CHAPTER 3

Corpus-based Studies of Lexical and Semantic Variation: The Importance of Both Corpus Size and Corpus Design

Mark Davies

Abstract

Small corpora (e.g. 1–5 million words) are often adequate for the study of high-frequency syntactic constructions, but they are typically inadequate for the study of lexical and semantic phenomena, especially for medium and lower-frequency words. "Mega corpora", on the other hand, may have billions of words of easily-obtainable web pages, but they are often just a huge "blob" of texts, which does not have a structure which lends itself to the study of variation. In this paper, we discuss three corpora of English – COCA, COHA, and GloWbE – which are very large (about 100 times the size of comparable corpora like ICE or the Brown family of corpora), but which also have a corpus design, architecture, and interface that lends itself to the in-depth study of variation. With such corpora, we are able to examine genre-based, historical, and dialectal variation in lexis and meaning in ways that would be difficult or impossible with comparable corpora.

Keywords

corpus – variation – historical – dialectal – lexical – semantic – collocates

1 Introduction

English corpus linguistics has a strong tradition of using small, carefully-crafted corpora (1–5 million words) to look at change and variation, and hundreds of insightful studies have been carried out with these corpora. Within the last decade or two, however, there has arisen a different model, which favors the use of very large corpora – some of which are now billions (or tens of billions) of words in size.

 | DOI:10.1163/9789004390652_004

The "small is beautiful" approach tends to focus on those phenomena where there would be enough tokens in a small corpus – such as modals and other auxiliaries. But it often ignores lexical and semantic variation, simply because there isn't enough data for such analyses. The "bigger is better" approach tends, not surprisingly, towards those phenomena where massive amounts of data is required, such as lexical analyses. But this approach also ignores *variation* in lexical and semantic phenomena, because these corpora are often just composed of immense "blobs" of easily-obtainable web pages, and there is no way to sub-divide this "blob" into meaningful sections.

In this paper, I will suggest that we can indeed have large corpora – fifty to one hundred times as big as what was available ten to twenty years ago. But they don't necessarily need to just be a huge, undifferentiated blob of newspapers or web pages. With the right corpus design, architecture, and interface we can compare the many distinctions inherent in these corpora – whether it be between genres, between dialects, or between historical periods.

In this paper I will consider three corpora that are available from the BYU suite of corpora – all of which allow us to carry out in-depth analyses on lexical and semantic variation.[1] COCA (the Corpus of Contemporary American English) contains 520 million words from 1990 to 2015, and it continues to grow by 20 million words each year (Davies 2009). Most importantly, it contains more than 100 million words from *each* of the genres of spoken, fiction, magazine, newspaper, and academic texts. COHA (the Corpus of Historical American English) contains 400 million words from the 1810s to the 2000s (Davies 2012). And GloWbE (the corpus of Global Web-Based English) contains 1.9 billion words from twenty different countries (Davies & Fuchs 2015).

All of these are very large corpora. COCA is more than five times as large as the British National Corpus, and COHA and GloWbE are each 50–100 times as large as comparable corpora like the Brown family of corpora (historical) or the International Corpus of English [ICE] (dialectal). And yet the unique corpus architecture and interface for these corpora allow us to examine variation in ways that is often ignored with even larger corpora, which – as mentioned – force us to analyze the entire corpus as one giant "blob" of data.

In terms of organization, Sections 2 and 3 will examine the issue of size, and show how corpora that are 50–100 times as large as earlier corpora really do allow us to examine many types of variation that could not be studied otherwise.

1 The focus of this paper is lexical and semantic change and variation. For examples of using COCA, COHA, and GloWbE to look at variation and change in syntax, see Davies 2014.

Section 4 then turns to the related issue of comparisons between corpora and also "data granularity" – the fact that once we move beyond very small corpora, we then have enough data to divide the corpus into different sections for meaningful comparisons.

At that point it might seem that size is the only thing that matters. In Section 5, however, we consider some data from very large corpora, which shows that without some meaningful divisions in the data, we have very little sense of exactly what we're looking at. Finally, in Sections 6, I look at what is needed in terms of corpus design, organization, interface, and architecture – in order to "de-blob-ify" the corpora and to carry out meaningful comparisons across the different sections of the corpus.

2 Size Matters: Lexis

As mentioned, small corpora like the Brown family of corpora (cf. Mair 1997) and the International Corpus of English [ICE] (Greenbaum 1996) have been very useful for looking at high frequency syntactic phenomena like modals and other auxiliaries, where even a small one million word corpus might have enough tokens. But when it comes to lexis, it is often a different story. Even for some moderately frequent words, a one million word corpus like the Brown corpus (or LOB, or FROWN, or FLOB) does not provide enough data for useful analyses.

Others who have attempted to use small corpora like these for lexical research have already noticed this limitation. As one of the most active researchers in this field notes (Baker 2011,70):

> [T]he 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 this section, I will provide some new data from the COCA and the Brown corpora, to show just how important size is for looking at lexical phenomena.

When we look at the highest frequency words, the million word Brown corpus is fairly sufficient. If we relax things (compared to Baker, above) and require only 50 tokens of a given word (actually lemma), we find that only 117 of the top 1,000 nouns, verbs, adjectives, and adverbs in COCA (all of which occur at least 40,000 times in COCA) appear less than 50 times in Brown, but these do include frequent words like *star, risk, sister, crime, challenge, lake, break,* and *partner* – to list just a few of the nouns. Things become more problematic for lower frequency words. 546 of the top 2,000 words in COCA (all of which appear 19,000 times or more in COCA) have a frequency of 50 or less in Brown, including *judge, weekend, league, beach, ice, lesson, prison, context* (nouns); *hurt, hide, earn, grab, blow, shut, cook, steal* (verbs); and *healthy, sorry, potential, dangerous, healthy, angry,* and *fast* (adjectives).

With the top 5,000 words in COCA (all of which occur at least 5,600 times in COCA), 3,286 of the words occur less than 50 times in Brown. These include words that would probably still be considered "core" words of English, such as *gap, offer, symptom, layer, prayer, juice, link, potato* (nouns); *kiss, display, bend, kick, evaluate, slide, analyze, whisper* (verbs); *lucky, silent, amazing, sad, violent, glad, pink, round* (adjectives); and *deeply, rarely, strongly, surely* (adverbs). Finally, a full 8,270 of the top 10,000 words in COCA (all of which occur at least 1,800 times in COCA) occur 50 times or less in Brown. These are not just "erudite" words, but rather they include words like *rejection, bargain, praise, rug, foreigner, duration* (nouns); *thrive, rob, dictate, curl, surrender, grip* (verbs); *vague, bizarre, crude, dull, fancy, unclear* (adjectives); and *seldom, abruptly, purely, namely* (adverbs).

