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

### Figure 1: Frequency of the word *pants* in COCA

![Figure 1: Frequency of the word *pants* in COCA](image)

### Figure 2: Frequency of the word *pants* in the BNC

![Figure 2: Frequency of the word *pants* in the BNC](image)

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):
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:

```
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:

```
Figure 4: ‘Side by side’ frequency of the word can 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**

<table>
<thead>
<tr>
<th>Approx. same in COCA and BNC</th>
<th>More frequent in COCA (American)</th>
</tr>
</thead>
<tbody>
<tr>
<td>head over heels in love</td>
<td>head (v) up</td>
</tr>
<tr>
<td>head-on</td>
<td>head (v) toward(s)</td>
</tr>
<tr>
<td>price on ~ head</td>
<td>head (v) back to</td>
</tr>
<tr>
<td>head for the hills</td>
<td>head (v) out</td>
</tr>
<tr>
<td>head and shoulders above</td>
<td>in over ~ head</td>
</tr>
<tr>
<td>talk over ~ head</td>
<td>(hit the) nail on the head</td>
</tr>
<tr>
<td>talk ~ head off</td>
<td>head ~ off at the pass</td>
</tr>
<tr>
<td>two heads are better (than one)</td>
<td>cooler heads (+ prevail)</td>
</tr>
<tr>
<td>use ~ head</td>
<td>go head-to-head</td>
</tr>
<tr>
<td>make ~ head spin</td>
<td>head start</td>
</tr>
<tr>
<td>put ~ heads together</td>
<td>heads or tails</td>
</tr>
<tr>
<td>bury ~ head (in the sand)</td>
<td>talking head</td>
</tr>
<tr>
<td>from head to toe</td>
<td>head game</td>
</tr>
<tr>
<td>have a head for (something)</td>
<td>head rush (n)</td>
</tr>
<tr>
<td>hanging over ~ head</td>
<td>head trip</td>
</tr>
<tr>
<td>off the top of ~ head</td>
<td>(like a) deer in the headlights</td>
</tr>
</tbody>
</table>
Powerful (yet simple) comparisons of a wide range of phenomena in British and American English

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):
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):

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, stamp 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:
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:

**Figure 9:** ‘Side by side’ display: verb + up

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

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:
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](image)

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**

<table>
<thead>
<tr>
<th></th>
<th>[have] proved</th>
<th>[have] proven</th>
<th>% proven</th>
</tr>
</thead>
<tbody>
<tr>
<td>COCA</td>
<td>3001</td>
<td>2616</td>
<td>47%</td>
</tr>
<tr>
<td>BNC</td>
<td>1669</td>
<td>82</td>
<td>4%</td>
</tr>
</tbody>
</table>

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**

<table>
<thead>
<tr>
<th></th>
<th>[have] proved</th>
<th>[have] proven</th>
<th>% proven</th>
</tr>
</thead>
<tbody>
<tr>
<td>Am: Brown/Frown</td>
<td>15</td>
<td>6</td>
<td>25%</td>
</tr>
<tr>
<td>Br: LOB/FLOB</td>
<td>47</td>
<td>1</td>
<td>2%</td>
</tr>
</tbody>
</table>

A final example is the simple past form of *sneak*: *sneaked* or *snuck*. Figure 12 shows that [pronoun] + *snuck* (e.g. *he snuck in the back window*) 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**

<table>
<thead>
<tr>
<th></th>
<th>[pronoun] sneaked</th>
<th>[pronoun] snuck</th>
<th>% snuck</th>
</tr>
</thead>
<tbody>
<tr>
<td>COCA</td>
<td>305</td>
<td>335</td>
<td>52%</td>
</tr>
<tr>
<td>BNC</td>
<td>43</td>
<td>1</td>
<td>2%</td>
</tr>
</tbody>
</table>

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**

<table>
<thead>
<tr>
<th></th>
<th>[pronoun] sneaked</th>
<th>[pronoun] snuck</th>
<th>% snuck</th>
</tr>
</thead>
<tbody>
<tr>
<td>Am: Brown/Frown</td>
<td>1</td>
<td>1</td>
<td>50%</td>
</tr>
<tr>
<td>Br: LOB/FLOB</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

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.

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:

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 \textit{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.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{Figure15.png}
\caption{\textit{end up V-ing} in COCA/BNC}
\end{figure}

Let us now examine an even more interesting and recent construction: the ‘quotative like’ construction, e.g. \textit{and I’m like, I don’t want it} (the search string is \texttt{[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.
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:
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 \[all the \text{NOUNs}\] and \[all of the \text{NOUNs}\] (e.g. \textit{all (of) the reasons}). Figure 19 shows the frequency of \[all \text{of the} \text{NOUNs}\] 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. \[all of the \text{NOUN}\] in COCA and BNC](image)

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

\begin{table}[h]
\centering
\begin{tabular}{|l|c|c|c|}
\hline
 & \text{all the} \text{[nn2]} & \text{all of the} \text{[nn2]} & \% all \text{of} \\
\hline
\text{COCA} & 58,345 & 10,170 & 14.8\% \\
\hline
\text{BNC} & 15,116 & 647 & 4.1\% \\
\hline
\end{tabular}
\end{table}

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

<table>
<thead>
<tr>
<th></th>
<th>all the [nn2]</th>
<th>all of the [nn2]</th>
<th>% all of</th>
</tr>
</thead>
<tbody>
<tr>
<td>Am: Brown/Frown</td>
<td>295</td>
<td>20</td>
<td>6.3%</td>
</tr>
<tr>
<td>Br: LOB/FLOB</td>
<td>293</td>
<td>22</td>
<td>6.9%</td>
</tr>
</tbody>
</table>

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):

![Figure 22: Collocates of scheme in COCA and BNC](image)

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*|*nappies*|*nappy*\}. 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*.

![Figure 23: Collocates of nappy+ in COCA and BNC](image)

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. *sales*, *bookings*, *prices*, *product*).
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.

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).

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.
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


