Powerful (yet simple) comparisons of a wide range of phenomena in British and American English

Mark Davies, Brigham Young University

1   Introduction

There has certainly been no lack of research on corpus-based comparisons of British and American English (cf. Biber *et al*. 1999; Algeo 2006; Mair 2006; Leech *et al*. 2009; Rohdenburg *et al*. 2010, to name just a few). As useful as these studies have been, for a number of reasons it has still been very difficult for most corpus linguists to carry out a wide range of research on the two varieties of English. The three main problems have been the following:

First, the corpora may not be publicly-available. For example, Biber *et al*. (1999) used the Longman corpus, which has been made available only to a few researchers, as they author Longman materials. Another example is Algeo (2006), who uses the Cambridge International Corpus (CIC, now called the *Cambridge English Corpus*), which is again limited to 'in-house' use. Finally, sometimes the corpus is publicly-available, but at a price that is so high that it renders it inaccessible for most researchers, such as the Bank of English (now called *WordBanks Online*), which costs more than $1,100 per year.

Second, the corpora may not be 'balanced' between a wide range of genres. For example, many of the studies in Rohdenburg *et al*. (2010) are based primarily on a handful of British and American newspapers. Mair (2006) is likewise based primarily on newspapers in the two dialects, as well as some web-based data. But there is little or no spoken material, or fiction, or popular magazines, or academic. To be fair, however, we should remember that prior to about five years ago, there were no large, publicly-available, genre-balanced corpora of American English that could be used for studies of the two dialects. Therefore, these books were simply using the best materials that they could 'cobble together', even if they only included a few sources, in just one or two genres.

Third, there are some studies that are based on corpora that are fairly well-balanced for genre, and which are fairly available to a wide range of researchers, but which use corpora that are quite small, and which therefore can be used to study only very high frequency phenomena. Many of the articles in the Leech *et*

al. (2009) volume fit into this category. They are based on the Brown family of corpora (Brown, LOB, Frown, and FLOB), which are each only about one million words in size. An investigation of the chapters in this book shows that more than half deal with just very high frequency phenomena like modals, progressives, passives, and high-frequency phenomena related to the noun phrase. So as insightful as these studies might be for high frequency syntactic studies (and these corpora have been of great value for studying certain types of syntactic change, during the past few decades), these corpora do not have enough data to be used for many more lexically-oriented studies (see Baker 2011; Davies forthcoming a, forthcoming b) or medium- and low-frequency syntactic constructions (see Davies 2012a, 2012b, and 2012c).

2 Interfaces and architectures for comparing British and American corpora

Researchers of British English have of course had access to the large (100 million word), genre-balanced, publicly-available British National Corpus (BNC) since the early 1990s. But as we have seen, there was not anything comparable for American English until relatively recently. In 2008, however, the Corpus of Contemporary American English (COCA) was released (see Davies 2009). It is large (starting at 385 million words, and currently [2013] at 450 million words), it is genre-balanced (divided evenly between spoken, fiction, popular magazines, newspapers, and academic writing), and it is freely available at http://corpus.byu.edu/coca (see Davies 2009, 2011, as well as books that are based in large part on COCA, such as Anderson and Corbett 2009; Lindquist 2010; Reppen 2010). In other words, it solves the three problems listed above.

The combination of the BNC and COCA is made even more attractive by the fact that at the http://corpus.byu.edu/ interface, users have been able for 2–3 years now to 'seamlessly' move back and forth between the two corpora. In other words, they could do a search in one corpus and then with just one click of the mouse, they could re-do the same search in the other corpus.

To take a trivial example, they could see the frequency of *pants* in the different genres of COCA (Figure 1) and the BNC (Figure 2). This shows (not surprisingly) that *pants* is more than five times as frequent (per million words) in COCA (overall) than in the BNC (29.9 in COCA, 5.6 in the BNC); for clearer versions of the ensuing figures in this article, please see the on-line version of this publication.

SECTION	ALL	SPOKEN	FICTION	MAGAZINE	NEWSPAPER	ACADEMIC	1990-1994	1995-1999	2000-2004	2005-2009	2010-2012
FREQ	13897	1336	6420	4026	1807	308	2477	3143	3297	3341	1639
PER MIL	29.95	13.98	70.99	42.13	19.70	3.38	23.82	30.38	32.03	32.74	31.57
SEE ALL SUB-SECTIONS AT ONCE											
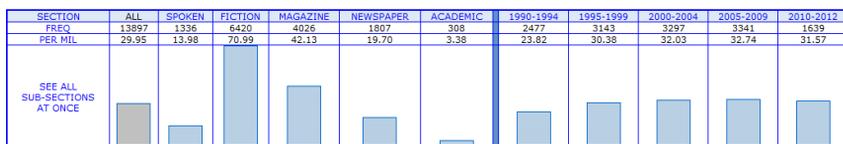
Figure 1: Frequency of the word pants *in COCA*

SECTION	ALL	SPOKEN	FICTION	MAGAZINE	NEWSPAPER	NON-ACAD	ACADEMIC	MISC
FREQ	536	56	213	93	45	38	16	75
PER MIL	5.61	5.62	13.39	12.81	4.30	2.30	1.04	3.60
SEE ALL SUB-SECTIONS AT ONCE								
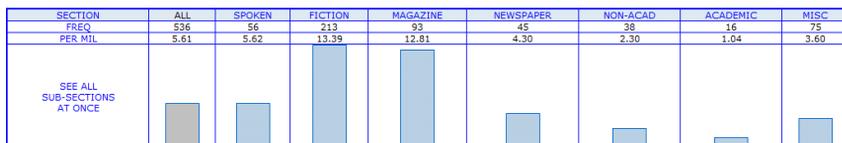
Figure 2: Frequency of the word pants *in the BNC*

Researchers could also search for *ism words in COCA and then in the BNC, and they would see the items given in Table 1 (note that these are separate searches, and not compared to each other in the corpus interface):

*Table 1: *ism words in the BNC and COCA*

BNC		COCA	
criticism	4573	*criticism*	16854
mechanism	2881	*terrorism*	15265
capitalism	1863	*mechanism*	8020
socialism	1612	*racism*	7864
tourism	1410	*journalism*	6468
racism	1061	*tourism*	6325
nationalism	996	*capitalism*	5707
communism	920	*optimism*	4862
realism	875	*nationalism*	4715
optimism	830	*communism*	4612

To take a final example, they could compare the collocates of a given word, such as noun and adjective collocates of *scheme* (again, these are two separate searches in the BNC and COCA interfaces; see Table 2):

Table 2: Collocates of *scheme* in the BNC and COCA

BNC		COCA	
pension	621	*color*	547
new	574	*things*	450
training	432	*ponzi*	273
other	263	*new*	252
government	243	*grand*	220
colour	193	*broader*	165
pilot	190	*other*	159
share	189	*part*	147
national	186	*pyramid*	141
classification	173	*classification*	139

The problem with this approach, however, is that users had to collect the data for the BNC, copy it to some other program (such as an Excel spreadsheet), and then re-do the search in COCA, and then copy it to the other program. They would then have to compare the two sets of data, taking into account the size of the corpus (or the genre in each corpus), and therefore the relative (normalized) frequencies in the two corpora. This could take several minutes for even a relatively simple search.

