

The Corpus of Contemporary American English as the first reliable monitor corpus of English

Mark Davies

Brigham Young University, Provo, UT, USA

Abstract

The Corpus of Contemporary American English is the first large, genre-balanced corpus of any language, which has been designed and constructed from the ground up as a 'monitor corpus', and which can be used to accurately track and study recent changes in the language. The 400 million words corpus is evenly divided between spoken, fiction, popular magazines, newspapers, and academic journals. Most importantly, the genre balance stays almost exactly the same from year to year, which allows it to accurately model changes in the 'real world'. After discussing the corpus design, we provide a number of concrete examples of how the corpus can be used to look at recent changes in English, including morphology (new suffixes *-friendly* and *-gate*), syntax (including prescriptive rules, quotative *like*, *so not* ADJ, the *get* passive, resultatives, and verb complementation), semantics (such as changes in meaning with *web*, *green*, or *gay*), and lexis—including word and phrase frequency by year, and using the corpus architecture to produce lists of all words that have had large shifts in frequency between specific historical periods.

Correspondence:

Mark Davies, Linguistics and English Language, Brigham Young University.

E-mail:

mark_davies@byu.edu

1 Introduction

One of the goals of corpus linguistics during the last fifteen to twenty years has been to develop and use large 'monitor corpora'. Unlike 'static' corpora like the Brown Corpus or the British National Corpus (BNC)—which are not updated once they are created—monitor corpora are dynamic, in the sense that new texts continue to be added to the corpus. The goal of creating such corpora is to allow users to search the continually expanding corpus to see how the language is changing.

With a reliable monitor corpus, we could answer questions like the following:

(Lexical) When (and perhaps why, as a result) have the following words increased most in usage: *globalization*, *adolescent*,

insurgent, *same-sex*, *upscale*, *old-school*, *wirelessly*, *online*, and the verbs *mentor*, *morph*, *download*, *freak out*, and *splurge*?

(Lexical) What are the fifty verbs, nouns, and adjectives that have increased the most in usage during the past five years?

(Morphology) Is the 'scandal' suffix *-gate* (*Watergate*, *zippergate*, *Irangate*) increasing or decreasing in usage since the 1990s?

(Syntax) Are the grammatical constructions 'end up V-ing' (*ends up paying*, *ended up working*) and the 'get passive' (*he got hired* versus *he was hired*) increasing or decreasing over time?

(Semantics) Words such as *hot*, *lame*, *green*, or *random* have recently changed meaning. What

have been the most important periods of change, and exactly how have the new meanings been acquired?

(Discourse analysis) What are we saying differently about *women*, or the *environment*, or *immigrants* than we were 15–20 years ago?

Although the development of monitor corpora has been a long-standing goal of corpus creation and use, the fact is that until recently there actually were no reliable monitor corpora of English. There are some corpora that approximated monitor corpora and which have been advertised as monitor corpora. As we will see, however, each of them has suffered from serious flaws. As a result, to this point it has not been possible for linguists to use structured monitor corpora to carry out research on a number of types of ongoing linguistic change in contemporary English, such as those mentioned above.

Fortunately, this has changed with the introduction of the Corpus of Contemporary American English (COCA), which we created and placed online in 2008. This freely available corpus is composed of more than 400 million words from 1990 to the present time, including twenty million words each year from 1990 to 2009. In addition, the corpus will continue to be updated—twenty million words each year—with the most recent texts having been added less than one month ago. And most importantly, the genre balance of the corpus stays the same from year to year, which—as we will see—allows us to be quite certain that the changes that we see in the corpus actually reflect linguistic changes in the ‘real world’.

In Section 2 we will discuss some “near misses”—corpora and text archives that were available before 2008, and which provided some data (although not sufficient data) on linguistic shifts in English. Section 3 discusses the composition of the COCA, with an emphasis on both the historical and the genre-based organization of the corpus. Section 4 provides a number of examples of insights from the corpus on changes in contemporary American English, including changes in morphology, syntax, lexis, and semantics.

2 Previous Corpora

As was mentioned in the introduction, many of the well-known corpora of English are static. This means that once they are created, no more texts are added to the corpus, which renders them useless as monitor corpora to look at linguistic change (although they certainly do have other important uses).