Consider that even high school students studying English probably know at least 2,000 words, but that more than 25% of these probably do not occur enough in Brown to carry out meaningful research (at least 50 tokens). Most college-level students would know at least 10,000 words, but the vast majority of these (83%) occur very infrequently in the Brown corpus. As we can see, we need something much larger than a one million word corpus to carry out meaningful lexical analyses of such words.

3 Size Matters: Semantic Phenomena (Via Collocates)

Collocates can provide useful insight into meaning and usage, following Firth's insight that "you shall know a word by the company it keeps" (1957,11). But collocates are very sensitive to corpus size. For example, Table 3.1 shows the number of collocates with different node words in COCA (520 million words), the BNC (100 million words), and the Brown corpus (1 million words).

There are 22 distinct adjectival collocate lemmas of *riddle* (NOUN) that occur three times or more in COCA (span = 1 left / 0 right), e.g. *great, ancient, cosmic.*

TABLE 3.1 Collocates in COCA, BNC, and Brown.

node word	collocates	COCA	BNC	Brown
riddle	adj	22	0	0
nibble	noun	112	13	0
witty	noun	63	4	0
serenely	verb	31	4	0

Note: These words were selected by querying the corpus databases to find words with contrasting frequencies in COCA and the BNC. A Table 3.with the raw frequency for 100,000+ words in COCA and BNC can be found at http://www.wordfrequency.info/100k.asp, which will help in replicating these tests.

There are 112 distinct NOUN collocate lemmas of *nibble* (VERB) that occur three times or more (span = 0L/4R), e.g. *edges, grass, ear*. Turning to collocates of adjectives, we find 63 distinct NOUN collocate lemmas of *witty* (ADJ) with a frequency of three or more (span = 0L/2R), e.g. *dialogue, repartee, banter*. Finally, there are 31 distinct VERB collocate lemmas of the adverb *serenely* that occur three times or more (span = 3L/3R), e.g. *smile, float, gaze*.

Because collocates are so sensitive to size, we find that these numbers decrease dramatically, even in a 100 million word corpus like the BNC. For example, these totals of 22, 112, 63, and 31 in COCA decrease to 0, 13, 4, and 4 (respectively) in the BNC. The situation becomes even more bleak in the Brown corpus. None of the four words have any collocates that occurred with the specified minimal level of frequency.

One might argue that the number of distinct collocates is just a function of the frequency of the node word. In other words, if a node word is ten times as frequent in one corpus than another, then it should have about ten times as many collocates (with a moderate frequency of four or five tokens). But as we will see, the effect of corpus size is often magnified in the case of collocates.

To provide a concrete example, let us consider four different collocate searches in GloWbE (1.9 billion words), COCA (520 million words), and the BNC (100 million words). Table 3.2 shows the number of tokens for four lemmas: BROWSE (verb), STEWARDSHIP (noun), OUTLANDISH (adjective), and RIGHTFULLY (adverb). It also shows the number of collocates with the indicated part of speech, which occur at least five times with the given node word (Note that the collocates are grouped by lemma, and that the collocates span was 4 left / 4 right in all cases).

TABLE 3.2 Frequency of node word and collocates in GloWbE, COCA, and BNC.

		Frequency of node word			# collocates		
Word	Collocates	BNC	COCA	GloWbE	BNC	COCA	GloWbE
stewardship	Adjective	169	1,612	5,179	0	43	123
browse	Noun	166	2,242	24,336	2	193	878
outlandish	Noun	97	842	3,115	0	32	168
rightfully	Verb	69	864	5,279	1	36	202

As we can see, the importance of corpus size for the number of collocates is magnified even more than what we would expect from the token frequency of the node word. For example, the overall frequency of *outlandish* in COCA is only about 8–9 times what it is in the BNC (842 COCA, 97 BNC). But in terms of noun collocates that occur at least five times, the difference is much greater – 32 different collocates in COCA, and none at all in the BNC. Or take the example of *browse*. Because the BNC is limited just to texts from before 1993 (when the Web really began to take off), there are relatively few tokens of *browse* in the BNC – 166 tokens. In COCA, there are about 13–14 times as many tokens of *browse* as in the BNC (2,242 vs 166). But the difference in the number of collocates is much greater – 193 noun collocates that occur at least five times in COCA compared to just 2 in the BNC.

An interesting use of collocates is their role in signaling "semantic prosody" (cf. Louw 1993), in which a word occurs primarily in a negative or positive context. For example, *budge* is nearly always preceded by negation (*it wouldn't budge*), and *cause* takes primarily negative objects (e.g. *death, disease, pain, cancer, problems*). In order to see such patterns, however, we need large corpora. In COCA, there are 1,645 tokens of *budge* and 1,432 different object noun collocates of *cause* that occur at least 10 times each (span = 0L/4R). This decreases to 164 tokens of *budge* and 358 noun collocates of *cause* in the BNC, and just 3 tokens of *budge* and 0 noun collocates of *cause* (occurring ten times or more) in Brown – again, simply not enough for insightful analyses.

The bottom line is that even when a given word has a moderate number of tokens (e.g. 200–400 tokens), that is often not enough when it comes to examining meaningful collocates of that word. A one million word corpus is very rarely sufficient for anything but the highest frequency words, and even a 100 million word corpus like the BNC often provides meager collocates data for moderately frequent words like *riddle, nibble, witty*, or *serenely* (see Table 3.1), which is probably not enough to really say much of interest about the meaning and usage of these words.

4 Comparisons and Data Granularity

In Sections 2 and 3, I examined the issue of size in terms of research on lexical and semantic phenomena, and I focused on the overall size of the corpus. In this section, I will focus on how this problem is compounded once we start making comparisons across small corpora, or (more seriously) within small corpora.

Turning first to comparisons between small corpora, I will take just one example – a comparison of lexical frequency in the 1960s and 1990s – based on the 2 million words of data in Brown (US, 1960s) and FROWN (US, 1990s). As a test case, I will compare this data to the data for the equivalent decades in COHA, which contains 52 million words total for the 1960s and 1990s. Although I will be considering lexical change in this section, the same principles would apply to the comparison of lexis in other small corpora, such as between different dialects of the same language (as with two 1,000,000 word corpora from the ICE corpora).

The lexical phenomena that I will consider are those adjectives that have (at least) doubled in (normalized) frequency from the 1960s to the 1990s. I first created a list of these adjectives from COHA, and I then examined how well the one million word Brown and FROWN corpora (US, 1960s and 1990s) did in providing comparable evidence for this increase in frequency. In other words, in the data below I will be considering adjectives like *overall*, *emerging*, and *motivated*, whose charts in COHA are shown below.

Table 3.3 shows that in COHA there are 15 adjectives that have a combined frequency of between 800–1600 tokens in COHA in the 1960s and 1990s (words such as *overall* (shown above), *amazing, long-term*, and *alternative*) and which have at least doubled in frequency during this time. There are another 127 types with a frequency of between 200–400 tokens in COHA in these two decades (e.g. *emerging* (shown above), *compelling, indoor, preferred,* and *unclear*), and 394 types with a frequency between of between 50 and 100 tokens (e.g. *motivated* (shown above), *first-time, blurry, impaired, viral, obnoxious,* and *luscious*).

Table 3.3 shows that for the 15 COHA adjectives that have at least doubled in frequency and which have a combined token frequency of 800–1600 in COHA

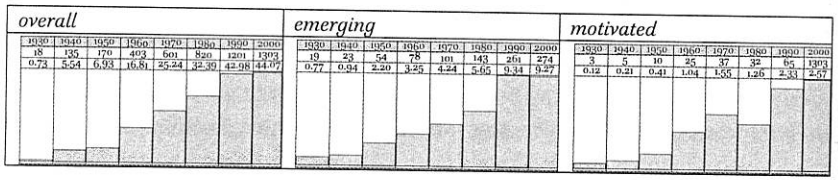


FIGURE 3.1 COHA: Adjectives doubling in frequency, 1960s–1990s.

TABLE 3.3 Evidence for increase in adjective frequency, COHA and Brown family.

COHA: token range			800–1600	200–400	50–100
COHA: # of types			15	127	394
# Brown/Frown tokens	0		0	8	114
	1–9		1	46	264
	>= 10	Support	6	50	12
	>= 10	???	5	15	0
	>= 10	Contradict	3	8	4
Brown/Frown "correct"			0.40	0.39	0.03

in the 1960s and 1990s, all of these occur at least once in Brown/Frown, which is encouraging. One word occurs between 1 and 9 times in Brown/Frown, and the other 14 occur at least 10 times (e.g. 3 tokens in Brown and 7 tokens in Frown), which is perhaps enough to show an increase from the 1960s to the 1990s. Of these 14 adjectives that occur at least 10 times, 6 do show frequency that has doubled from the 1960s–1990s (e.g. Brown 6, Frown 12, which is shown as "Support" (COHA) in Table 3.3 above). Another 5 adjectives show an increase, but less than the doubling in COHA (e.g. 6 Brown and 7 Frown, shown as "? ? ?" above). And in 3 cases, the Brown/Frown data actually shows a decrease from the 1960s to the 1990s (e.g. 7 Brown, 4 Frown; shown as "Contradict" above). Overall, then, 6 of the 15 types (40%) of these high-frequency adjectives in Brown / Frown show the same doubling in frequency that is shown in the robust data (800–1600 tokens) in COHA.

The situation is even less encouraging for the 127 medium-frequency adjectives (token count of 200–400 for the 1960s/1990s in COHA). Of these, 8 do not occur at all in Brown/Frown and 46 occur just 1–9 times, which is probably too few to see an increase. Of those occurring 10 times or more in Brown/Frown, 50 show a doubling, 15 show a smaller increase, and 8 show a decrease.

The situation with lower-frequency words is very poor. Remember, these are adjectives like *first-time, blurry, impaired, viral, obnoxious, luscious*, and *motivated* – less common to be sure, but certainly still the type of adjectives that most speakers of English would be familiar with. Of the 394 types in COHA with a frequency of between 50–100 and which have at least doubled in frequency, 114 of these do not occur at all in Brown/Frown, and another 264 occur

less than 10 times – probably too few to be useful. As a result, Brown/Brown provides evidence for doubling in frequency for only about 3% of all of these lower-frequency adjectives from COHA.

We have seen the problems that arise when we try to compare the frequency of words and phrases in two small corpora. The situation becomes much worse if we try to carry out comparisons across an even larger number of divisions within the corpus. As we have seen above, even with just two data points (1961 and 1991), there is often not enough data to make meaningful comparisons of lexical frequency. But suppose that we wanted to see the frequency of a word or phrase in each year from 1961 to 1991. Here the problem would be greatly magnified – there just wouldn't be enough data from 1964 and 1975 and 1986 and any other years in our 1961–1991 corpus. We would only have about 30,000–35,000 words from each year, and that is far too small to look at thousands of different words.

As a result, there are few options beyond just staying with just two data points – 1961 and 1991. But because there are texts from only every 30 years (due to the small size), any changes that take place in between these years is 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. (Note that in COHA, we have robust data from not only each decade, but also from each year. For example, there are 75,377,000 words of data for the thirty years from 1955 to 1985 – more than 2,400,000 words *each year* for this thirty 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 very small corpora.

As a second example, consider the case of *normalcy*:

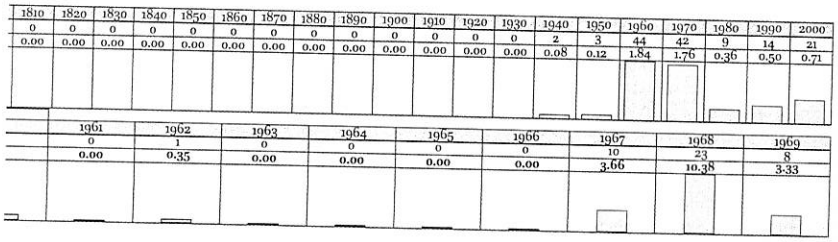


FIGURE 3.2 COHA: *groovy*

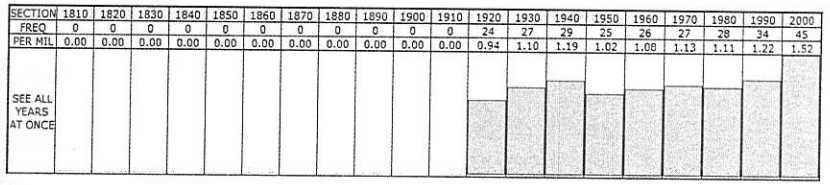


FIGURE 3.3 COHA: *normalcy*

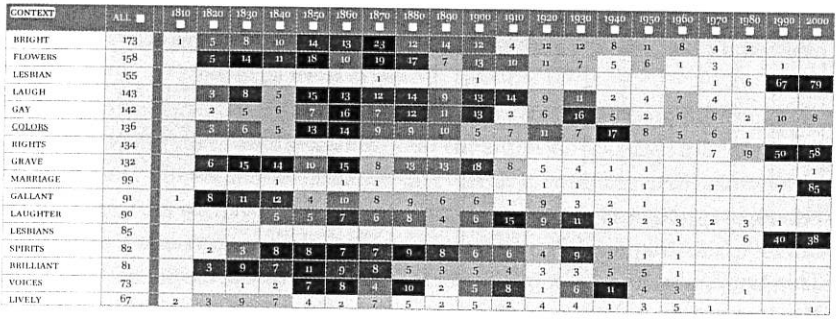
This word was famously "rescued" from obscurity by President Warren G. Harding in 1920, who (according to purists) mistakenly used it instead of the more "correct" *normality*. The word caught on with a public tired of World War I and other foreign involvements, and Harding went on to win the election. But imagine that we only had two small corpora from 1901 and 1931 (as with the 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 if that predated Harding, whether his campaign caused the increase in usage, or whether it was after his time. Corpora that have texts that are spaced decades apart may be adequate for looking at more gradual grammatical change, but they are much more problematic in looking at lexical change, which can occur quite suddenly.

In the previous sections, we considered the issue of size in terms of comparing different sections of the corpus. We have seen that 1–5 million word corpora are usually too small to look at a wide range of lexical phenomena. In this section we will see that the situation is even more serious when it comes to collocates and their use in looking at semantic change. In this case, even corpora that are in the range of 10–50 million words may be too small, once we begin to compare many different sections in the corpus.

As a concrete example, consider the following data from the COHA corpus. Figure 3.4 shows the frequency by decade (1810s–2000s) for the most frequent collocates of the word *gay*, where the word has a Mutual Information score of at least 3.0.

As we can see, the collocates that were most frequent in the 1800s were words like *bright, laugh, colors, gallant, spirits*, and *voices*, which relate to the earlier meaning of "happy". Since at least the 1980s, however, the most frequent collocates are words like *lesbian(s), rights*, and *marriage*, which deal with the more common contemporary meaning that is related to sexual orientation. Clearly the collocates from the 400 million word COHA corpus are useful in terms of examining this semantic shift.

Suppose, however, that we had just a four million word corpus, or in other words about 1/100th the size of COHA. We would expect there to be roughly



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		3	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	6	15	9	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				
BRILLIANT	81		3	9	7	11	9	8	5	3	5	4	3	3	5	5	1				
VOICES	73			1	2	7	8	4	10	2	5	8	1	6	11	4	3		1		
LIVELY	67	2	3	9	7	4	2	7	5	2	5	2	4	4	1	3	5	1			1

FIGURE 3.4 COHA: collocates of *gay*

1/100th the number of tokens for a given collocate as well. For example, rather than 10, 14, 13, 23 tokens of *bright* as a collocate in the 1840s, 1850s, 1860s, and 1870s, we would be lucky to have even one token of *bright* in any of these decades (e.g. 1850s = 14 / 100 = 0.14). This is in spite of the fact that there might be approximately 160–170 tokens of *gay* itself in a four million word corpus (based on the total of 16,438 tokens in COHA – a corpus one hundred times that size).

A final example comes from the GloWbE corpus, and concerns the collocates of *scheme* (noun). As Figure 3.5 shows, the adjectival collocates of *scheme* in British English (right) are quite neutral – *approved, mentoring, eligible*, etc. But in American English, they are much more negative: *evil, fraudulent, nefarious, illegal, (get) rich quick*, etc. This shows that *scheme* has a much more negative connotation in American English, where it usually means "conspiracy, intrigue, ruse".

But the point is that here also, the collocates are very sensitive to size. In GloWbE, the US and GB sections of the corpus are about 770 million words total. If we were comparing these two dialects in the International Corpus of English (ICE), we would have only 2 million words. In other words, we would have about 1/385th the amount of data. If one divides the token count for the collocates above, one can see that very few of the collocates shown in Figure 3.5 would appear in a corpus that size.

In summary, we would agree with Baker (2011) that small 1–4 million word corpora – while useful for high frequency grammatical constructions – are in most cases inadequate for lexical studies, except for perhaps a handful of extremely frequent words. We have provided data primarily from the domain of lexical change, but the same issue would arise anytime we are trying to compare lexical frequency across a large number of small corpora, such as several one million word corpora from 15–20 different countries (as with the ICE corpora).

	SEC 1: 386,809,355 WORDS							SEC 2: 387,615,074 WORDS						
	WORD/PHRASE	TOKENS 1	TOKENS 2	PM 1	PM 2	RATIO		WORD/PHRASE	TOKENS 2	TOKENS 1	PM 2	PM 1	RATIO
1	BLOCKING	42	1	0.11	0.00	42.09	1	APPROVED	92	1	0.24	0.00	91.81
2	URI	80	6	0.21	0.02	13.36	2	OCCUPATIONAL	88	1	0.23	0.00	87.82
3	OFFENSIVE	61	6	0.16	0.02	10.19	3	MENTORING	53	1	0.14	0.00	52.89
4	DEFENSIVE	89	13	0.23	0.03	6.86	4	FLAT	36	1	0.09	0.00	35.93
5	SOCIALIST	20	3	0.05	0.01	6.68	5	ELIGIBLE	31	1	0.08	0.00	30.94
6	ALLEGED	26	5	0.07	0.01	5.21	6	OVERSEAS	31	1	0.08	0.00	30.94
7	EVIL	48	10	0.12	0.03	4.81	7	DEFINED	127	5	0.33	0.01	25.35
8	FRAUDULENT	62	18	0.16	0.05	3.45	8	GENEROUS	50	2	0.13	0.01	24.95
9	NEFARIOUS	27	9	0.07	0.02	3.01	9	LABOUR	25	1	0.06	0.00	24.95

FIGURE 3.5 GloWbE: collocates of *scheme* in US and GB.

We have also seen that the problem becomes even more serious when it comes to collocates, where even 10–50 million words of data might not be enough.

5 Size Alone is not Enough

To this point, I have made the argument that size is very important when we are examining lexical and semantic variation. In this section, however, I will show that size alone is not enough. This is important to understand, because it is increasingly common to find corpora that are composed of billions or even tens of billions of words of data, from easily obtainable newspapers or other sites on the Web. (For example, virtually all of the corpora over 100 million words in size in Sketch Engine are based exclusively on web pages.)

But the question is – how representative are web pages, in terms of the full range of variation in the language? Does the data from a web-only corpus contain the same range of variation that we would find in a carefully designed corpus like the BNC or COCA, where there are texts from spoken, fiction, magazines, newspapers, and academic? And if not, which of these traditional genres are web pages most similar to?

To answer these questions, we should first consider some data from COCA, which shows variation across genres for a number of syntactic and morphological phenomena. Figure 3.6 shows how much more common *–al* adjectives are in academic (adjectives that are at least ten letters in length, e.g. *international, additional, psychological, institutional*). Figures 7–10 show a number of grammatical phenomena where there are significant variations between genres: preposition stranding with *to* (e.g. *the man I was talking to*), the *get* passive (e.g. *John got fired from his job*), *real* instead of *really* before adjectives (e.g. *he was real sick*), and the quotative *like* (e.g. *and I'm like, what's the problem?*).

When we compare these morphological and syntactic phenomena in a web-only corpus (like GloWbE) to a more genre-balanced corpus (like COCA), the situation becomes very confusing. For example, the normalized frequency

SPOKEN	FICTION	MAGAZINE	NEWSPAPER	ACADEMIC
142906	54665	234038	193870	642274
1,495.38	604.50	2,449.15	2,113.77	7,052.83

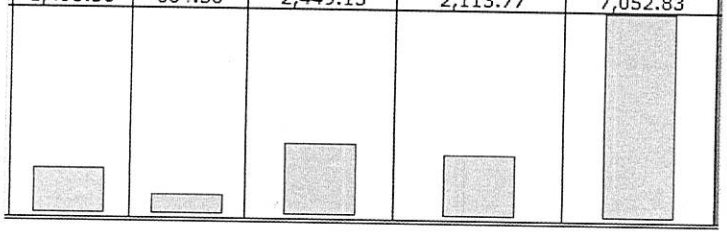


FIGURE 3.6 *al.[j*]

SPOKEN	FICTION	MAGAZINE	NEWSPAPER	ACADEMIC
6086	3423	1760	1623	1519
63.68	37.85	18.42	17.70	16.68

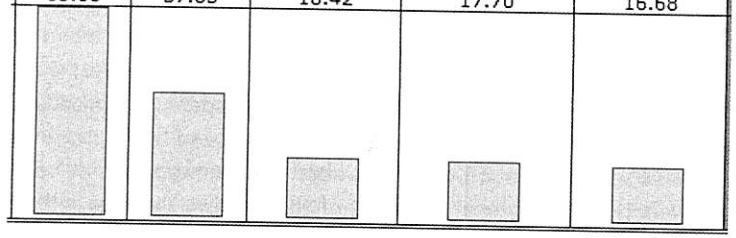


FIGURE 3.7 [vv*] to ,

of *–al* adjectives is 2,244 per million words in GloWbE-US (the 385 million words from the US in GloWbE), which places it between COCA magazine and newspapers (see Figure 3.6 above). But the normalized frequency of the *get* passive (239.4) is most similar to spoken (Figure 3.8), the frequency of preposition stranding (31.1) places it between fiction and magazines (Figure 3.7), and the frequency of the "quotative *like*" (2.5) is most similar to news (Figure 3.10). And strangely enough, the normalized frequency of *real* + ADJ in Figure 3.9 (0.41) is most like COCA Academic.

As we can see, depending on the particular phenomena that we are studying, the web corpora are "all over the map" in terms of which of the "traditional" genres they best represent. As a result, it would be difficult to know ahead of time – for any particular phenomena – how representative of "standard" genres (like spoken or fiction or academic) a web-only corpus would be. Likewise, it would probably be unwise to carry out studies on the language of these large web-only corpora, and then assume that we have mapped out the range of variation that we would find in a traditional, genre-balanced corpus.

SPOKEN	FICTION	MAGAZINE	NEWSPAPER	ACADEMIC
23643	16169	14120	13262	3218
247.40	178.80	147.76	144.60	35.34

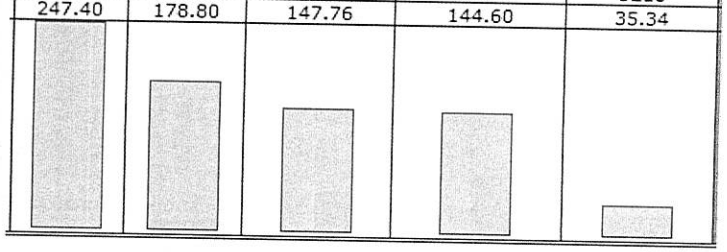


FIGURE 3.8 *get* passive

SPOKEN	FICTION	MAGAZINE	NEWSPAPER	ACADEMIC
512	390	146	264	21
5.36	4.31	1.53	2.88	0.23

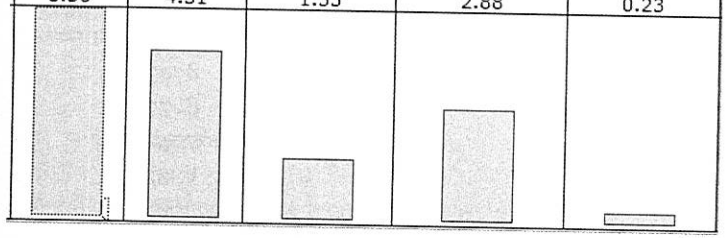


FIGURE 3.9 [be] real [j*] [y*]

Because the focus of this paper is on lexical and semantic variation, let us consider some additional phenomena that compare the lexis from web-only corpora to more genre-balanced corpora. In this case, we compare word frequency in COCA and the BNC to the 1.9 billion word GloWbE corpus, which again is based (like most Sketch Engine corpora) just on web pages. In this comparison, we will see how many words in a 100,000 word list of English[2] (which is based on COCA and BNC) have roughly the same normalized frequency in GloWbE as in different genres of COCA and the BNC. For example, there are 13,386 words (from among the 100,000 total in the list) whose normalized frequency in COCA Newspapers is roughly the same as that of GloWbE – i.e. the ratio is between 0.8 and 1.2. (In other words, if the frequency of a given word is 40 tokens per million words in GloWbE, then it would be between 32 and 48 tokens per million words in COCA Newspapers.)

2 http://www.wordfrequency.info. See also Davies and Gardner (2010).

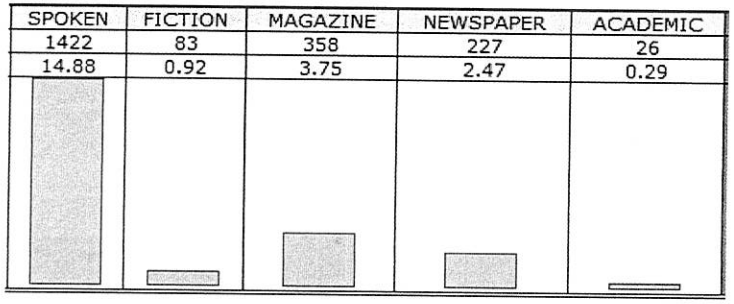


FIGURE 3.10 quotative *like*: [c*] [p*] [be] like ,|'

TABLE 3.4 Similarity of lexis in web-based GloWbE and genres in COCA and BNC.

COCA	# words	BNC	# words
Newspaper	13836	Magazine	8743
Magazine	13349	Newspaper	8677
Academic	11828	Academic	7032
Spoken	10793	Fiction	6335
Fiction	8804	Spoken	4667

As can be seen, at least in terms of lexis, the web-only corpus is most like newspapers and magazines, but "web" lexis does a much poorer job of representing the lexis of the academic genre, or especially fiction and spoken. This may be why at times even very large web-only corpora do not improve significantly on the data from a well-balanced corpus (like COCA or the BNC). Even a corpus like the 11.2 billion word Sketch Engine enTenTen12 corpus provides only minimally better data for words that are most common in genres like fiction.