In 2013, however, I created an improved interface for these corpora, which allows for much quicker and easier (and more powerful) comparisons between corpora such as the BNC and COCA. With this new architecture and interface, with one click (on the button 'SIDE BY SIDE') it is now possible for users to compare any set of data in the BNC and COCA – whether lexical, morphological, syntactic, or semantic (via collocates). With such searches, all of the normalizations for corpus size and the calculations of ratios between the two corpora are done automatically via the corpus interface, and the data is displayed side-by-side in the corpus interface.

With this new interface, researchers can now carry out powerful comparisons of the data in COCA and the BNC. This data in turn allows them to easily and quickly compare data for a wide range of phenomena in the two dialects in ways that were not possible previously, and this allows them to move beyond the small set of phenomena that have been studied previously. In the sections that follow, I will provide a few examples of such comparisons – for differences between lexis, phraseology, morphology, syntax, word meaning, and discourse – between the two dialects.

3   Lexical and phraseological differences

While the real power of the new corpus interface is found in comparing in the two dialects all matching words or collocates for a given string, for specific words and phrases it is also somewhat more useful than the two separate displays shown in Figures 1–2 above. As Baker (2011) and Davies (forthcoming a, forthcoming b) have noted, the Brown family of corpora (Brown, LOB, Frown, FLOB) have few if any tokens for medium and lower frequency words, which makes it difficult to see differences between British and American English. A few examples of this are *trash can* (0 tokens in the four corpora) / *rubbish bin* (0 tokens); *vacuum* (v) (0 tokens) / *hoover* (v) (0 tokens); *freeway* (18 tokens) / *motorway* (12 tokens); or *diaper* (8 tokens) / *nappy* (2 tokens).

COCA and the BNC are really the only large corpora of British and American English, which are genre-balanced and which are publicly-available, that can be used to compare such words. To do so, users simply input a word or phrase (such as *diaper* as a noun) into COCA, for example, and then click on 'SIDE BY SIDE / BNC' in the corpus interface to see the same search in the BNC. The 3,300 tokens (compare 10 in Brown+) show that *diaper* is more than 14 times as frequent in American English than in British English (7.09 tokens per million words in COCA, and 0.45 in the BNC); see Figure 3:


SECTION	ALL	SPOKEN	FICTION	MAGAZINE	NEWSPAPER	ACADEMIC	1990-1994	1995-1999	2000-2004	2005-2009	2010-2012
FREQ	3274	493	862	1309	489	121	661	785	684	757	387
PER MIL	7.06	5.16	9.53	13.70	5.33	1.33	6.36	7.59	6.64	7.42	7.46

BNC (COMPARE ⌄) [HELP...]

SECTION	ALL	SPOKEN	FICTION	MAGAZINE	NEWSPAPER	NON-ACAD	ACADEMIC	MISC
FREQ	43	1	9	1	1	17		13
PER MIL	0.45	0.10	0.57	0.14	0.10	1.03	0.07	0.62
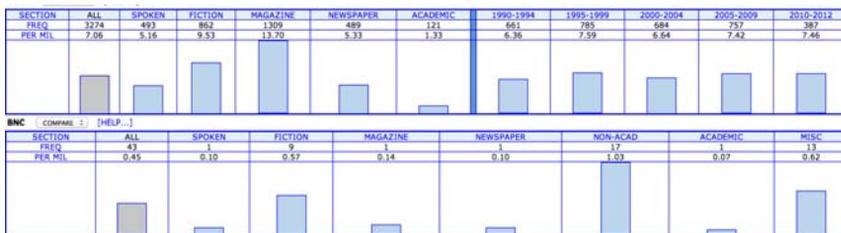
Figure 3: 'Side by side' frequency of the word diaper *in COCA / BNC*

Another example is *can* of [NOUN] in COCA (which would probably be *tin of* [NOUN] in the BNC), which is nearly 50 percent more frequent in COCA than the BNC (4.3 per million vs 2.8); see Figure 4:

COCA	COMPARE	[HELP...]								
SECTION	ALL	SPOKEN	FICTION	MAGAZINE	NEWSPAPER	ACADEMIC	1990-1994	1995-1999	2000-2004	2005-2009	2010-2012
FREQ	1996	220	982	397	339	58	441	481	477	375	222
PER MIL	4.30	2.30	10.86	4.15	3.70	0.64	4.24	4.65	4.63	3.67	4.28

BNC	COMPARE	[HELP...]						
SECTION	ALL	SPOKEN	FICTION	MAGAZINE	NEWSPAPER	NON-ACAD	ACADEMIC	MISC
FREQ	268	32	76	29	52	21	5	53
PER MIL	2.80	3.21	4.78	3.99	4.97	1.27	0.33	2.54

*Figure 4: 'Side by side' frequency of the phrase [*can of NOUNs] in COCA / BNC*

Examples like *diaper* and *can of* [NOUN] are somewhat trivial, in the sense that most native speakers of British or American English already know that there is a difference between the two dialects. But these two simple examples show that COCA and the BNC do provide the expected data for these words, and the same is true for hundreds of less obvious words as well.

Of course, it is possible to search for much more than just specific, individual words. For example, we could search for all of the following idioms with *head* in the phrase (see Table 3). In doing so, we would see that the words on the left have roughly the same frequency in the BNC and COCA (see Figure 5), whereas those on the right are quite a bit more common in American English (COCA; see Figure 6). (Note that ~ = {*his, her, my*}, etc.).

Table 3: Idioms with *head* in COCA the BNC

Approx. same in COCA and BNC	More frequent in COCA (American)
head over heels in love	*head* (v) *up*
head-on	*head* (v) *toward(s)*
price on ~ head	*head* (v) *back to*
head for the hills	*head* (v) *out*
head and shoulders above	*in over ~ head*
talk over ~ head	*(hit the) nail on the head*
talk ~ head off	*head ~ off at the pass*
two heads are better (than one)	*cooler heads (+ prevail)*
use ~ head	*go head-to-head*
make ~ head spin	*head start*
put ~ heads together	*heads or tails*
bury ~ head (in the sand)	*talking head*
from head to toe	*head game*
have a head for (something)	*head rush* (n)
hanging over ~ head	*head trip*
off the top of ~ head	(*like a*) *deer in the headlights*

SECTION	ALL	SPOKEN	FICTION	MAGAZINE	NEWSPAPER	ACADEMIC	1990-1994	1995-1999	2000-2004	2005-2009	2010-2012
FREQ	69	10	41	11	5	2	10	14	15	20	10
PER MIL	0.15	0.10	0.45	0.12	0.05	0.02	0.10	0.14	0.15	0.20	0.19

BNC COMPARE [HELP...]

SECTION	ALL	SPOKEN	FICTION	MAGAZINE	NEWSPAPER	NON-ACAD	ACADEMIC	MISC
FREQ	12	0	9	0	0	1	0	2
PER MIL	0.13	0.00	0.57	0.00	0.00	0.06	0.00	0.10
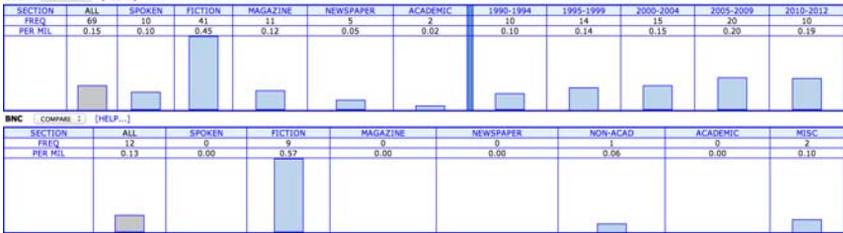
Figure 5: 'Side by side' frequencies for make ~ head spin (COCA 0.15, BNC 0.13)

COCA COMPARE [HELP...]

SECTION	ALL	SPOKEN	FICTION	MAGAZINE	NEWSPAPER	ACADEMIC	1990-1994	1995-1999	2000-2004	2005-2009	2010-2012
FREQ	198	36	71	55	34	2	35	54	41	46	22
PER MIL	0.43	0.38	0.79	0.58	0.37	0.02	0.34	0.52	0.40	0.45	0.42

BNC COMPARE [HELP...]

SECTION	ALL	SPOKEN	FICTION	MAGAZINE	NEWSPAPER	NON-ACAD	ACADEMIC	MISC
FREQ	1	0	0	0	1	0	0	0
PER MIL	0.01	0.00	0.00	0.00	0.10	0.00	0.00	0.00
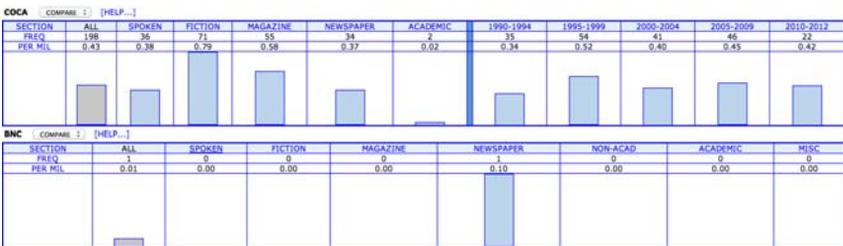
Figure 6: 'Side by side' frequencies for in over ~ head (COCA 0.43, BNC 0.01)

The real power of the dialect comparisons, however, is the ability to move beyond specific words and phrases, and to compare all matching words in the two dialects. For example, the following figure shows *head words (after a/an) in American and British English. This search is done by simply entering 'a|an *head' in the COCA search form, and then clicking on '[SIDE BY SIDE: BNC' to re-do the search in the BNC and compare the results, side-by-side. It shows that *trailhead*, for example, occurs 44 times in COCA but 0 times in the BNC, and that other words like *crackhead, knucklehead, knucklehead, bobblehead, pothead*, and *hothead* are fairly common in COCA, but almost completely absent in the BNC (see Figure 7):

COCA: 450,000,000 WORDS						BNC: 100,000,000 WORDS					
	WORD/PHRASE	1: COCA	2: BNC	PM 1	PM 2	RATIO	WORD/PHRASE	2: BNC	1: COCA	PM 2	PM 1	RATIO
1	A BEACHHEAD	77	1	0.17	0.01	17.11	A ROUNDHEAD	3	1	0.03	0.00	13.50
2	A TRAILHEAD	44	0	0.10	0.00	9.78	A FATHEAD	2	1	0.02	0.00	9.00
3	A BUCKHEAD	43	0	0.10	0.00	9.56	A DICKHEAD	10	10	0.10	0.02	4.50
4	A THUNDERHEAD	34	0	0.08	0.00	7.56	A RED-HEAD	2	2	0.02	0.00	4.50
5	A STEELHEAD	31	0	0.07	0.00	6.89	A BACK-OF-THE-HEAD	1	1	0.01	0.00	4.50
6	A BULKHEAD	88	3	0.20	0.03	6.52	AN AXE-HEAD	1	1	0.01	0.00	4.50
7	A CRACKHEAD	28	0	0.06	0.00	6.22	A BEACH-HEAD	1	1	0.01	0.00	4.50
8	A KNUCKLEHEAD	27	0	0.06	0.00	6.00	A BIGHEAD	1	1	0.01	0.00	4.50
9	A REDHEAD	149	6	0.33	0.06	5.52	A BIG-HEAD	1	1	0.01	0.00	4.50
10	A WELLHEAD	24	1	0.05	0.01	5.33	A DUMBHEAD	1	1	0.01	0.00	4.50
11	AN AIRHEAD	21	1	0.05	0.01	4.67	A PAN-HEAD	1	1	0.01	0.00	4.50
12	A COPPERHEAD	18	0	0.04	0.00	4.00	A PIN-HEAD	1	1	0.01	0.00	4.50
13	A BOBBLEHEAD	17	0	0.04	0.00	3.78	A SNAKEHEAD	1	1	0.01	0.00	4.50
14	A POTHEAD	17	0	0.04	0.00	3.78	A WOLFSHEAD	1	1	0.01	0.00	4.50
15	A HOTHEAD	50	3	0.11	0.03	3.70	A DEATH'S-HEAD	3	4	0.03	0.01	3.38

Figure 7: 'Side by side' display: a(n) * head

Another example are the words ending in **ism* in COCA and the BNC. Again, in just 1–2 seconds, we have an interesting comparison like the following. We see that the **ism* words that are more common in American English (COCA) include *counterterrorism, bioterrorism, volunteerism*, and *Islamism*, whereas those in British English (BNC) include *normativism, Labourism, Owenism*, and *Toryism* (see Figure 8):

COCA: 450,000,000 WORDS						BNC: 100,000,000 WORDS					
	WORD/PHRASE	1: COCA	2: BNC	PM 1	PM 2	RATIO	WORD/PHRASE	2: BNC	1: COCA	PM 2	PM 1	RATIO
1	COUNTERTERRORISM	1354	1	3.01	0.01	300.89	TSARISM	34	1	0.34	0.00	153.00
2	BIOTERRORISM	426	0	0.95	0.00	94.67	EQUESTRIANISM	29	1	0.29	0.00	130.50
3	VOLUNTEERISM	403	1	0.90	0.01	89.56	ISOMERISM	27	1	0.27	0.00	121.50
4	NEOLIBERALISM	306	1	0.68	0.01	68.00	NORMATIVISM	27	1	0.27	0.00	121.50
5	ISLAMISM	286	1	0.64	0.01	63.56	LABOURISM	72	3	0.72	0.01	108.00
6	PENTECOSTALISM	204	1	0.45	0.01	45.33	OWENISM	17	1	0.17	0.00	76.50
7	ANTITERRORISM	189	0	0.42	0.00	42.00	TORYISM	47	3	0.47	0.01	70.50
8	ECOTOURISM	411	3	0.91	0.03	30.44	VOLUNTARYISM	13	1	0.13	0.00	58.50
9	BOOSTERISM	123	0	0.27	0.00	27.33	JACOBITISM	64	6	0.64	0.01	48.00
10	GLOBALISM	197	2	0.44	0.02	21.89	LAMARCKISM	55	7	0.55	0.02	35.36
11	ASTERISM	98	0	0.22	0.00	21.78	FABIANISM	21	3	0.21	0.01	31.50
12	AUTISM	2909	30	6.46	0.30	21.55	ORALISM	20	3	0.20	0.01	30.00
13	WAHHABISM	92	0	0.20	0.00	20.44	RETRIBUTIVISM	20	3	0.20	0.01	30.00
14	ORIGINALISM	83	0	0.18	0.00	18.44	TEETOTALISM	13	2	0.13	0.00	29.25