Perhaps the most famous example of this is the 100 million word BNC. The corpus was completed in 1993 and contains texts from the 1970s through the early 1990s, but no more texts have been added since that time (nor will they be). Therefore, there is no way to use the corpus to look at linguistic changes since the early 1990s, and—due to the non-systematic way in which the corpus was created for each year from the 1970s to the early 1990s—there really is really no way to use the corpus to look at changes from the 1970s to the 1990s either. However, this is not really a criticism of the BNC, since it was never really designed as a monitor corpus per se, and it certainly has been a useful corpus for many other types of research (see Burnard, 2002).

In the early 1990s an attempt was made to update the Brown (US) and LOB (UK) corpora—which are based on texts from 1961—to show language change from 1961 to 1991. With the introduction of the Frown corpus (Freiburg-Brown) and FLOB (Freiburg-LOB) corpora in the early 1990s, researchers could now compare equivalent corpora thirty years apart. As admirable as these efforts were, this could not result in true monitor corpora. First and perhaps most obviously, the corpora are still ‘static’ in the sense that nothing has been added to them since 1991, so there is no way to see changes in English since that time. Second, because the corpora are spaced thirty years apart, they do not have the ‘granularity’ needed to see intervening changes. To take a very simple example, the word *hippy* hardly occurs in 1961 and is likewise very infrequent in 1991, and any changes with this word in the late 1960s are in essence ‘invisible’ to these corpora.

Finally, the ‘Brown family’ of corpora are far too small to be useful for looking at most types of language change. While they might show changes in

frequent grammatical constructions (modals, or pronominal usage), they are quite inadequate for many other types of syntactic research (such as verbal subcategorization: *John helped Mary* (\pm *to clean the room*)), and most types of semantic and lexical research, such as changes in meaning of words like *gay*, *hot*, or *green*, or the overall increase or decrease in mid- and low-frequency words—which would occur only one or two times in one million words. As we consider several specific linguistic changes in Section 4, we will see that even with 400 million words in the corpus, there are often only 100–200 tokens for some of these phenomena. With one million word corpora—like those in the Brown family—there would only be one or two tokens—far too small to look at any meaningful change.

2.1 The bank of English

Until recently, the only corpus of English with both the size and the diachronic extension to possibly be used as a monitor corpus was the Bank of English (BoE), also known as the Cobuild Corpus, and (in its most recent incarnation online) as Word Banks Online (<http://wordbanks.harpercollins.co.uk>). The corpus was started in the 1980s as the basis for the Collins Cobuild dictionaries, and texts continued to be added to the corpus each year for over twenty years, resulting in a corpus of about 455 million words by 2005.

Although (as far as we are aware) the creators of the BoE have never claimed themselves that the BoE could be used as a monitor corpus, this claim has been made on its behalf in a number of introductory and survey books on corpus linguistics (e.g. McEnery and Wilson, 2001, pp. 30–31; Meyer, 2002, p. 15; Hunston, 2002, pp. 30–31; Sampson and McCarthy, 2004, pp. 396–98; McEnery *et al.*, 2006, pp. 67–70; Baker *et al.*, 2006, p. 65).

As we will see, however, the BoE has a particular flaw which—in spite of claims to the contrary—creates serious problems in terms of its use as a monitor corpus. It is perhaps for this reason that—although the corpus has been billed as a *potentially* useful monitor corpus—in fact relatively little actual diachronic work with the corpus has actually been done.

Table 1 Composition of the BoE by period

| Time period | Fiction | Total | % fiction |
|----------------|------------------|-------------------|--------------|
| 1960–79 | 1,030,000 | 1,414,000 | 72.8% |
| 1980–89 | 3,087,000 | 8,792,000 | 35.1% |
| 1990–94 | 6,049,000 | 20,833,000 | 29.0% |
| 1995–99 | 3,100,000 | 19,187,000 | 16.2% |
| 2000–4 | 18,800,000 | 123,055,000 | 15.3% |

Values in bold are discussed in the text.

Table 2 Anomalous results for fiction-oriented words in the BoE

| | All | | Fiction | |
|---------------------------|---------------|---------------|---------------|---------------|
| | 1990–94 | 1995–99 | 1990–94 | 1995–99 |
| <i>Mutter</i> (all forms) | 378 | 269 | 326 | 159 |
| | 18.1 | 14.0 | 53.9 | 51.3 |
| <i>Pale</i> (all forms) | 707 | 402 | 421 | 202 |
| | 33.9 | 21.0 | 69.6 | 65.2 |
| <i>had</i> + VBN | 56,239 | 31,125 | 21,590 | 10,418 |
| (e.g. <i>had seen</i>) | 2669.5 | 1622.2 | 3569.2 | 3360.7 |

Normalized (per million) values are in bold.

In order to understand the weakness of the BoE as a monitor corpus, suppose that we take a worst-case example, and that we have a corpus had only newspapers from the 1990s and then only fiction from the 2000s. For any change that we see from the 1990s to the 2000s, we would not know if the change had actually occurred in the language as a whole, or if it is just an ‘artifact’ of the changing genre composition from one period to the next.

Although things are not this serious with the BoE, they are quite problematic. It appears that in the creation of the corpus, little if any attention was paid to the issue of keeping the genre balance the same from one year or historical period to another. For example, Table 1 shows the percentage of the US sub-corpus in different historical periods, which comes from fiction.

Notice how the percentage of fiction decreases by nearly 50% from the early 1990s to the late 1990s. Let us briefly look at how this distorts the corpus data for these periods.

Notice that the three forms in Table 2 (*mutter*, *pale*, and *had* + VBN) are characteristic of fiction (see Biber *et al.*, 1999, pp. 36–572 (lexical items by

genre) and pp. 467–70 for the past perfect in fiction and other genres). Notice also that in just the US fiction part of the BoE (the two rightmost columns), the frequency per million words stays about the same from 1990–94 to 1995–99, as we would expect (the lower figure in each cell is the normalized frequency, with the top one being raw frequency). But in the entire US part of the BoE (the first two columns), the normalized frequency (per million words) decreases much more from 1990–94 to 1995–99. For example, *had* + VBN decreases by about 40%. Why is this?

To find the answer, note that in Table 1 the percentage of the US sub-corpus in the BoE that is fiction decreased by about 55% during the same period (1990–94 to 1995–99). In other words, the decrease of the phenomena in Table 2 is simply a function of the change in genre balance, rather than any change in ‘real world’ language. (After all, it would be quite strange if people really did all of the sudden say *had eaten*, *had noticed*, etc. only 50% as much in the late 1990s as the early 1990s!) This is one simple example that shows how crucial it is to keep the genre composition the same from year to year¹.

In terms of genres, in addition to the issue of balance there is also the issue of the lack of informal texts in the BoE. Ideally, a corpus will contain at least some moderately informal spoken texts, since this is often where language change originates. In looking at the spoken texts in the BoE, however, we find that (for American English, at least) they are limited to transcripts from the *Voice of America* radio broadcasts. As a result, they do not model very well informal American speech. To give just one example, consider the following, which shows

the number of tokens of the ‘quotative *like*’ construction (*and she’s like*, *I don’t know*) (Table 3).

It is quite strange that there are so few examples of this colloquial construction in the BoE. As we will see when we consider this same construction in the COCA (Table 16), the frequencies (per million words) are anywhere from 3 to 66 times as common in COCA as in the BoE.

Even beyond these serious problems of genre balance and the lack of informal texts, it appears that there might be an even more fundamental problem with the BoE. To see what this is, consider Table 4.

Table 4 shows the frequency of two common words (*is* and *and*) and one grammatical construction (*was* VVN: *was seen*, *was considered*) in the BoE in three period—1990–94, 1995–1999, and 2000–2004. (The raw frequency data are in parentheses, while the normalized value per million words is in bold.) The following two columns (90–94 > 95–99 and 95–99 > 00–04) shows the percentage change (for the normalized figures) between 1990–94 and 1995–99 and for 1995–99 and 2000–04. For example, in the BoE the frequency of the passive ‘decreased’ 31% between 1990–94 and 1995–1999, and then ‘increased’ 36% between 1995–99 and 2000–04.

Table 3 Frequency of ‘quotative *like*’ in the US portion of the BoE

| Years | Tokens | Size | Per million |
|---------|--------|-------------|-------------|
| 1990–94 | 5 | 20,883,000 | 0.24 |
| 1995–99 | 1 | 19,187,000 | 0.05 |
| 2000–04 | 173 | 123,055,000 | 1.41 |

Normalized (per million) values are in bold.

Table 4 Frequency of common words (by period) in the BoE

| | Periods | | | Change | |
|---------|-------------------------|-------------------------|---------------------------|-------------------|-------------------|
| | 1990–94 | 1995–99 | 2000–4 | 1990–94 > 1995–99 | 1995–99 > 2000–04 |
| was VVN | 1550 (32,370) | 1071 (20,558) | 1458 (179,367) | 0.69 | 1.36 |
| is | 6443 (134,551) | 8225 (157,808) | 6558 (810,686) | 1.28 | 0.80 |
| and | 22,400 (467,783) | 22,517 (432,037) | 18,580 (2,286,364) | 1.01 | 0.83 |

Values in bold are discussed in the text.

Notice that the normalized values for these very frequent words and phrases varies widely from one period to the next. One might wonder why the passive would increase or decrease 30–35% per cent between two adjacent five-year periods, or why a very common word like *is* or *and* would vary by 20–30% from one period to the next. And notice that it is not just a problem with corpus sizes and bad calculations—with one word the frequency might increase dramatically between two periods in the BoE, while with another word it might decrease dramatically during the same period. Since the frequency statistics are so strange for common, predictable words, it is very difficult to have confidence that the BoE will provide accurate data for other words, phrase, and grammatical constructions that we might be researching. But again, perhaps we are expecting too much of the BoE, since it is not clear that the corpus creators ever explicitly designed to accurately map out changes over time.

2.2 The Oxford English corpus

The Oxford English corpus (OEC) (<http://www.askoxford.com/oec/>) has received much less attention than the BoE. This is partly because it is much more recent than the BoE. More importantly, however, it is probably because the OEC has had even more restricted access than the BoE. The OEC was designed for the express purpose of materials development and research for Oxford University Press, and very few outside of the OUP have had access to the corpus. For the purposes of this paper, however, we were granted access to the corpus.

In general terms, one of the real advantages of the corpus is its size. At about 1.9 billion words, it is

almost 19 times as large as the BNC, and 4 to 5 times as large as the COCA and the BoE. One of its drawbacks, however, is the very limited time period that it covers. As can be seen in the following table, it is limited to just 2000–06. No work has been done on the expanding the corpus since that time.

In terms of genre distribution, Table 5 shows the corpus size and the number of words from fiction for just the US portion of the corpus, which comprises a little more than one half of the corpus – 927,000,000 of the 1,889,000,000 words.

As can be seen, the OEC suffers from the same problem as the BoE—the lack of genre balance from one period to the next. For example, notice that in 2000 about 10% from the corpus is from fiction texts, whereas this increases to 82% six years later in 2006. Even from one year to the next, the genre balance can change quite dramatically, such as the 50% drop in fiction from 2003 to 2004.

As with the BoE (see Table 2), this lack of genre balance has serious implications in terms of the data. As can be seen in the Table 6, for example, the frequency of ‘fiction’ words like *mutter* vary

Table 5 Composition by year in the US portion of the Oxford English corpus

| Year | Fiction | Total | Fiction (%) |
|-------------|------------|-------------|-------------|
| 2000 | 6,479,988 | 66,455,562 | 9.8 |
| 2001 | 14,326,315 | 89,913,492 | 15.9 |
| 2002 | 36,938,545 | 142,621,850 | 25.9 |
| 2003 | 61,788,465 | 191,239,937 | 32.3 |
| 2004 | 53,462,736 | 240,840,436 | 22.2 |
| 2005 | 57,083,698 | 180,930,648 | 31.6 |
| 2006 | 12,740,916 | 15,442,798 | 82.5 |

Values in bold are discussed in the text.

Table 6 Fiction-oriented words and constructions in the BoE

| | Entire corpus | | | Fiction | | |
|---|------------------------|--------------------------|-------------------------|-------------------------|--------------------------|-------------------------|
| | 2001 | 2004 | 2006 | 2001 | 2004 | 2006 |
| <i>Mutter</i> (all forms) | 1669 18.6 | 8552 44.7 | 1652 107.0 | 1557 110.1 | 5927 110.9 | 1647 129.3 |
| <i>Pale</i> (all forms) | 2186 24.3 | 6543 27.2 | 1203 77.9 | 1174 81.9 | 4223 79.0 | 1190 93.4 |
| <i>had</i> + VBN (e.g. <i>had seen</i>) | 81,811 909.9 | 245,966 1021.3 | 32,178 2083.7 | 36,135 2522.3 | 135,952 2542.9 | 30,535 2396.6 |

Normalized (per million) values are in bold.

widely from one year to the next (the bolded figures in the ‘Entire corpus’ section to the left). In terms of fiction, however (the right side), the frequency stays almost flat from one year to the next. The effect is not quite as noticeable with grammatical construction like *had* + VVN (*had seen*, *had paid*), but it is still present. The normalized frequency (per million words) increases somewhat from 2001 to 2004, and then increases dramatically from 2004 to 2006, as the overall percentage of fiction texts in the corpus increases from 22.5 to 82.5% (see Table 5). When we look just at the fiction texts, however, the normalized frequency stays almost perfectly flat from 2001 to 2006 (right around 2400–2500 occurrences per million words).

In summary—as with the BoE—the frequency of words, phrases, and grammatical constructions in the OEC varies widely from one year to the next, but it is probably just an artifact of the changing genre balance, rather than any indication of real change in the language (as we would hope would be the case with a real monitor corpus).

In terms of range of genres, and particularly texts that represent informal English, the OEC is about on par with the BoE, but it has many fewer informal, spoken texts than the COCA. For example, in the OEC the normalized figures for the ‘quotative *like*’ construction (*and I’m like*, *I’m not going*) range from 0.44 to 0.98 tokens per million words, or an overall figure of 0.76 (Table 7).

In the BoE (Table 3) it was between 0.05 and 1.41, with an overall figure of 1.10. In the COCA, on the other hand, it ranges from a low of 1.3 tokens per million words in the early 1990s to 6.9 in the late 2000s, for an overall figure of 3.94 (see Table 16). Therefore a colloquial construction like this appears about 4 to 5 times as frequently in the COCA as in either the BoE or the OEC.

In summary, the OEC is very impressive because of its size—nearly two billion words. On the downside, it is limited to just the period 2000–2006, about one-third the chronological range of the COCA. As with the BoE, it is not genre-balanced from year to year, and it has much less colloquial material than COCA (which is the type of language where changes often start).

Table 7 Frequency of ‘quotative *like*’ in the US portion of the OEC

| Years | Tokens | Size | Per million |
|-------|--------|-------------|-------------|
| 2000 | 45 | 66,455,562 | 0.68 |
| 2001 | 40 | 89,913,492 | 0.44 |
| 2002 | 111 | 142,621,850 | 0.78 |
| 2003 | 121 | 191,239,937 | 0.63 |
| 2004 | 202 | 240,840,436 | 0.84 |
| 2005 | 177 | 180,930,648 | 0.98 |
| 2006 | 12 | 15,442,798 | 0.78 |

To be fair to the COE, however, we should not that it may have never been explicitly designed to model change over time or to contain texts from a wide range of genres. And of course, all of these issues are mere details compared to the fact that the OEC is not available—except in rare cases—to anyone outside of Oxford University Press, whereas the COCA is freely available on the web.

2.3 Text archives

With all of these problems with ‘structured corpora’, one might be tempted to use ‘unstructured corpora’ like the Web or ‘text archives’ of newspaper and magazine articles instead. After all, these text archives have the advantage of being much larger than structured corpora, and many of them extend back 15–20 years or more. Unfortunately, such archives are in fact not a reasonable option for use as monitor corpora.

The search engine for such archives is not oriented towards linguistic research, and users therefore can typically only search for exact words and phrases. There is little ability to search by substrings (thus hindering studies of morphological change), no ability to search by part of speech tags (problematic for studies of syntactic change), and little if any ability to look at changes in collocates (thus hindering studies of semantic change). Even in the domain of lexis, the search interface would need to break down the frequency by year, to see whether the frequency is increasing or decreasing (and we would then have to normalize these raw frequencies, by knowing the overall size of the corpus each year). This is also not possible with any interface that we are aware of.

So although unstructured text archives might look promising, they are in fact not a reasonable option for monitor corpora. When we combine this with the fact that there are no structured corpora of English that can serve as truly reliable monitor corpora, we can see that until the last year or two, there were really very few (if any) corpus-based tools to look at recent changes in English

3. The COCA

In this section we will briefly discuss the composition of the COCA, which is—as we will claim—the first reliable monitor corpus of English, and the first balanced monitor corpus of any language.²

A crucial aspect of the design of the COCA is that the corpus is divided almost equally between spoken, fiction, popular magazines, newspapers, and academic journals—20% in each genre (see Davies 2009b for a more complete overview of the textual corpus, and Davies (2005) and Davies (2009a) for information on earlier versions of the corpus architecture).

As of August 2009, there are more than 160,000 texts in the corpus, and they come from a variety of sources:

Spoken: (83 million words) Transcripts of unscripted conversation from more than 150 different TV and radio programs (examples: *All Things Considered* (NPR), *Newshour* (PBS), *Good Morning America* (ABC), *Today Show* (NBC), *60 Minutes* (CBS), *Hannity and Colmes* (Fox), *Jerry Springer*, *Oprah*, etc).

Fiction: (79 million words) Short stories and plays from literary magazines, children's magazines, popular magazines, first chapters of first edition books 1990-present, and movie scripts.

Popular magazines: (84 million words) Nearly 100 different magazines, with a good mix (overall, and by year) between specific domains (news, health, home and gardening, women, financial, religion, sports, etc). A few examples are *Time*, *Men's Health*, *Good Housekeeping*, *Cosmopolitan*, *Fortune*, *Christian Century*, *Sports Illustrated*, etc.

Newspapers: (79 million words) Ten newspapers from across the US, including: *USA Today*,

New York Times, *Atlanta Journal Constitution*, *San Francisco Chronicle*, etc. There is also a good balance between different sections of the newspaper, such as local news, opinion, sports, financial, etc.

Academic journals: (79 million words) Nearly 100 different peer-reviewed journals. These were selected to cover the entire range of the Library of Congress classification system (e.g. a certain percentage from B (philosophy, psychology, religion), D (world history), K (education), T (technology), etc.), both overall and by number of words per year.

In terms of its use as a monitor corpus, the crucial point is that the genre balance stays almost exactly the same from year to year. In other words, in each year from 1990 to 2009, 20% of the corpus is from spoken, 20% from fiction, 20% from popular magazines, 20% from newspapers, and 20% from academic journals. In addition, the balance between sub-genres (e.g. Newspaper–Sports or Academic–Medicine) stays roughly the same from year to year as well.

This balance between genres from year to year results in a corpus that provides data that is quite different from the BoE and the OEC, shown above. For example, compare Table 8 to Tables 2 and 6.

Here we see that—unlike the BoE and the OEC—the frequency of these three ‘fiction’-oriented words and constructions stays essentially flat in the four periods from the early 1990s to the current time. This is because the percentage of the corpus that is fiction stays almost exactly the same—20% in each year from 1990 to 2009. Consider also Table 9.

As in Table 4, this shows the normalized frequencies (per million words) for four common words, phrases, and grammatical constructions in COCA from the early 1990s to 2004 (data are also available for 2005–09 but we have omitted it here, to enable easier comparison with Table 4). Notice that the frequency of these words is essentially flat over

Table 8 Results for fiction-oriented words in COCA

| | 1990–94 | 1995–99 | 2000–04 | 2005–09 |
|---------|---------|---------|---------|---------|
| mutter | 14.9 | 13.4 | 14.8 | 15.9 |
| had VVN | 1173.1 | 1066.2 | 1059.0 | 1095.5 |

time (as we would expect it to be), and that we do not have the strange anomalies that are found in the BoE.

The foregoing is meant to provide an introduction to COCA, in terms of its balance from year to year. Let us now provide a number of more in-depth examples of how the COCA data can be used to answer interesting questions about changes in English morphology, syntax, semantics, and discourse during the past 15 to 20 years—in ways that are not possible with any other existing corpus, text archive, or linguistic resource.

4. Recent Changes in English

4.1 Lexical change

At the most basic level, COCA can give the frequency of any word or phrase in each of the five time periods (1990–94, 1995–99, 2000–04, and 2005–09), as well as the five main genres (spoken, fiction, popular magazine, newspaper,

and academic). For example, Table 10 shows the frequency for all forms of *morph* as a verb.

We see the increase in each five year block since the early 1990s and we also see that *morph* is used the most in popular magazines. By clicking on (See All Sections), users can also see the frequency in each of the seventy sub-genres of the corpus (e.g. Magazines–Entertainment or Newspaper–Sports), as well as each individual year since 1990. For example, Table 11 shows partial entries by year, including 2008 and 2009 (when it was used the most), about ten years ago, and then the early 1990s.

This type of frequency information is far superior to what is typically done in terms of tracking word usage. Suppose that a word occurs once in some collection of texts in 1990, but then it does not really increase until the late 1990s, and it then decreases after 2005 or so. In a typical dictionary like the OED, great care would be taken to find the ‘first attestation’ (1990), but then there would be little or no information about frequency after that time. With COCA, all of this is easily and almost instantaneously available to even inexperienced users.

Table 9 Frequency of common words (by period) in COCA

| | Periods | | | Change | |
|---------|---------|---------|---------|--------------------|--------------------|
| | 1990–94 | 1995–99 | 2000–04 | 1990–94 1995–99 | 1995–99 2000–04 |
| was VVN | 1305 | 1235 | 1234 | 0.95 | 1.00 |
| to be | 1560 | 1517 | 1490 | 0.97 | 0.98 |
| is | 9549 | 9414 | 9190 | 0.99 | 0.98 |
| and | 26606 | 26731 | 26782 | 1.00 | 1.00 |

Table 11 Frequency of *morph* in COCA by year

| Rank | Year | Per million | Token | Corpus size |
|------|------|-------------|-------|-------------|
| 1 | 2008 | 5.3 | 85 | 15,920,933 |
| 2 | 2009 | 4.3 | 48 | 11,102,803 |
| 10 | 2001 | 2.2 | 44 | 20,110,099 |
| 11 | 1999 | 1.7 | 36 | 20,607,309 |
| 18 | 1993 | 0.2 | 4 | 20,761,353 |
| 19 | 1991 | 0.0 | 1 | 20,639,513 |
| 20 | 1990 | 0.0 | 0 | 20,532,370 |

Table 10 Frequency of *morph* in COCA by genre and time period

| SECTION | SPOK | FIC | MAG | NEWS | ACAD | 1990–94 | 1995–99 | 2000–04 | 2005–09 |
|------------------|------|------|------|------|------|---------|---------|---------|---------|
| SEE ALL SECTIONS | | | | | | | | | |
| PER MIL | 0.9 | 2.2 | 3.4 | 2.4 | 0.9 | 0.3 | 0.9 | 3.0 | 3.9 |
| SIZE (MW) | 81.7 | 78.8 | 83.3 | 79.4 | 79.3 | 103.3 | 102.9 | 102.6 | 93.6 |
| FREQ | 77 | 174 | 282 | 187 | 69 | 31 | 96 | 310 | 363 |

With the COCA interface, however, it is possible to do much more than just find the frequency of a specific word or phrase. Because the corpus architecture stores the frequency of all words and *n*-grams (up to 10-grams) for each section of the corpus (genre and year), users can query the corpus to find words that have a given frequency in one section of the corpus (e.g. 2005–09) and which are much more frequent than in another section (e.g. 1990–94). Setting up a query like this takes just a few clicks of the mouse and less than five seconds.

Table 12 gives an example of such a query. To the left it shows verbs that are much more common in 2005–09 than in 1990–94, while to the right it shows verbs that are much more common in 1990–94 than in 2005–09. In both cases, it shows the frequency in the two sections, and (RATIO) indicates how much more frequent the verb is (normalized, per million

words). (Note also that in these and similar tables in this article, some additional frequency information that is seen in the web interface is deleted for reasons of space. Note also that the results are ‘smoothed’ for words where the frequency is 0 in the other time period.)

Finally, consider Table 13. This shows phrasal verbs that are much more common in 2005–09 than in 1990–94 (to the left), and those which have decreased in frequency since 1990–94 (to the right) (see Martin, 1991 and Gardner and Davies, 2006 for recent data on phrasal verbs in English). Note that in order to simplify this display, we have searched for just the infinitival form, rather than all verb forms.

These data should resonate with native speakers of the language. There really is more of a colloquial feel with most of the words to the left (*hit someone up for \$5, change up the order, switch up your style,*

Table 12 Comparison of verbs, 2005–09 and 1990–94

| Word | 2005–09 | 1990–94 | Ratio | Word | 1990–94 | 2005–09 | Ratio |
|-----------|---------|---------|-------|--------------|---------|---------|-------|
| Blog | 51 | 0 | 54.51 | Propitiate | 13 | 1 | 11.77 |
| Multitask | 39 | 0 | 41.68 | Excoriate | 11 | 1 | 9.96 |
| Email | 35 | 0 | 37.41 | Gurgle | 11 | 1 | 9.96 |
| Morph | 70 | 3 | 25.77 | Demilitarize | 18 | 2 | 8.15 |
| Teleport | 21 | 1 | 23.19 | Moralize | 17 | 2 | 7.70 |
| Upload | 107 | 7 | 16.88 | Deemphasize | 15 | 2 | 6.79 |
| