

FIGURE 2.38 Google Books (BYU): **ism* words by decade

SEC 1: 32.8 BILLION WORDS (1860-1919)

	WORD/PHRASE	1: 1860-1919	2: 1970-2009	P/BIL 1	P/BIL 2	RATIO
1	aneurism	75,991	4,072	2,313.7	53.4	43.31
2	traumatism	30,871	2,180	939.9	28.6	32.86
3	heathenism	48,315	8,048	1,471.0	105.6	13.93
4	galvanism	17,644	2,963	537.2	38.9	13.82
5	Mohammedanism	35,944	6,424	1,094.4	84.3	12.98
6	Romanism	36,846	7,110	1,121.8	93.3	12.03
7	bimetallism	16,714	3,729	508.9	48.9	10.40
8	Pantheism	23,926	7,368	728.5	96.7	7.54
9	rheumatism	203,355	64,562	6,191.5	847.1	7.31
10	pauperism	43,132	15,642	1,313.2	205.2	6.40
11	despotism	212,283	98,543	6,463.4	1,292.9	5.00
12	Brahmanism	12,874	6,289	392.0	82.5	4.75
13	Congregationalism	25,192	12,629	767.0	165.7	4.63

SEC 2: 76.2 BILLION WORDS (1970-2009)

	WORD/PHRASE	2: 1970-2009	1: 1860-1919	P/BIL 2	P/BIL 1	RATIO
1	consumerism	86,941	1	1,140.7	0.0	37,464.05
2	existentialism	66,111	1	867.4	0.0	28,488.12
3	environmentalism	47,385	1	621.7	0.0	20,418.84
4	Surrealism	46,800	1	614.0	0.0	20,166.75
5	isolationism	42,459	1	557.1	0.0	18,296.16
6	Racism	161,705	5	2,121.6	0.2	13,936.17
7	racism	818,513	27	10,738.9	0.8	13,063.27
8	Sexism	46,226	2	606.5	0.1	9,959.70
9	McCarthyism	35,259	2	462.6	0.1	7,596.79
10	minimalism	17,005	1	223.1	0.0	7,327.68
11	sexism	193,193	12	2,534.7	0.4	6,937.46
12	Pentecostalism	24,987	2	327.8	0.1	5,383.62
13	surrealism	45,507	4	597.1	0.1	4,902.40

FIGURE 2.39 Google Books (BYU): **ism* words, 1860s–1910s vs 1970s–2000s

WORD(S)	CHARTS	TOTAL	1810	1820	1830	1840	1850	1860	1870	1880	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000
gay men	G	248927	5	3	17	38	47	14	12	25	24	45	68	41	31	24	50	47	2592	20837	107564	117443
gay people	G	65106	17	21	71	74	150	117	135	174	232	255	274	211	146	200	230	236	4898	7252	24546	25867
gay community	G	52955					1		4	3	2	7	4	2	4	2	5	20	2055	6446	22481	21919
gay rights	G	48261																2	1079	4387	16990	25803
gay man	G	43532	17	28	59	96	124	87	92	96	106	94	76	77	58	78	78	98	496	2790	16539	22443
gay life	G	27995	18	30	124	139	312	316	390	581	791	1107	1406	1122	948	715	872	1269	1965	2666	6761	6463
gay liberation	G	23215									1					1		2	4049	3421	7675	8066
gay world	G	19378	131	218	531	701	1059	661	658	847	1120	1092	867	436	314	231	301	462	2052	1844	3257	2596
gay bars	G	16267									1			1	3	5	19	163	1999	2197	5547	6332
gay identity	G	16081																	346	1338	6568	7829
gay bar	G	15991									1		1	1	3	3	29	135	1643	1884	5357	6934
gay marriage	G	15746			1	6	3	4		17	4	6	5			8	2	2	131	156	1770	13631
gay movement	G	12936						10	7	23	17	19	20	14	23	22	24	15	895	1432	4915	5500
gay couples	G	12461				4		4	2		7	7	7	15	8	10	14	44	290	1230	3882	6937
gay colors	G	10365		4	75	196	489	433	546	887	819	916	879	840	723	693	765	787	443	269	288	313

FIGURE 2.40 Google Books (BYU): NOUN collocates of *gay* by decade

SEC 1: 22.6 BILLION WORDS (1810-1899)

	WORD/PHRASE	1: 1810-1899	2: 1980-2009	P/Bil 1	P/Bil 2	RATIO
1	gay court	845	66	37.4	1.1	35.25
2	gay birds	770	65	34.1	1.0	32.21
3	gay plumage	1,051	90	46.5	1.4	32.15
4	gay companions	2,005	174	88.7	2.8	31.72
5	gay attire	2,606	230	115.3	3.7	31.19
6	gay dresses	1,329	137	58.8	2.2	26.70
7	gay season	936	104	41.4	1.7	24.78
8	gay flowers	1,928	230	85.3	3.7	23.08
9	gay dress	850	103	37.6	1.7	22.72
10	gay throng	1,434	181	63.5	2.9	21.81
11	gay company	2,661	349	117.8	5.6	20.99
12	gay appearance	955	128	42.3	2.1	20.54
13	gay spirits	1,490	231	65.9	3.7	17.76
14	gay clothing	779	124	34.5	2.0	17.29

SEC 2: 62.6 BILLION WORDS (1980-2009)

	WORD/PHRASE	2: 1980-2009	1: 1810-1899	P/Bil 2	P/Bil 1	RATIO
1	gay liberation	19,162	1	308.0	0.0	6,960.74
2	gay bar	14,175	1	227.9	0.0	5,149.17
3	gay bars	14,076	1	226.3	0.0	5,113.21
4	gay culture	9,381	1	150.8	0.0	3,407.72
5	gay parents	6,838	1	109.9	0.0	2,483.95
6	gay communities	6,713	1	107.9	0.0	2.428.55
7	gay community	50,846	10	817.3	0.4	1,847.02
8	gay history	3,024	1	48.6	0.0	1,098.49
9	gay rights	47,180	0	758.4	0.0	758.41
10	gay sensibility	2,075	1	33.4	0.0	753.76
11	gay individuals	1,702	1	27.4	0.0	618.26
12	gay newspaper	1,396	1	22.4	0.0	507.11
13	gay men	245,844	185	3,951.9	8.2	482.73
14	gay partners	1,051	1	16.9	0.0	381.78

FIGURE 2.41 *gay* + NOUN, 1800s vs 1980s–2000s

SEC 1: 35.8 BILLION WORDS (1850-1919)

	WORD/PHRASE	1: 1850-1919	2: 1970-2009	P/Bil 1	P/Bil 2	RATIO
1	feeble women	816	149	22.8	2.0	11.66
2	fair women	8,212	1,529	229.4	20.1	11.44
3	Fair women	524	104	14.6	1.4	10.73
4	chief women	699	139	19.5	1.8	10.71
5	delicate women	2,651	605	74.1	7.9	9.33
6	defenceless women	1,474	341	41.2	4.5	9.20
7	tender women	816	190	22.8	2.5	9.14
8	noblest women	975	229	27.2	3.0	9.07
9	handsomest women	1,037	249	29.0	3.3	8.87
10	nervous women	1,895	476	52.9	6.2	8.48
11	Grecian women	600	154	16.8	2.0	8.30
12	honourable women	731	204	20.4	2.7	7.63
13	fairest women	547	156	15.3	2.0	7.47
14	agreeable women	500	147	14.0	2.5	7.24
15	amiable women	654	194	18.3	2.0	7.18

SEC 2: 76.2 BILLION WORDS (1870-2009)

	WORD/PHRASE	2: 1970-2009	1: 1850-1919	P/Bil 2	P/Bil 1	RATIO
1	bisexual women	10,784	1	141.5	0.0	5,064.71
2	battered women	83,346	10	1,093.5	0.3	3,914.35
3	heterosexual women	21,388	4	280.6	0.1	2,511.22
4	academic women	4,253	1	55.8	0.0	1,997.42
5	negative women	3,696	2	48.5	0.1	867.91
6	urban women	10,521	7	138.0	0.2	705.88
7	Black women	102,960	69	1,350.8	1.9	700.80
8	overweight women	4,287	3	56.2	0.1	671.13
9	Inuit women	1,163	1	15.3	0.0	546.20
10	Jamaican women	1,153	1	15.1	0.0	541.51
11	Thai women	3,344	3	43.9	0.1	523.50
12	pioneering women	1,075	1	14.1	0.0	504.87
13	glamorous women	1,051	1	13.8	0.0	493.60
14	untreated women	1,008	1	13.2	0.0	473.41
15	indigenous women	7,204	8	94.5	0.2	422.92

FIGURE 2.42 ADJ + *women*, 1850s–1910s vs 1970s–2000s

Examining Very Recent Variation and Change With the COCA and NOW Corpora

COHA and the Google Books corpus are useful for looking at long-term shifts, but for greater detail on very recent changes, many researchers prefer the Corpus of Contemporary American English (COCA) and the NOW (News on the Web) corpus. COCA (which was released in 2008; see Davies, 2009, 2011) currently contains 520 million words of text from 1990–2015, and 20 million words are added each year. NOW (which was released in 2016) currently has about 3.7 billion words of text (as of October 2016), and it grows by 4–5 million words *each day*.

Using COCA, users can input any word, phrase (or even syntactic construction) and see the frequency in five year blocks from 1990 to the present (for many more examples, see Davies, 2011). For instance, Figure 2.43 shows the decrease in the word *retarded* (which is now considered a very insensitive term) as well as the term *blacks* (compared to *African Americans* or *people of color*).

Or consider the shift from *global warming* to *climate change* (Figure 2.44), which is of course related to contentious issues about what is actually happening over time.

As with GloWbE and COHA, we can carry out much more sophisticated searches as well. For example, we can examine changes in collocates to investigate semantic change. Figure 2.45, for example, shows the collocates of *web* in 1990–1994 (left) and 2005–2015 (right), and we can clearly see how the growth of the (World Wide) Web during this time has affected what are the most common collocates (e.g., *site, page, browser, search, company, content*).

Another search might consider how *green* has acquired the new meaning "environmentally friendly" during this time (Figure 2.46). In the early 1990s (left), *green* often referred to physical objects that were literally green (*chimneys, cross, cloak, meadows, ink, coat*), and people from that time would have been confused by *green* with *zone, jobs, technology*, or *economy*, which are some of the most common collocates in 2005–2015.

Finally, we can use collocates to see what people are talking about for more general topics. For example, Figure 2.47 compares words "nearby" the word *crisis* in the early 1990s (left) and 2005–2015 (right). In the early 1990s, people were concerned about the *Persian GulfWar, Kuwait*, and *Saddam Hussein*, the Savings and Loan crisis, and events related to *Gorbachev* and the former *Yugoslavia*. During the past ten years, however, people have been worrying and talking about the *2008 subprime mortgage crisis* (which resulted in many *foreclosures*), as well as problems with the *Euro* in the *Eurozone* (especially in *Greece*) and other hotspots like *Ukraine* and *Darfur*.

With the NOW corpus, the level of detail is increased even more. As has been mentioned, every day 4–5 million words (from about 10,000 newspaper and magazine articles) are added to the corpus or about 130 million words per month and 1.5 billion words per year. Users can see the frequency of any word or phrase in time periods as short as ten days (currently about 250 ten-day periods from

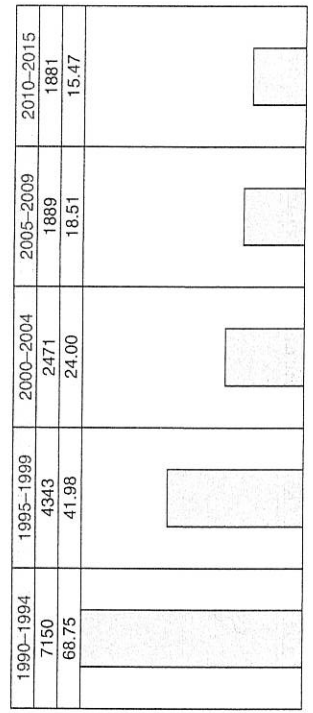


FIGURE 2.43B COCA: *blacks*

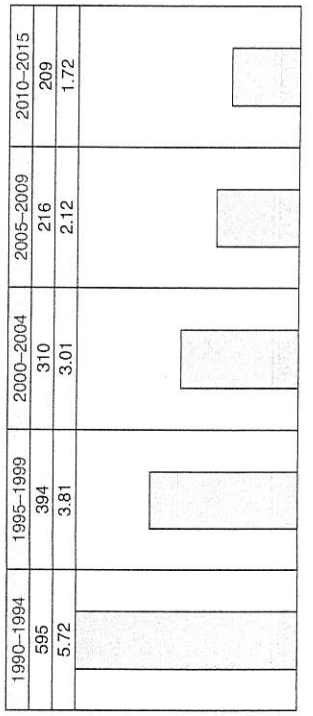


FIGURE 2.43A COCA: *retarded*

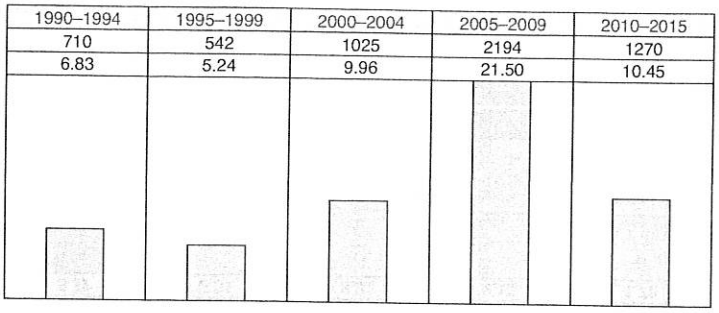


FIGURE 2.44A COCA: *global warming*

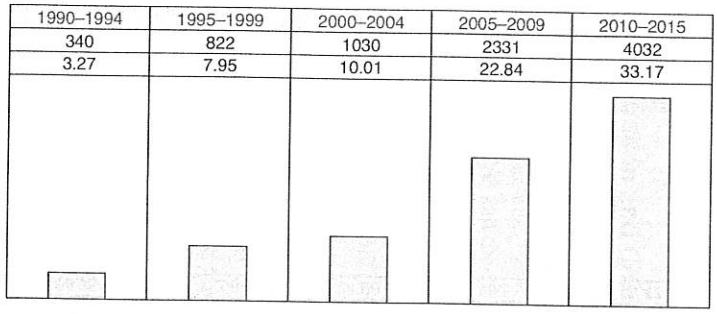


FIGURE 2.44B COCA: *climate change*

SEC 1 (1990-1994): 103,999,130 WORDS

	WORD/PHRASE	TOKENS 1	TOKENS 2	PM 1	PM 2	RATIO
1	SILVER	6	0	0.1	0.0	5.8
2	STRANDS	10	4	0.1	0.0	5.4
3	INTRIGUE	8	4	0.1	0.0	4.3
4	IMAGE	16	9	0.2	0.0	3.8
5	SPIDER	162	182	1.6	0.8	1.9
6	NETWORKS	15	18	0.1	0.1	1.8
7	DECEIT	6	9	0.1	0.0	1.4
8	RELATIONSHIPS	20	37	0.2	0.2	1.2
9	POWER	9	18	0.1	0.1	1.1
10	RELATIONS	9	19	0.1	0.1	1.0
11	LIFE	35	74	0.3	0.3	1.0
12	LINES	6	15	0.1	0.1	0.9
13	LIES	7	25	0.1	0.1	0.6
14	SPIDERS	12	47	0.1	0.2	0.5

SEC 2 (2005-2009, 2010-2015): 223,609,936 WORDS

	WORD/PHRASE	TOKENS 2	TOKENS 1	PM 2	PM 1	RATIO
1	SITE	8190	0	36.6	0.0	3,662.6
2	SITES	2047	1	9.2	0.0	952.0
3	PAGES	362	0	1.6	0.0	161.9
4	E-MAIL	358	0	1.6	0.0	160.1
5	PAGE	526	2	2.4	0.0	122.3
6	BROWSER	251	0	1.1	0.0	112.2
7	HTTP	198	0	0.9	0.0	88.5
8	SEARCH	186	0	0.8	0.0	83.2
9	VIDEO	174	1	0.8	0.0	80.9
10	COMPANY	173	1	0.8	0.0	80.5
11	TOOLS	174	0	0.8	0.0	77.8
12	CONTENT	171	0	0.8	0.0	76.5
13	ADDRESS	164	1	0.7	0.0	76.3
14	RESOURCES	160	1	0.7	0.0	74.4

FIGURE 2.45 COCA: collocates of *web*, 1990–1994 vs 2005–2015

SEC 1 (1990-1994): 103,999,130 WORDS

	WORD/PHRASE	TOKENS 1	TOKENS 2	PM 1	PM 2	RATIO
1	GREEN CHIMNEYS	20	2	0.2	0.0	21.5
2	GREEN PLAN	30	5	0.3	0.0	12.9
3	GREEN CONSUMER	11	0	0.1	0.0	10.6
4	GREEN CROSS	40	10	0.4	0.0	8.6
5	GREEN CLOAK	29	9	0.3	0.0	6.9
6	GREEN ISLAND	21	7	0.2	0.0	6.5
7	GREEN SEAL	36	13	0.3	0.1	6.0
8	GREEN POINT	12	5	0.1	0.0	5.2
9	GREEN KNIGHT	14	6	0.1	0.0	5.0
10	GREEN MEADOWS	13	6	0.1	0.0	4.7
11	GREEN PEPPER	95	60	0.9	0.3	3.4
12	GREEN INK	12	9	0.1	0.0	2.9
13	GREEN TRUCK	12	9	0.1	0.0	2.9
14	GREEN COAT	18	15	0.2	0.1	2.6

SEC 2 (2005-2009, 2010-2015): 223,609,936 WORDS

	WORD/PHRASE	TOKENS 2	TOKENS 1	PM 2	PM 1	RATIO
1	GREEN ZONE	268	0	1.2	0.0	119.9
2	GREEN JOBS	167	0	0.7	0.0	74.7
3	GREEN BUILDING	264	2	1.2	0.0	61.4
4	GREEN GAZETTE	99	0	0.4	0.0	44.3
5	GREEN TECHNOLOGY	69	1	0.3	0.0	32.1
6	GREEN GARLIC	59	1	0.3	0.0	27.4
7	GREEN ROOFS	61	0	0.3	0.0	27.3
8	GREEN ECONOMY	61	0	0.3	0.0	27.3
9	GREEN ENERGY	175	3	0.8	0.0	27.1
10	GREEN LANTERN	39	1	0.2	0.0	18.1
11	GREEN ARROW	34	1	0.2	0.0	15.8
12	GREEN DESIGN	33	1	0.1	0.0	15.3
13	GREEN BUILDINGS	30	0	0.1	0.0	13.4
14	GREEN COFFEE	28	1	0.1	0.0	13.0

FIGURE 2.46 COCA: collocates of *green*, 1990–1994 vs 2005–2015

SEC 1 (1990-1994): 103,999,130 WORDS

	WORD/PHRASE	TOKENS 1	TOKENS 2	PM 1	PM 2	RATIO
1	PERSIAN	371	1	3.6	0.0	797.7
2	GULF	1216	16	11.7	0.1	163.4
3	S&L	51	1	0.5	0.0	109.7
4	VOICE-OVER	35	1	0.3	0.0	75.3
5	SADDAM	29	1	0.3	0.0	62.4
6	KOPPEL	44	0	0.4	0.0	42.3
7	TED	37	2	0.4	0.0	39.8
8	CONVERSATIONS	16	1	0.2	0.0	34.4
9	&L	62	4	0.6	0.0	33.3
10	KUWAIT	32	0	0.3	0.0	30.8
11	SAM	14	1	0.1	0.0	30.1
12	HUSSEIN	27	0	0.3	0.0	26.0
13	GORBACHEV	25	0	0.2	0.0	24.0
14	YUGOSLAVIA	25	0	0.2	0.0	24.0

SEC 2 (2005-2009, 2010-2015): 223,609,936 WORDS

	WORD/PHRASE	TOKENS 2	TOKENS 1	PM 2	PM 1	RATIO
1	2008	226	0	1.0	0.0	101.1
2	MORTGAGE	144	0	0.6	0.0	64.4
3	FORECLOSURE	111	0	0.5	0.0	49.6
4	UKRAINE	99	1	0.4	0.0	46.0
5	SUBPRIME	75	0	0.3	0.0	33.5
6	EURO	67	0	0.3	0.0	30.0
7	CREDIT	189	3	0.8	0.0	29.3
8	OBAMA	62	0	0.3	0.0	27.7
9	GREEK	50	0	0.2	0.0	22.4
10	GREECE	46	0	0.2	0.0	20.6
11	CLIMATE	87	2	0.4	0.0	20.2
12	DARFUR	38	0	0.2	0.0	17.0
13	ASIAN	71	2	0.3	0.0	16.5
14	EUROZONE	36	0	0.2	0.0	16.1

FIGURE 2.47 COCA: collocates of *crisis*, 1990–1994 vs 2005–2015

January 2010 to October 2016 and continuing on into the future as well). There is no other large corpus of English that allows users to compare ongoing changes in English to this degree. In addition, the NOW corpus includes texts from the same 20 countries as GloWbE, and we can thus limit searches to particular countries and compare the frequency across countries.

For example, Figure 2.48 shows that the term *gig economy* ("the economic sector consisting of freelancers who take on a series of small jobs, particularly when those jobs are contracted online using a website or app") increased significantly in the latter half of 2015. (Note that because of the format for figures in the NOW corpus online, Figures 2.48–50 were generated in Excel rather than as screenshots from the web, as is the case with all the other figures in this chapter.) Figure 2.48 also shows that this new phrase is (at least to the present) much more common in the US than in other Inner Circle varieties of English.

Sometimes a new word or phrase is limited primarily to the Inner Circle countries, which raises interesting questions about the extent to which concepts do or do not spread across language varieties, even in the Internet Age. For example, consider the data for *precariat* ("people whose lives are precarious because they have little or no job security"). As Figure 2.49 shows, the word originated (at least in the NOW corpus) in late 2010 and really starting increasing in about 2012. But as of late 2016 (when this chapter was written), it still had not extended much beyond the Inner Circle countries (Figure 2.49):

In the case of *locavore* "a person who eats only locally grown food" (Figure 2.50), the frequency spiked three or four years ago and has been decreasing since then. But again, it never really gained a foothold outside of the Inner Circle countries.

Perhaps *precariat* never really extended to South Asia and Africa and other varieties of English, simply because it is almost a "redundant" concept in those areas, where so many have lived on the edge for so long. And perhaps the same is true for *locavore*, either because so much food in South Asia and Africa is already grown locally, or because people there do not have the luxury of demanding locally grown food for those items where it is more economical to produce it elsewhere.

Finally, as with the other BYU corpora, it is possible to look at societal and cultural differences across language varieties by looking at differences in collocates. For example, Figure 2.51 shows adjectives that occur with *marriage* in the US and India in the NOW corpus from 2010 to the current time. Notice the numerous references to same-sex marriage in the US (on the left: *anti-gay, Biblical, constitutional, opposite-sex, straight, pro-gay, same-sex*, etc.). In India, on the other hand (right), there are few if any references to same-sex marriage. Rather the emphasis seems to be on religious issues: *inter-caste, Hindu, inter-religious, inter-, Sikh*, etc.

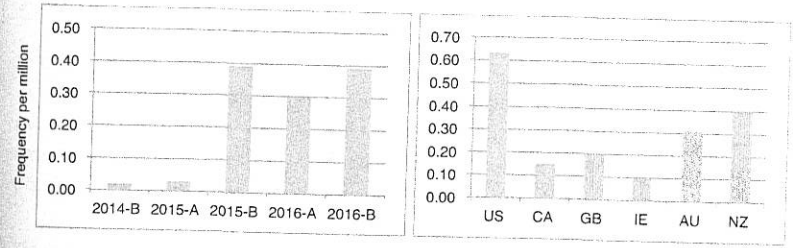


FIGURE 2.48 NOW: *gig economy*

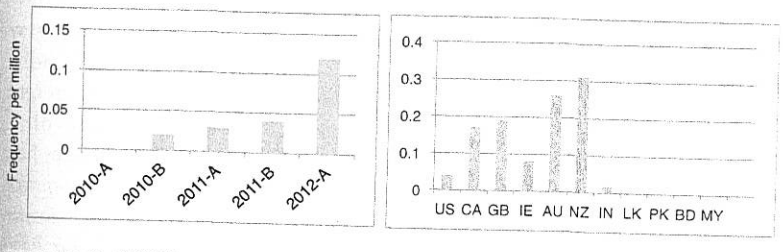


FIGURE 2.49 NOW: *precariat*

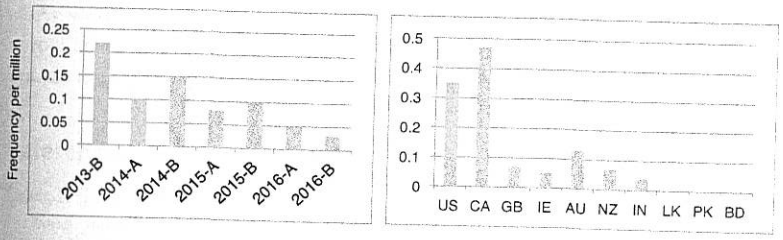


FIGURE 2.50 NOW: *locavore*

The important point is that with the NOW corpus, we can track very recent changes in the language, both across time and space. We can see how a word or phrase (or even meaning and usage, using collocates) starts in one country and then spreads to other countries over time, and we can do this with a very good level of detail (months or even ten-day periods). Such a corpus could potentially revolutionize the way that we study variation and change in English.

SEC 1 (United States): 585,845,861 WORDS

	WORD/PHRASE	TOKENS 1	TOKENS 2	PM 1	PM 2	RATIO
1	ANTI-GAY	108	1	0.2	0.0	44.4
2	BIBLICAL	83	1	0.1	0.0	34.1
3	FUNDAMENTAL	63	1	0.1	0.0	25.9
4	CONSTITUTIONAL	239	5	0.4	0.0	19.6
5	OPPOSITE-SEX	44	1	0.1	0.0	18.1
6	FEDERAL	340	9	0.6	0.0	15.5
7	LICENSED	85	0	0.1	0.0	14.5
8	STRAIGHT	72	0	0.1	0.0	12.3
9	PRO-GAY	70	0	0.1	0.0	11.9
10	MILITARY	81	3	0.1	0.0	11.1
11	SAME-SEX	6516	254	11.1	1.1	10.5

SEC 2 (India): 240,804,113 WORDS

	WORD/PHRASE	TOKENS 2	TOKENS 1	PM 2	PM 1	RATIO
1	ENTERTAINING	141	3	0.6	0.0	114.3
2	INTER-CASTE	341	8	1.4	0.0	103.7
3	HINDU	548	16	2.3	0.0	83.3
4	INTER-RELIGIOUS	68	2	0.3	0.0	82.7
5	INTER	33	1	0.1	0.0	80.3
6	SIKH	29	1	0.1	0.0	70.6
7	ACCUSED	26	1	0.1	0.0	63.3
8	RAMPANT	25	1	0.1	0.0	60.8
9	SUPERB	141	0	0.6	0.0	58.6
10	IRRETRIEVABLE	61	4	0.3	0.0	37.1
11	GRAND	30	2	0.1	0.0	36.5

FIGURE 2.51 NOW: ADJ collocates of *marriage*, US vs India

Conclusion

In this chapter we have seen a number of cases in which these new corpora enable us to investigate societal and cultural differences between varieties of English and over time in ways that would have been quite impossible even six or seven years ago. We have seen that in many cases, this is due to the very large size of the corpora, compared to what was possible 10–15 years ago.

As has been mentioned, however, we should keep in mind that size is not everything. Imagine that we had a 10–20 billion-word corpus (much larger than COCA, GloWbE, or even NOW), but the texts were all taken from the same country at essentially the same time (or that the architecture and interface did not allow us to compare and contrast across the different countries). This is the case with many huge corpora that are based on newspapers or web pages, where it is very easy to create a massive corpus in a short amount of time with relatively little effort. But in this case, the massive "blob" of data would provide very little insight into meaningful and interesting differences in the language.

In other words, corpora that are meaningful and useful for looking at societal and cultural differences need to be large enough to provide information about lower-frequency words, as well as data to compare the collocates of different words. On the other hand, the "textual corpus" needs to be varied enough that it allows users to compare across time or space. And most importantly, it needs to have an underlying architecture and interface that allows users to make meaningful comparisons across the different sections of the corpus (e.g., time periods or countries).

As we have seen, COCA, COHA, GloWbE, NOW, and the Google Books corpus all have both the size and the architecture to provide such data and, thus, unique insight into language variation and change.

References

Baker, P. (2009). The BE06 Corpus of British English and recent language change. *International Journal of Corpus Linguistics, 14*(3), 312–337.

Baker, P. (2010). Will Ms ever be as frequent as Mr? A corpus-based comparison of gendered terms across four diachronic corpora of British English. *Gender and Language, 4*(1), 125–129.

Baker, P. (2011). Times may change but we'll always have money: A corpus driven examination of vocabulary change in four diachronic corpora. *Journal of English Linguistics, 39*, 65–88.

Baron, A., Rayson, P., & Archer, D. (2009). Word frequency and key word statistics in corpus linguistics. *Anglistik, 20*, 41–67.

Davies, M. (2009). The 385+ million word Corpus of Contemporary American English (1990–2008+): Design, architecture, and linguistic insights. *International Journal of Corpus Linguistics, 14*, 159–190.

Davies, M. (2011). The Corpus of Contemporary American English as the first reliable monitor corpus of English. *Literary and Linguistic Computing, 25*, 447–465.

Davies, M. (2012). Expanding horizons in historical linguistics with the 400 million word Corpus of Historical American English. *Corpora*, 7, 121–157.

Davies, M. (2014). Making Google Books n-grams useful for a wide range of research on language change. *International Journal of Corpus Linguistics*, *19*(3), 401–416.

Davies, M. (2015). Corpora: An introduction. In Biber, D. & Reppen, R. (Eds.), *Cambridge handbook of English corpus linguistics* (pp. 11–31). Cambridge: Cambridge University Press.

Davies, M., & Chapman, D. (2016). The effect of representativeness and size in historical corpora: An empirical study of changes in lexical frequency. In Chapman, D., Moore, C., & Wilcox, M. (Eds.), *Studies in the history of the English language VII: Generalizing vs. particularizing methodologies in historical linguistic analysis* (pp. 131–150). Berlin: Mouton De Gruyter.

Davies, M., & Fuchs, R. (2015). Expanding horizons in the study of World Englishes with the 1.9 billion Word Global Web-Based English Corpus (GloWbE). *English World-Wide*, *36*, 1–28.

Hofland, K., & Johansson, S. (1982). *Word frequencies in British and American English*. Bergen/London: Norwegian Computing Centre for the Humanities/Longman.

Leech, G., & Fallon, R. (1992). Computer corpora—What do they tell us about culture? *ICAME Journal*, *16*, 29–50.

Louw, W. (1993). Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies. In Baker, M., Francis, G., & Tognini-Bonelli, E. (Eds.), *Text and technology: In honour of John Sinclair* (pp. 157–176). Amsterdam: John Benjamins.

Michel, J., Kui Shen, Y., Presser Aiden, A., Veres, A., Gray, M., The Google Books Team, . . . Lieberman Aiden, E. (2011). Quantitative analysis of culture using millions of digitized books. *Science*, *331*, 176–182.

Oakes, M., & Farrow, M. (2007). Use of the chi-square test to examine vocabulary differences in English-language corpora representing seven different countries. *Literary and Linguistic Computing*, *22*(1), 85–100.

Sigley, R., & Holmes, J. (2002). Looking at *girls* in corpora of English. *Journal of English Linguistics*, *30*(2), 138–157.

3

USING CORPUS-BASED ANALYSIS TO STUDY REGISTER AND DIALECT VARIATION ON THE SEARCHABLE WEB

Douglas Biber, Jesse Egbert, and Meixiu Zhang

Introduction

The language of the internet is innovative in ways that are noticeable to even a casual observer. Linguists have been eager to describe these innovations, including the special linguistic features associated with internet language (e.g., the use of emoticons, abbreviations, contractions, and acronyms) as well as the 'new' registers found on the internet (e.g., blogs, internet forums, instant messages, and tweets). The book-length treatments by Crystal (2001) and Baron (2008) are good examples of this type.

Similar research issues have also been investigated using corpus analysis. One specialized research approach—multi-dimensional (MD) analysis—has been especially useful for analyzing the linguistic characteristics of internet registers (see, e.g., Biber, Conrad, Reppen, Byrd, & Helt, 2007; Grieve, Speelman, & Geeraerts, 2011; Hardy & Friginal, 2012; Titak & Roberson, 2013). These studies focus on the use of core grammatical features, rather than special linguistic features associated with internet language. Thus, MD studies consider the use of features like pronouns, nouns, verb tenses, and others.

However, most of these MD studies have been similar to other previous research in their focus on the new internet registers. These registers are especially interesting because they are not found in other discourse domains. However, because most previous research has focused on those special registers, we know surprisingly little about the full range of other registers found on the web.[1]

Over the last few years, we have been working on a major project to fill this gap. Rather than focusing on a few special internet registers, we attempted to construct a representative random sample of the entire searchable web: the full range of registers found on the web. MD analysis was used to describe the patterns of linguistic variation among those registers (see Biber & Egbert, 2016). In part, the goal of the present chapter is to describe the corpus-based methods used for these analyses and summarize the major patterns of register variation emerging from the analysis.