For example, COCA has 112 noun collocates of the verb *nibble* that occur at least three times (see Table 3.1 above), but the 11.2 billion word Sketch Engine enTenTen12 corpus (which is about 20 times as large as COCA) only had 96 such collocates. Likewise, COCA has 31 verb collocates of *serenely* that occur at least three times, while enTenTen12 improve this only slightly to 36 different collocates. The very large Sketch Engine corpora are great when we are looking at lexis that is most like the lexis from newspapers and magazines, but it is only marginally better (or perhaps even worse) for other genres like fiction.

To look at this a different way, consider Figure 3.11, which shows verbs that are much more common in fiction (left) than newspapers (right) in COCA. Imagine that we had a corpus composed only of web-based newspapers

SEC 1: 90,429,400 WORDS

	WORD/PHRASE	TOKENS 1	TOKENS 2	PM 1	PM 2	RATIO
1	FUCK	984	0	10.88	0.00	1,088.14
2	PISS	389	8	4.30	0.09	49.32
3	WHIMPER	90	2	1.00	0.02	45.64
4	HISS	84	3	0.93	0.03	28.40
5	SHRIEK	114	5	1.26	0.05	23.12
6	FIDGET	65	3	0.72	0.03	21.98
7	SNORE	86	4	0.95	0.04	21.81
8	GLARE	170	8	1.88	0.09	21.55
9	THROB	101	5	1.12	0.05	20.49
10	SOB	238	12	2.63	0.13	20.12
11	UNDRESS	213	11	2.36	0.12	19.64
12	TREMBLE	485	25	5.36	0.28	18.92
13	PEE	465	26	5.14	0.28	18.14
14	CRUMBLE	52	3	0.58	0.03	17.58
15	UNZIP	67	4	0.74	0.04	16.99

SEC 2: 91,717,452 WORDS

	WORD/PHRASE	TOKENS 2	TOKENS 1	PM 2	PM 1	RATIO
1	RE-SIGN	201	1	2.19	0.01	198.18
2	REFINANCE	214	3	2.33	0.03	70.33
3	REZONE	62	0	0.68	0.00	67.66
4	BLITZ	53	0	0.58	0.00	57.79
5	REDEVELOP	92	2	1.00	0.02	45.35
6	TELEVISE	89	2	0.97	0.02	43.88
7	OUTPERFORM	79	2	0.86	0.02	38.95
8	OVERCOOK	71	2	0.77	0.02	35.00
9	RESTRUCTURE	278	8	3.03	0.09	34.26
10	PRIVATIZE	169	5	1.84	0.06	33.33
11	DEREGULATE	61	2	0.67	0.02	30.07
12	DIVERSIFY	263	9	2.87	0.10	28.81
13	RETOOL	82	3	0.89	0.03	26.95
14	LEGALIZE	147	6	1.60	0.07	24.16
15	TOUT	138	6	1.50	0.07	22.68

FIGURE 3.11 COCA: Verbs in fiction and newspapers

(which are very easy to obtain). In this case, words like those on the left would be almost completely absent in the corpus, while those on the right would be massively over-represented.

In summary, the web-based corpus only provide data on a very narrow "slice" of the language, and there is often no way to generalize the results from that corpus to the language as a whole.

6 Creating Variation-aware Corpora

As we have seen, corpus size is crucial to most lexical and semantic studies. But size is not enough. If all we have is a huge one billion or ten billion word "blob" of web texts, then we are very limited in terms of understanding vital aspects of variation in the language. In this section, I will consider ways in which we can design corpora that allow us to have the best of both worlds – size, plus the ability to meaningfully analyze variation. I will focus on COCA, COHA, and GloWbE from the BYU family of corpora, and I will focus on both aspects of the "textual corpus" (the texts in the corpus) as well as the corpus architecture and interface.

In the case of the BYU corpora, several of the textual corpora were designed "from the ground up" to facilitate the study of variation in English. As has been mentioned, COCA has at least 100 million words each of spoken, fiction, magazine, newspaper, and academic texts, and the relative frequency of these genres (and sub-genres such as Magazine-Sports, Magazine-Children, Academic-Legal, or Academic-Engineering) stays roughly the same from year to year. I have argued elsewhere that COCA is the only large corpus of English that continues to be updated and which maintains the same genre (and sub-genre) balance from year to year (see Davies 2011).

Turning to COHA, we find that it too was designed from the ground up for the study of historical variation in English. As mentioned, it also maintains roughly the same genre balance from decade to decade (fiction, newspaper,

magazine, and non-fiction books). For example, the percentage of fiction in each decade is always between 48–52% of the total for that decade. In addition, the balance by sub-genre (e.g. Non-Fiction History, Non-Fiction Domestic Arts, Non-Fiction Religion) also stays roughly the same from decade to decade (see Davies 2012).

Finally, GloWbE was also designed from the ground up to look at dialectal variation in English. We used Google's country identification for the categorization of the texts, and this identification uses advanced heuristics including IP, country of origin of the links to the website, and country of origin of the visitors to the site.

Of course the BYU corpora are not completely unique in the sense of being the only large corpora that are designed to look at variation. It is true that COHA is the only large (> 30–40 million words) structured corpus of historical English and that GloWbE is the only large, structured corpus from different countries. But there are a handful of other large corpora that focus on genre-based variation, in addition to the 520 million word COCA corpus. The most well known is undoubtedly the British National Corpus (100 million words), as well as the 2.5 billion word Oxford English Corpus (OEC). The OEC was designed to include texts from many different domains – mostly from the web pages, but supplemented by other copyrighted texts from the OUP. Unfortunately, the OEC is generally available only to researchers at Oxford University Press, although other researchers who can demonstrate a strong need may apply for access.

Finally, we should remember that it is possible to have a "variation-aware" corpus, even when the corpus was not initially designed that way. For example, researchers of "Web as Corpus" and "Web for Corpus" often create large corpora of web-based texts (simply taking any and all web pages) and then attempt to categorize the texts after the fact, according to domain (sports, recipes, news, personal blogs, etc). Unfortunately, such post-hoc categorization is both very time-intensive and very expensive, and the corpora tend to be quite small. Perhaps the best example of a genre-categorized corpus of web texts is the CORE corpus (Corpus of Online Registers of English; http://corpus.byu.edu/core). The creators of CORE have used Mechanical Turk to obtain judgments from hundreds of thousands of people about the genre of nearly 50,000 texts (50 million words), and they have also subjected these texts to sophisticated analysis of linguistic features in order to determine register (see Biber et al. 2015a, 2015b).

Creating a "textual corpus" that is composed of texts from many different decades, dialects, or genres is only half of the battle. There also needs to be

SEC 1: 129,755,748 WORDS | **SEC 2: 106,640,094 WORDS**

	WORD/PHRASE	TOKENS 1	TOKENS 2	PM 1	PM 2	RATIO		WORD/PHRASE	TOKENS 2	TOKENS 1	PM 2	PM 1	RATIO
1	PAUPERISM	301	1	2.32	0.01	247.38	1	RACISM	994	0	9.32	0.00	932.11
2	ROMANISM	154	0	1.19	0.00	118.68	2	TOURISM	756	0	7.09	0.00	708.93
3	HEATHENISM	165	2	1.27	0.02	67.80	3	ACTIVISM	404	0	3.79	0.00	378.84
4	PROPAGANDISM	58	0	0.45	0.00	44.70	4	ANTI-SEMITISM	302	1	2.83	0.01	367.46
5	CONGREGATIONALISM	106	2	0.82	0.02	43.56	5	METABOLISM	348	0	3.26	0.00	326.33
6	BIMETALLISM	48	0	0.37	0.00	36.99	6	MARXISM	342	0	3.21	0.00	320.70
7	GALVANISM	44	0	0.34	0.00	33.91	7	FASCISM	278	0	2.61	0.00	260.69
8	SENSUALISM	41	1	0.32	0.01	33.70	8	RELATIVISM	253	0	2.37	0.00	237.25
9	DEMAGOGISM	39	1	0.30	0.01	32.05	9	FEMINISM	250	0	2.34	0.00	234.43

FIGURE 3.12 COHA: **ism* words, 1870s–1890s and 1970s–2000s

SEC 1: 1,239,817,686 WORDS | **SEC 2: 233,866,709 WORDS**

	WORD/PHRASE	TOKENS 1	TOKENS 2	PM 1	PM 2	RATIO		WORD/PHRASE	TOKENS 2	TOKENS 1	PM 2	PM 1	RATIO
1	CONSERVATIVISM	189	1	0.15	0.00	35.65	1	QADIANISM	100	1	0.43	0.00	530.14
2	EUROSCEPTICISM	183	1	0.15	0.00	34.52	2	NAXALISM	145	2	0.62	0.00	384.35
3	PRESENTEEISM	166	1	0.13	0.00	31.31	3	MARXISM-LENINISM-MAOISM	127	3	0.54	0.00	224.43
4	NIMBYISM	154	1	0.12	0.00	29.05	4	CASTEISM	249	12	1.06	0.01	110.00
5	ANABAPTISM	146	1	0.12	0.00	27.54	5	TALIBANISM	32	2	0.14	0.00	84.82
6	LABOURISM	143	1	0.12	0.00	26.97	6	VAISNAVISM	47	3	0.20	0.00	83.06
7	PIETISM	142	1	0.11	0.00	26.79	7	GANDHISM	45	3	0.19	0.00	79.52
8	MONETARISM	252	2	0.20	0.01	23.77	8	SHAIVISM	56	4	0.24	0.00	74.22
9	PRESBYTERIANISM	120	1	0.10	0.00	22.64	9	VAISHNAVISM	83	6	0.35	0.00	73.34

FIGURE 3.13 GloWbE: **ism* words: "Inner Circle" and South Asian countries

some way to efficiently compare across the different sections of the corpus. As has been mentioned, many very large corpora (billions or tens of billions of words) do not have the analysis of variation as one of their goals, and so there is no easy way to make these comparisons. In the case of the BYU corpora, the architecture and interface is designed from the ground up to facilitate such research.

All of the BYU corpora are stored as relational databases – a structure that lends itself to very powerful and efficient comparisons across the corpora. For example, suppose that we wanted to compare **ism* words in the two periods of the 1870s–1890s and the 1980s–2000s. There is a "sources" metadata Table 3.in the database that includes information on each of the 100,000+ texts in the corpus – date, genre, author, etc. If we want to compare **ism* words in the two time periods, we simply select the two periods in the search interface. After the user submits the queries, we use advanced SQL commands to store (in turn) the **ism* words from the 1870s–1890s and the 1980s–2000s in two temporary tables. Further SQL commands are then used to find the words that are common in one period but not in the other, and then they are displayed in tables like those in Figure 3.12 (1870s–1890s on the left, 1970s–2000s on the right).

Of course this is not limited just to COHA. Similar queries can be carried out in GloWbE, or COCA, or any of the other BYU corpora. For example, Figure 3.13 shows the **ism* words that are more frequent in the "inner circle" varieties (on the left; US, UK, Canada, Ireland, Australia, New Zealand) compared to

the South Asian varieties on the right (India, Sri Lanka, Bangladesh, Pakistan). Note the more secular words in the Inner Circle varieties, and the focus on religious words in South Asia.

Or consider Figure 3.14 from COCA, which finds adjectives in Academic-Medicine (left) compared to Academic (overall), or verbs in Magazine-Religion (right) compared to Magazines (overall):

In addition to these comparisons, users can also see the frequency of all matching words in all decades. For example, Figure 3.15 shows the frequency of all **ism* in COHA by decade.

This shows the higher frequency of words like *patriotism, despotism*, and *heroism* in the 1800s, the high frequency of *communism* in the mid-1900s, and the recent increase of *mechanism, journalism*, and *terrorism*.

The point is that because of the way that the data is stored in the relational database, these searches are very fast – even for a 400 million word corpus like COHA. A query like the comparison of **ism* words in the 1870s–1890s and the 1980s–2000s takes only 1.0–1.5 seconds. Even a comparison of collocates (like *gay*; see Figure 3.14 above) typically takes only 2–3 seconds – to find all occurrences of a given node word, find nearby collocates, store them in temporary tables, and compare them across different sections of the corpus, and then display them by section (decade, dialect, or genre).

	WORD/PHRASE	TOKENS 1	TOKENS 2
1	PAROTID	389	1
2	TONSILLAR	185	0
3	PARATHYROID	132	1
4	OTOTOXIC	116	0
5	BRANCHIAL	229	2
6	OTOLOGIC	110	1
7	ONCOCYTIC	94	0
8	OSSICULAR	89	0
9	PAPILLARY	92	1
10	STAPEDIAL	87	0

	WORD/PHRASE	TOKENS 1	TOKENS 2
1	SIN	44	18
2	MINISTER	76	57
3	ORDAIN	36	28
4	REPENT	33	41
5	PRAY	483	603
6	BAPTIZE	22	28
7	PREACH	141	183
8	BLESS	73	101
9	WORSHIP	110	157
10	AFFIRM	92	133

FIGURE 3.14 COCA: ADJ in Academic-Medicine and Magazine-Religion

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	356	343	341	370	408	689	684	706	1016	1212	1123	860	771	922	974	975	895	659	541
PATRIOTISM	4924	25	147	430	359	332	406	359	307	357	320	481	290	221	179	114	116	135	170	125	113
COMMUNISM	4790				6	14	4	57	101	26	15	34	168	441	496	1450	940	362	279	320	147
MECHANISM	4538		20	71	141	96	306	110	98	151	285	270	381	358	291	376	329	338	267	492	343
SOCIALISM	3540				17	55	10	148	92	181	213	440	269	397	331	279	295	311	304	136	56
ORGANISM	3423		3	2	69	36	82	174	228	179	321	289	374	273	256	343	191	214	136	126	132
JOURNALISM	2031		3	1	57	17	34	87	50	96	108	150	196	131	201	177	182	207	202	283	346
CAPITALISM	2509								2	7	11	65	335	270	247	217	208	259	453	439	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	103	54	47	46	39	27	41	22	7
BAPTISM	2106		48	40	217	531	286	131	162	110	100	61	38	44	49	47	54	49	42	43	79
HEROISM	2031	18	61	71	115	169	193	115	179	131	108	179	111	83	90	83	61	69	95	60	80
REALISM	2012		5		1	13	24	49	122	120	122	114	198	147	139	237	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	1847			1	2	3	7	9	2	6	12	19	57	55	50	30	62	220	387	135	746

FIGURE 3.15 COHA: **ism* words by decade

Sometimes there are too many possible categories for the texts in a corpus. For example, in the BYU Wikipedia Corpus (http://corpus.byu.edu/wiki) there are 4.4 million texts on a wide range of categories – science, technology, history, companies, sports, pop culture, and so on. And unfortunately, there is no single way to categorize all of the texts. For example, the page on Bill Gates might be categorized as biography or technology, and the page for the London Eye might be categorized according to geography (London) or purpose (attractions).

In 2015, we developed the functionality to create "virtual corpora" within any of the BYU corpora. Users can create these corpora "on the fly" using either words within the texts, or in the title of the text, or any combination of these. In just a matter of 3–4 seconds, users can create a virtual corpus of the top 1000 (or 10,000 or more) texts dealing with any topic – investments, molecular biology, basketball, Buddhism, or anything else – and pointers to all of these texts are stored for their account on the corpus server. The users can then limit their searches (specific words or phrases, substrings, collocates, etc) to any of these virtual corpora; they can compare the frequency across their different virtual corpora; and (perhaps most useful) they can create lists of keywords from each of these virtual corpora.

Perhaps the most straightforward use of these virtual corpora is for corpora like Wikipedia, where there is no single way to categorize all of the texts. But even for the other corpora, these virtual corpora can be quite useful. For example, in the 1.6 billion word Hansard Corpus (www.hansard-corpus.org), users can create customized corpora from the 7.6 million speeches (1803–2005) in the British Parliament by speaker, date, and topic (e.g. speeches by Winston Churchill from 1939–1945, which mention the word *Germany*). Or in COCA, they could create a virtual corpus of all texts (from among the 220,000+ texts in the corpus) that mention *Monica Lewinsky* and which appear in the New York Times or the Washington Post in 1998 or 1999.

The bottom line is that it is possible to "de-blob-ify" corpora, and to carry out meaningful comparisons across sections of the corpora. We might do this as we assign sections as we create the corpus (genres and sub-genres in COCA, decades and years in COHA, and countries in GloWbE), or post-hoc via linguistic features (as with the CORE corpus), or via user-defined "virtual corpora" (as with Wikipedia, Hansard, and now any of the BYU corpora).

7 Conclusion

As we have seen, we need two things to carry out meaningful lexical and semantic comparisons in corpora. First, the corpora need to be quite large. Small

corpora like the Brown family of corpora (4 million words total) or even the ICE corpora (~15 million words total) may not be large enough for meaningful comparisons of lexis. And as we have seen, size is even more important for analysis of meaning (via collocates), where sometimes even 100 million words is not enough. We have also seen that even larger corpora are needed once we begin to compare across different sections of the corpora, such as the 20 decades of COHA or the 20 countries in GloWbE.

But size alone is not sufficient. As we have seen, huge corpora containing billions (or tens of billions) of words are often just immense "blobs" of data, which don't provide much insight into important variation in the language. Without the right corpora and corpus architecture and interface, we wouldn't know that *muffled* and *frowned* are more common in fiction and that *validity* and *correlate* are more common in formal academic writing; that *bestow* and *swell* (adjective) sound old-fashioned and that *morph, freak out,* and *throw someone under the bus* are quite recent; or that *fortnight* isn't used much in the US, and that *banjaxed* is found almost exclusively in Ireland.

And these are just the simplest of comparisons. With the right corpora and corpus architectures and interfaces, we could find, for example, all verbs that are more common in sports reporting than in newspapers overall, all adjectives that are more common in the 1920s–1940s than the 1950s–1970s, or all words that are more frequent in Australia or in South Asia than in other varieties of English. Using "virtual corpora", we can find keywords in texts related to Buddhism or biology, or in speeches by Churchill in the World War II years.

Finally, we can use collocates to compare meaning and usage across time periods, genres, and dialects. We can compare the collocates of *chain* in fiction and academic, collocates of *woman* in the 1800s and the late 1900s, or collocates of *wife* in the "Inner Circle" compared to the "Outer Circle" varieties of English.

In summary, lexical and semantic comparisons often require very large corpora, and they require corpora that are designed from the ground up to look at variation, and which have a useful architecture and interface. With such corpora, we can gain insight into lexical and semantic phenomena in ways that help us to understand the full range of variation in the language.

References

Baker, Paul. 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.

Biber, Douglas, Jesse Egbert & Mark Davies. 2015a. Exploring the composition of the web: A corpus-based taxonomy of web registers. *Corpora* 10(1). 11–45.

Biber, Douglas, Jesse Egbert & Mark Davies. 2015b. Developing a bottom-up, user-based method of web register classification. *Journal of the Association for Information Science and Technology* (JASIST) 66. 1817–1831.

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

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

Davies, Mark. 2014. Examining syntactic variation in English: The importance of corpus design and corpus size. *English Language and Linguistics* 19(3). 1–35.

Davies, Mark & Dee Gardner. 2010. *A frequency dictionary of American English: Word sketches, collocates, and thematic lists*. London: Routledge.

Davies, Mark & Robert Fuchs. 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.

Firth, John R. 1957. A synopsis of linguistic theory 1930–1955. In *Studies in linguistic analysis*, 1–32. Oxford: Blackwell.

Greenbaum, Sidney (ed.). 1996. *Comparing English worldwide: The International Corpus of English*. Oxford: Oxford University Press.

Louw, Bill. 1993. Irony in the text or insincerity in the writer?: The diagnostic potential of semantic prosodies. In Mona Baker, Gill Francis, John McHardy Sinclair & Elena Tognini-Bonelli, *Text and technology: In honour of John Sinclair,* 157–76. Philadelphia, PA & Amsterdam: John Benjamins.

Mair, Christian. 1997. Parallel corpora: A real-time approach to the study of language change in progress. In Magnus Ljung (ed.), *Corpus-based studies in English*, 195–209. Amsterdam: Rodopi.