Figure 8: 'Side by side' display: *ism *words*

A final example – this time dealing with differences in phraseology in the two dialects – are phrasal verbs occurring with the adverbial particle *up* in British and American English. As we can see, *ratchet up, suit up, crank up, boot up*, and *muster up* are much more common in American English, whereas British English prefers *nip up, stump up, phone up, cash up, tot up*, and *pluck up* (all of which sound quite 'exotic' to this speaker of American English); see Figure 9:

		COCA: 450,000,000 WORDS							BNC: 100,000,000 WORDS					
	WORD/PHRASE	1: COCA	2: BNC	PM 1	PM 2	RATIO		WORD/PHRASE	2: BNC	1: COCA	PM 2	PM 1	RATIO
1	RATCHET UP	143	1	0.32	0.01	31.78	1	NIP UP	15	1	0.15	0.00	67.50
2	SUIT UP	124	1	0.28	0.01	27.56	2	STUMP UP	28	4	0.28	0.01	31.50
3	FESS UP	94	0	0.21	0.00	20.89	3	PLANT UP	6	1	0.06	0.00	27.00
4	HIKE UP	76	0	0.17	0.00	16.89	4	TOP UP	72	12	0.72	0.03	27.00
5	CRANK UP	118	2	0.26	0.02	13.11	5	PHONE UP	50	10	0.50	0.02	22.50
6	BOOT UP	53	0	0.12	0.00	11.78	6	CASH UP	5	1	0.05	0.00	22.50
7	MUSTER UP	52	1	0.12	0.01	11.56	7	TOT UP	4	1	0.04	0.00	18.00
8	CHEW UP	51	1	0.11	0.01	11.33	8	PLUCK UP	25	7	0.25	0.02	16.07
9	PUNCH UP	49	0	0.11	0.00	10.89	9	ALLOCATE UP	3	1	0.03	0.00	13.50
10	LISTEN UP	48	0	0.11	0.00	10.67	10	BUGGER UP	3	1	0.03	0.00	13.50
11	BUST UP	46	0	0.10	0.00	10.22	11	COCK UP	3	1	0.03	0.00	13.50
12	SCARE UP	46	0	0.10	0.00	10.22	12	FEATURE UP	3	1	0.03	0.00	13.50
13	HIT UP	45	1	0.10	0.01	10.00	13	MANAGE UP	3	1	0.03	0.00	13.50
14	CUDDLE UP	44	1	0.10	0.01	9.78	14	MUG UP	3	1	0.03	0.00	13.50
15	BUCKLE UP	43	1	0.10	0.01	9.56	15	STRENGTHEN UP	3	1	0.03	0.00	13.50

Figure 9: *'Side by side' display:* verb + up

4   *Syntactic and morphological differences*

As I noted previously, the vast majority of studies on differences between British and American English have focused on high-frequency syntactic differences (such as modals and other auxiliaries), because that is the type of data that one can get from small 1–4 million word corpora. In this section, I will provide some evidence for these high frequency constructions as well. But because we are dealing with corpora whose combined frequency is more than 100 times that of the smaller corpora, we can look at many other constructions as well. I will begin by looking at morphological differences, and then move to syntax.

One of the obvious morphological differences between British and American English is the past participle of *get*, where *gotten* is frequent only in American English. The BNC/COCA data show this contrast quite clearly, where *gotten* is 60–70 times as frequent in COCA (38.7 per million in COCA, 0.6 in the BNC); see Figure 10:

COCA	COMPARE ⇕	[HELP...]									
SECTION	ALL	SPOKEN	FICTION	MAGAZINE	NEWSPAPER	ACADEMIC	1990-1994	1995-1999	2000-2004	2005-2009	2010-2012
FREQ	17966	5745	5739	2720	3197	565	3770	3847	3859	4185	2305
PER MIL	38.72	60.12	63.46	28.46	34.86	6.20	36.25	37.19	37.49	41.01	44.40

BNC	COMPARE ⇕	[HELP...]							
SECTION	ALL	SPOKEN	FICTION	MAGAZINE	NEWSPAPER	NON-ACAD	ACADEMIC	MISC
FREQ	61	10	14	10	1	13	2	11
PER MIL	0.64	1.00	0.88	1.38	0.10	0.79	0.13	0.53

Figure 10: *'Side by side' display:* [have] + gotten

Another example is the past participle of *prove*: *proved* or *proven*. Figure 11 shows that *proven* is much more common in American English (5.6 per million in COCA, 0.9 in the BNC):
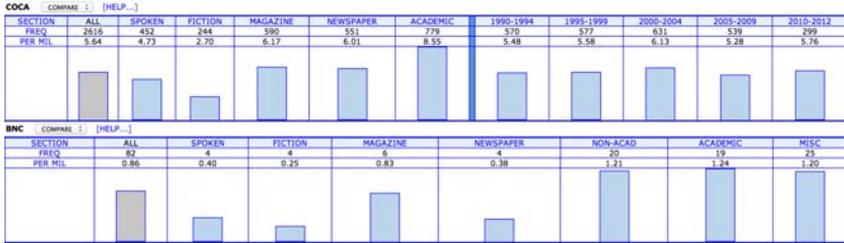

Figure 11: [have] + proven

Table 4 shows the frequency of both *proved* and *proven* in the two dialects, and a Chi-square calculation shows that this difference is significant at p < .000001:

Table 4: [have] + proved / proven in COCA/BNC

	[*have*] *proved*	[*have*] *proven*	*% proven*
COCA	3001	2616	47%
BNC	1669	82	4%

The data from the four Brown corpora is given in Table 5. Obviously, there are many fewer tokens, and while the difference between the two dialects is still significant, it is only significant at p < .0035:

Table 5: [*have*] + *proved* / *proven* in Brown family of corpora

	[*have*] *proved*	[*have*] *proven*	*% proven*
Am: Brown/Frown	15	6	25%
Br: LOB/FLOB	47	1	2%

A final example is the simple past form of *sneak*: *sneaked* or *snuck*. Figure 12 shows that [pronoun] + *snuck* (e.g. <u>*he snuck* in the back window</u>) is much more common in American English (0.72 per million in COCA, 0.01 in the BNC):
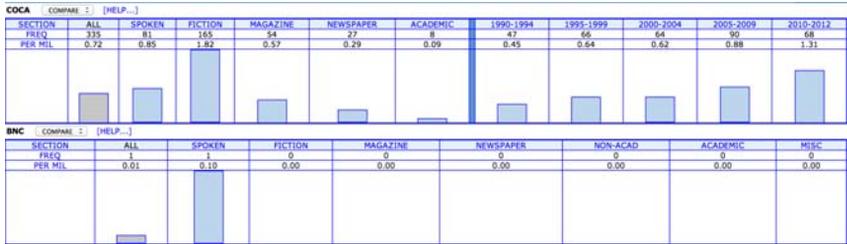
Figure 12: PRON + snuck *in COCA/BNC*

Table 6 shows the frequency of both *sneaked* and *snuck* in the two dialects, and a Chi-square calculation shows that this difference is significant at p < .000001:

Table 6: PRON + *snuck* in COCA/BNC

	[*pronoun*] *sneaked*	[*pronoun*] *snuck*	*% snuck*
COCA	305	335	52%
BNC	43	1	2%

The data from the four Brown corpora is extremely sparse (only two tokens total for either construction in any corpus), and as a result, the Chi-square calculation shows no significant difference between the two dialects (see Table 7). This is typical of most medium- and low-frequency morphological phenomena with these smaller corpora – there just are not enough tokens to compare the two dialects.

Table 7: PRON + *sneaked* / *snuck* in Brown family of corpora

	[*pronoun*] *sneaked*	[*pronoun*] *snuck*	*% snuck*
Am: Brown/Frown	1	1	50%
Br: LOB/FLOB	0	0	

Turning to syntactic differences, we find in the BNC and COCA the same type of data for highly-frequent constructions that others have already found in much smaller corpora like Brown+. For example, the data given in Figure 13 shows that *must* + lexical verb (e.g. *they must admit that…*) is more common in British than American English (245 tokens per million in the BNC; 177 in COCA).

Note also that in COCA, it is least common in the most informal dialect (Spoken) and the most common in the most formal dialect (Academic), and that its frequency is decreasing in each five-year period since the early 1990s.
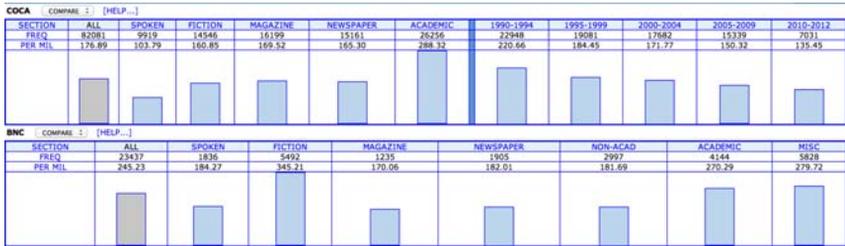
Figure 13: must + *lexical verb in COCA/BNC*

Let us now turn to a somewhat less frequent construction – post-verbal negation with the verb *need* (e.g. *they need not concern you*). The Brown family of corpora have 45 tokens in the US corpora (Brown and Frown) and 69 in the British corpora (LOB and FLOB). In COCA and the BNC there are nearly 6,000 tokens. They show that the construction is more than twice as common in the BNC, and that in COCA, the construction is associated mainly with the more formal genres (e.g. eight times as common in Academic as Spoken), and that the construction is decreasing in frequency over time; see Figure 14:
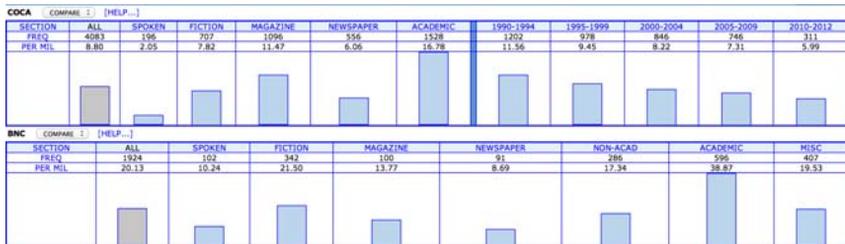
Figure 14: need + NEG + VERB in COCA/BNC

Turning to an even less frequent construction, we find that there are only 31 tokens of the [*end up* V-ing] construction in the Brown corpora (e.g. *they ended up paying too much*). Even with this small amount of data, however, it looks like

the construction is more common in the US (21 vs 10 tokens) and that it is increasing from the 1960s to the 1990s (3 vs 28 tokens).

Of course, the data from COCA and the BNC is much more robust. There are nearly 13,000 tokens, and they show that the [*end up* V-ing] construction is more than twice as common in the US as in the UK, that in the US (but not UK) it is the most common in the informal genres, and that it is increasing in frequency in each five-year period in the US (of course there is no such diachronic data for the BNC, since it is not designed to be used as a historical or monitor corpus); see Figure 15.

Remember, however, that the BNC is limited to texts from a generation ago (the 1980s and early 1990s), whereas COCA is added to year-by-year (and thus currently included texts through 2012). If the construction is increasing over time, then any more recent corpus (e.g. COCA, which alone includes texts from the last 20 years) will have more tokens. This is an issue that I will return to in the conclusion to this paper.

COCA	COMPARE	[HELP...]									
SECTION	ALL	SPOKEN	FICTION	MAGAZINE	NEWSPAPER	ACADEMIC	1990-1994	1995-1999	2000-2004	2005-2009	2010-2012
FREQ	11887	3893	1597	2994	2536	867	2056	2492	2598	3054	1687
PER MIL	25.62	40.74	17.66	31.33	27.65	9.52	19.77	24.09	25.24	29.93	32.50

BNC	COMPARE	[HELP...]						
SECTION	ALL	SPOKEN	FICTION	MAGAZINE	NEWSPAPER	NON-ACAD	ACADEMIC	MISC
FREQ	1047	173	155	173	178	117	57	194
PER MIL	10.96	17.36	9.74	23.82	17.01	7.09	3.72	9.31

*Figure 15: [*end up V-ing] in COCA/BNC*

Let us now examine an even more interesting and recent construction: the 'quotative *like*' construction, e.g. *and I'm like, I don't want it* (the search string is *[c*] [p*] [be] like ,|'*). The data given in Figure 16 from COCA and the BNC show that it is nearly ten times as frequent in COCA (4.6 per million COCA and 0.5 in the BNC). In addition, it is most common in the more informal genres in COCA, and it is increasing in each five-year period in COCA.

COCA	COMPARE	[HELP...]									
SECTION	ALL	SPOKEN	FICTION	MAGAZINE	NEWSPAPER	ACADEMIC	1990-1994	1995-1999	2000-2004	2005-2009	2010-2012
FREQ	2116	1422	83	358	227	26	128	336	436	711	505
PER MIL	4.56	14.88	0.92	3.75	2.47	0.29	1.23	3.25	4.24	6.97	9.73

BNC	COMPARE	[HELP...]						
SECTION	ALL	SPOKEN	FICTION	MAGAZINE	NEWSPAPER	NON-ACAD	ACADEMIC	MISC
FREQ	48	44	0	3	0	0	1	0
PER MIL	0.50	4.42	0.00	0.41	0.00	0.00	0.07	0.00

Figure 16: Quotative like *construction in COCA/BNC*

Again, however, we have to worry about the fact that we are 'comparing apples and oranges' to some degree as we use COCA (continually updated; current as of 2012) and the BNC (now a generation old). Any construction that is increasing over time has the potential to appear more common in American English by the mere fact that COCA is a more modern corpus.

Interestingly, if we look at a corpus whose texts in British and American English are completely contemporaneous, this huge gap with the 'quotative *like*' construction is much smaller. For example, the 1.9 billion word GloWbE corpus (web pages from 20 English-speaking countries, 2012–2013; see http://corpus2.byu.edu/glowbe) shows that 'quotative *like*' is only slightly more frequent in American than British English (2.5 per million in US and 1.9 per million in GB (Great Britain)), and the KWIC lines following that provide examples of the construction from the GB portion of the corpus; see Figures 17 and 18:

SECTION	ALL	US	CA	GB	IE	AU	NZ	IN	LK	PK	BD	SG	MY	PH	HK	ZA	NG	GH	KE	TZ	JM
FREQ	3114	966	309	740	126	203	79	69	13	35	27	74	43	55	34	48	84	62	36	40	71
PER MIL	1.56	2.50	2.29	1.91	1.25	1.37	0.97	0.72	0.28	0.68	0.68	1.72	1.03	1.27	0.84	1.06	1.97	1.60	0.88	1.14	1.79
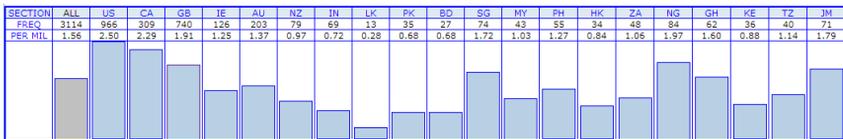
Figure 17: Quotative like *construction in GloWbE*

1	GB G	guru.bafta.org	A B C	can attest to that. They were very excited about it, **so I was like ,** ' Holy shit, now I have to do a play!' because
2	GB G	guru.bafta.org	A B C	' Well sure, it's all downhill that way.' **And I was like ,** ' Oh, okay. He's got a repertoire, and doesn't
3	GB G	femalefirst.co.uk	A B C	who I love, and I was showing him my stomach, **and he was like ,** ' Errr...' and I was like,' I'll do a
4	GB G	femalefirst.co.uk	A B C	my stomach, and he was like,' Errr...' **and I was like ,** ' I'll do a handstand for you!' And I kicked him
5	GB G	eurogamer.net	A B C	, but over here is the area of effect fear. Things **where it 's like ,** in some dungeons this might be really good, but in this encounter this
6	GB G	blokely.com	A B C	' # '' But touring is all I've ever done **since I was like ,** 18, 20 years old. I couldn't do anything else I used
7	GB G	eurogamer.net	A B C	light the lamps with whale-oil tanks, " says Smith. " **And we were like ,** I know, what if this was a combatant who was made to burn
8	GB G	dailymail.co.uk	A B C	money. # ' The public seem to be picking on her **and I 'm like ,** ' Carry on, keep on picking on Dorries, she deserves everything she
9	GB G	uk.answers.yahoo.com	A B C	UK? I had my induction day at college on the 17th **and it was like ,** meh. I didn't know anyone there and I thought everyone else was
10	GB G	dailymail.co.uk	A B C	just said, do you actually wan na be my girlfriend? **And I was like ,** yeah. I really liked him. I knew Gemma then for about two

Figure 18: Concordance lines for 'quotative like*' in British portion of GloWbE*

48

Of course, not all of the dialectal differences in syntax are due to the fact that COCA is a generation more recent than the BNC. For example, consider the data with the two competing constructions [*all the* NOUNs] and [*all of the* NOUNs] (e.g. *all* (*of*) *the reasons*). Figure 19 shows the frequency of [*all **of the*** NOUNs] in COCA and the BNC, and we see that it is much more common in COCA. Notice, however, the genre patterning in COCA, where the construction is not limited to primarily formal or informal genres, and note also that the frequency is fairly static over time. Nevertheless, the construction is more than three times as frequent in COCA as in the BNC (21.9 tokens per million in COCA, 6.8 in the BNC).

*Figure 19. [*all of the *NOUN] in COCA and BNC*

If we compare the frequency of the two constructions in COCA and the BNC, we see that the construction with *of* is much more common in American English, and this difference is significant (using Chi square) at p < .00001; see Table 8:

Table 8: [*all (of) the* NOUN] in COCA and BNC

	all the [nn2]	all **of** the [nn2]	% all **of**
COCA	58,345	10,170	14.8%
BNC	15,116	647	4.1%

Again, however, the data from the much smaller Brown family of corpora is much less helpful. In this case, the results from the two dialects are virtually the same, and (using Chi square) there is no significant difference between the two dialects; see Table 9:

Table 9: [*all (of) the* NOUN] in the Brown family of corpora

	all the [nn2]	all of the [nn2]	% all **of**
Am: Brown/Frown	295	20	6.3%
Br: LOB/FLOB	293	22	6.9%

5   *Semantic differences*

Suppose we want to compare the meaning and usage of a word in American (COCA) and British (BNC) English. How would we do this? One possibility is to use standard concordance / KWIC lines, and to look for nearby words. For example, Figure 20 gives a few lines for *scheme* in COCA, and Figure 21 gives a few lines from the BNC:

Figure 20: Concordance lines for scheme *in COCA*

Figure 21: Concordance lines for scheme *in the BNC*

As we examine the concordance lines, we notice that COCA uses somewhat negative words like *helter-skelter, insane, manipulated*, and *money-laundering*, whereas these are largely absent from the BNC. Our hypothesis, then, might be that the word *scheme* has a somewhat more negative 'semantic prosody' (cf. Louw 1993) in American than in British English.

Notice, however, that this analysis required us to compare concordance lines – one by one – to see this difference. In corpora the size of COCA and the BNC, there is another option. We can simply search for the collocates of a given word in one corpus, and then click on [SIDE BY SIDE] to find the collocates in the other dialect, and then compare the two sets of collocates. As an example of

this, let us compare the collocates of *scheme* side-by-side in the two dialects (see Figure 22):

| COCA: 450,000,000 WORDS | | | | | | BNC: 100,000,000 WORDS REMOVE COCA | | | | | |
WORD/PHRASE	1: COCA	2: BNC	PM 1	PM 2	RATIO	WORD/PHRASE	2: BNC	1: COCA	PM 2	PM 1	RATIO
1 PONZI	258	1	0.57	0.01	57.33	1 GRADUATED	32	1	0.32	0.00	144.00
2 DEFENSIVE	56	0	0.12	0.00	12.44	2 OCCUPATIONAL	156	5	1.56	0.01	140.40
3 BROADER	165	4	0.37	0.04	9.17	3 DISABLED	19	1	0.19	0.00	85.50
4 RISKY	39	1	0.09	0.01	8.67	4 REGULAR	19	1	0.19	0.00	85.50
5 HAZARDOUS	36	0	0.08	0.00	8.00	5 HYDRO-ELECTRIC	16	1	0.16	0.00	72.00
6 OFFENSIVE	30	1	0.07	0.01	6.67	6 ELIGIBLE	15	1	0.15	0.00	67.50
7 BLOCKING	26	0	0.06	0.00	5.78	7 SPECIAL	110	8	1.10	0.02	61.88
8 AGGRESSIVE	20	0	0.04	0.00	4.44	8 SHELTERED	26	2	0.26	0.00	58.50
9 ADVANCE-FEE	19	0	0.04	0.00	4.22	9 ENGLISH	13	1	0.13	0.00	58.50
10 COSMIC	19	1	0.04	0.01	4.22	10 RELEVANT	13	1	0.13	0.00	58.50
11 NEUTRAL	17	1	0.04	0.01	3.78	11 SCOTTISH	13	1	0.13	0.00	58.50
12 REGULATORY	102	7	0.23	0.07	3.24	12 DEFINED	12	1	0.12	0.00	54.00
13 GET-RICH-QUICK	58	4	0.13	0.04	3.22	13 OVERSEAS	23	2	0.23	0.00	51.75
14 CLASSIC	14	1	0.03	0.01	3.11	14 CONTINUING	11	1	0.11	0.00	49.50
15 DIABOLICAL	13	0	0.03	0.00	2.89	15 SUITABLE	11	1	0.11	0.00	49.50
16 EVIL	13	1	0.03	0.01	2.89	16 BRITISH	63	6	0.63	0.01	47.25

Figure 22: Collocates of scheme *in COCA and BNC*

As we can see, the collocates of *scheme* in American English (COCA) include the 'negative' words *Ponzi, risky, hazardous, aggressive, get-rich-quick, diabolical*, and *evil* (and many more are found beyond the first sixteen collocates, shown here). But these negative collocates are absent from the BNC, and we can therefore see that *scheme* has a much more neutral meaning in British English.

Let us briefly consider a few more examples of comparing collocates between the two corpora. Figure 23 gives collocates for the words {*napkin|napkins|nappy|nappies*}. Note that the BNC has more collocates referring to children (e.g. *baby, children, rash, toy, child*), showing that this word has roughly the same meaning as the American *diaper*. In American English, though, it refers to the British *serviette*, and this shows up with collocates referring to food and dining, like *cocktail, silverware, plates*, and *cups*.

| COCA: 450,000,000 WORDS | | | | | | BNC: 100,000,000 WORDS REMOVE COCA | | | | | |
WORD/PHRASE	1: COCA	2: BNC	PM 1	PM 2	RATIO	WORD/PHRASE	2: BNC	1: COCA	PM 2	PM 1	RATIO
1 COCKTAIL	149	1	0.33	0.01	33.11	1 BABY	27	6	0.27	0.01	20.25
2 HAIR	53	0	0.12	0.00	11.78	2 CHILDREN	4	1	0.04	0.00	18.00
3 EYES	32	0	0.07	0.00	7.11	3 RASH	4	1	0.04	0.00	18.00
4 SILVERWARE	32	0	0.07	0.00	7.11	4 TOYS	4	1	0.04	0.00	18.00
5 CHIN	24	1	0.05	0.01	5.33	5 CHILD	6	2	0.06	0.00	13.50
6 BAR	22	0	0.05	0.00	4.89	6 NIGHTS	3	1	0.03	0.00	13.50
7 FLOOR	21	1	0.05	0.00	4.67	7 PILES	3	1	0.03	0.00	13.50
8 HOLDER	21	0	0.05	0.00	4.67	8 PRODUCTS	3	1	0.03	0.00	13.50
9 PLATES	62	3	0.14	0.03	4.59	9 BED	3	1	0.03	0.00	13.50
10 CUPS	20	0	0.04	0.00	4.44	10 BOTTOM	3	1	0.03	0.00	13.50
11 DISPENSER	20	0	0.04	0.00	4.44	11 ONES	3	1	0.03	0.00	13.50
12 STACK	19	0	0.04	0.00	4.22	12 CHANGE	2	1	0.02	0.00	9.00

Figure 23: Collocates of nappy+ *in COCA and BNC*

Sometimes the difference in meaning and usage appears to be much more subtle. For example, consider the collocates for *boost* (verb) (see Figure 24). It looks like in British English, it refers primarily to 'increasing' something (e.g.

finances, figures), whereas in American English it has expanded its meaning to 'improvement' (e.g. *mood, spirits, security*). But notice also how few tokens we have more many of the collocates in the BNC (perhaps just 3–4 tokens), since it is a much smaller corpus than COCA, which is four or five times as large.

COCA: 450,000,000 WORDS

	WORD/PHRASE	1: COCA	2: BNC	PM 1	PM 2	RATIO
1	METABOLISM	58	1	0.13	0.01	12.89
2	MOOD	54	0	0.12	0.00	12.00
3	ODDS	44	0	0.10	0.00	9.78
4	BRAIN	43	1	0.10	0.01	9.56
5	SPIRITS	41	1	0.09	0.01	9.11
6	LEVEL	39	1	0.09	0.01	8.67
7	FUEL	38	1	0.08	0.01	8.44
8	RETURNS	37	1	0.08	0.01	8.22
9	SCORES	36	1	0.08	0.01	8.00
10	PERCENT	35	0	0.08	0.00	7.78
11	SECURITY	33	1	0.07	0.01	7.33
12	STUDENT	33	0	0.07	0.00	7.33

BNC: 100,000,000 WORDS

	WORD/PHRASE	2: BNC	1: COCA	PM 2	PM 1	RATIO
1	SQUAD	6	1	0.06	0.00	27.00
2	COUNCIL	4	1	0.04	0.00	18.00
3	CUP	4	1	0.04	0.00	18.00
4	POLICE	4	1	0.04	0.00	18.00
5	USERS	4	1	0.04	0.00	18.00
6	CO-OPERATION	3	1	0.03	0.00	13.50
7	GRANT	3	1	0.03	0.00	13.50
8	MEDIA	3	1	0.03	0.00	13.50
9	VISITORS	3	1	0.03	0.00	13.50
10	WALL	3	1	0.03	0.00	13.50
11	FINANCES	3	1	0.03	0.00	13.50
12	FIGURES	9	3	0.09	0.01	13.50

Figure 24: Collocates of boost *(v) in COCA and BNC*

As with *boost* (v), sometimes the difference in meaning and usage is quite subtle. For example, Figure 25 shows the collocates of *flip* (v) in the two dialects. The list of noun collocates in COCA (e.g. *light, hair, phone, bird, head, channels*) suggests that in American English it refers to a quick movement of the hand or finger, but that meaning is not found in British English (or at least the BNC, from a generation ago).

COCA: 450,000,000 WORDS

	WORD/PHRASE	1: COCA	2: BNC	PM 1	PM 2	RATIO
1	LIGHT	155	1	0.34	0.01	34.44
2	HAIR	93	0	0.21	0.00	20.67
3	PHONE	83	0	0.18	0.00	18.44
4	BIRD	77	0	0.17	0.00	17.11
5	HEAD	68	0	0.15	0.00	15.11
6	LIGHTS	66	1	0.15	0.01	14.67
7	CHANNELS	65	0	0.14	0.00	14.44
8	TV	49	0	0.11	0.00	10.89
9	PAGE	177	4	0.39	0.04	9.83
10	MAGAZINE	35	0	0.08	0.00	7.78

BNC: 100,000,000 WORDS

	WORD/PHRASE	2: BNC	1: COCA	PM 2	PM 1	RATIO
1	HEART	3	1	0.03	0.00	13.50
2	NUMBER	2	1	0.02	0.00	9.00
3	AXIS	2	1	0.02	0.00	9.00
4	WEEK	2	1	0.02	0.00	9.00
5	MIND	3	2	0.03	0.00	6.75
6	WIPERS	10	7	0.10	0.02	6.43
7	HALF	2	2	0.02	0.00	4.50
8	ARM	1	1	0.01	0.00	4.50
9	SUPPORT	1	1	0.01	0.00	4.50
10	TRANSMITTER	1	1	0.01	0.00	4.50

Figure 25: Collocates of flip *(v) in COCA and BNC*

Finally, consider the collocates of *web* in COCA and the BNC; see Figures 26. Because COCA currently covers the period 1990–2012 (and continues to grow each year), it contains many collocates relating to the World Wide Web, which had not yet really appeared by 1992, as finishing touches were being put on BNC. Therefore the meaning of *web* in BNC is limited primarily to the {spider / network} meaning of web, with collocates like *arms, bird, bodies*, and *greed*.

COCA: 450,000,000 WORDS						BNC: 100,000,000 WORDS					
	WORD/PHRASE	1: COCA	2: BNC	PM 1	PM 2	RATIO		WORD/PHRASE	2: BNC	1: COCA	PM 2	PM 1	RATIO
1	SITE	15446	0	34.32	0.00	3,432.44	1	ARMS	2	1	0.02	0.00	9.00
2	SITES	4166	1	9.26	0.01	925.78	2	BIRD	2	1	0.02	0.00	9.00
3	WORLD	1463	1	3.25	0.01	325.11	3	BLOCK	2	1	0.02	0.00	9.00
4	PAGE	1144	0	2.54	0.00	254.22	4	BODIES	2	1	0.02	0.00	9.00
5	PAGES	749	0	1.66	0.00	166.44	5	GREED	2	1	0.02	0.00	9.00
6	E-MAIL	621	0	1.38	0.00	138.00	6	OUTLINE	2	1	0.02	0.00	9.00
7	COMPANY	447	1	0.99	0.01	99.33	7	FEMALE	4	4	0.04	0.01	4.50
8	BROWSER	430	0	0.96	0.00	95.56	8	DEW	2	2	0.02	0.00	4.50
9	INTERNET	350	0	0.78	0.00	77.78	9	FABRIC	2	2	0.02	0.00	4.50
10	ADDRESS	346	0	0.77	0.00	76.89	10	ALLEGIANCES	1	1	0.01	0.00	4.50
11	INFORMATION	616	2	1.37	0.02	68.44	11	BRIDGES	1	1	0.01	0.00	4.50
12	PEOPLE	287	1	0.64	0.01	63.78	12	CLAY	1	1	0.01	0.00	4.50
13	ACCESS	233	1	0.52	0.01	51.78	13	COMMITTEES	1	1	0.01	0.00	4.50
14	SEARCH	233	0	0.52	0.00	51.78	14	CONTINENT	1	1	0.01	0.00	4.50
15	CONTENT	220	1	0.49	0.01	48.89	15	CHEEK	1	1	0.01	0.00	4.50

Figure 26: Collocates of web *in COCA and BNC*

Obviously, in this case we are not really talking about dialectal differences, but rather differences in meaning over time. With the word *web*, the difference is quite easy to see, but with some words where the semantic change has been more subtle (perhaps *boost* or *flip*?), the semantic difference between the two corpora will be much more subtle as well.

6   Summary, and a few concluding comments

As we have seen, to this point in time it has been difficult for most researchers to carry out a wide range of comparisons on British and American English. Only a handful of researchers have access to large, 'in-house' corpora of these two dialects. As a result, some researchers have chosen to focus on just a narrow range of genres, such as a few British and American newspapers. The disadvantage of this approach, of course, is that our understanding of the dialectal differences hinges on whether that one easily-available genre can in fact serve as a proxy for all genres, such as fiction or academic.

The most common approach to looking at dialectal differences, however, has been to use rather small corpora like the Brown family of corpora (Brown, LOB, Frown, FLOB). For example, this is the approach taken by the majority of the studies in Leech *et al*. (2009). The serious downside of this approach, however, is that with just 2–4 million words total size, the range of phenomena that can be studied is rather small. As I have noted, more than half of the studies in Leech *et al*. (2009) deal with high frequency constructions like modals, other auxiliaries, and passives, which have already been considered in great detail in many other studies over the past 10–15 years. So as insightful as these studies are, there is very little on medium and low-frequency syntactic constructions, and nothing on lexical, morphological, or semantic differences between the two dialects.

In this paper, I have shown how we can compare British and American English using the BNC and COCA. We have seen that the data from these two genre-balanced corpora is rich enough to enable us to examine a wide range of linguistic phenomena – whether lexical, morphological, syntactic, or semantic. However, there are three issues related to these two corpora that I should mention in conclusion.

First, while the genres in these two corpora are similar, they are not identical. For example, half of the spoken material in the BNC comes from 'everyday conversation', whereas COCA Spoken comes from *unscripted* conversation on national TV and radio programs, which tends to be more formal (especially in terms of topics and therefore lexis). The other four genres, however – fiction, popular magazines, newspapers, and academic – are very similar in the two corpora. Second, some of the differences in lexis and collocates between the two dialects could be rather trivial, in the sense that it might just come from orthographic differences, such as *color* / *colour* or *center* / *centre*. Fortunately, as the corpus interface compares the two dialects, it takes these superficial spelling differences into account, and removes them from the results.

The most serious challenge to using the BNC and COCA to compare the two dialects, however, is the fact that they represent two slightly different time periods. The BNC contains texts from the late 1980s / early 1990s – a full generation ago – whereas COCA continues to be updated; the most recent texts are from 2012. As we have seen, this can result in rather trivial differences, such as the fact that the collocates of *web* in COCA refer to the World Wide Web, whereas they do not in the BNC, since the Web was not around when the corpus was created. Likewise, any other change in the language (or in culture or society since the early 1990s), would result in spurious differences as well. Unfortunately, until and unless someone extends the BNC to include texts from the last 20 years, there is no easy solution to this problem.

Overall, however, we have seen that the 'Compare Corpora' functionality for the corpora from http://corpus.byu.edu/ can in general produce very useful data to compare British and American English, in ways that probably are not possible with any other resource. This is due in part to the corpora themselves, and the way that they accurately model the two dialects. Equally as important, however, is that the new corpus interface (which has been available since just late 2012), allows users to compare a wide range of phenomena in the two varieties, with just one click. Although corpus architectures and interfaces are sometimes overlooked when people are comparing corpora, in this case the architecture and interface are key to allowing us to carry out insightful comparisons of British and American English.

References

Algeo, John. 2006. *British or American English? A handbook of word and grammar patterns*. Cambridge: Cambridge University Press.

Anderson, Wendy and John Corbett. 2009. *Exploring English with online corpora: An introduction*. London: Palgrave Macmillan.

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, Stig Johansson, Geoffrey Leech, Susan Conrad and Edward Finegan. 1999. *Longman grammar of spoken and written English*. Harlow: Pearson Education.

Davies, Mark. Forthcoming a. Corpora: An introduction. In D. Biber and R. Reppen (eds.). *Cambridge handbook of English corpus linguistics*. Cambridge: Cambridge University Press.

Davies, Mark. Forthcoming b. A corpus-based study of lexical developments in Early and Late Modern English. In M. Kytö and P. Pahta (eds.). *Cambridge handbook of English historical linguistics*. Cambridge: Cambridge University Press.

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

Davies, Mark. 2012b. Examining recent changes in English: Some methodological issues. In T. Nevalainen and E. Closs Traugott (eds.). *The Oxford handbook of the history of English*, 263–287. Oxford: Oxford University Press.

Davies, Mark. 2012c. Recent shifts with three nonfinite verbal complements in English: Data from the 100 million word TIME Corpus (1920s–2000s). In B. Aarts, J. Close, G. Leech and S. Wallis (eds.). *Current change in the English verb phrase,* 46–67. Cambridge: Cambridge University Press.

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

Leech, Geoffrey, Marianne Hundt, Christian Mair and Nick Smith. 2009. *Change in contemporary English: A grammatical study.* Cambridge: Cambridge University Press

Lindquist, Hans. 2010. *Corpus linguistics and the description of English*. Edinburgh: Edinburgh University Press.

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

Mair, Christian. 2006. *Twentieth-century English: History, variation, and standardization*. Cambridge: Cambridge University Press.

Reppen, Randi. 2010. *Using corpora in the language classroom*. Cambridge: Cambridge University Press.

Rohdenburg, Günter and Julia Schlüter (eds.). 2010. *One language, two grammars? Differences between British and American English*. Cambridge: Cambridge University Press.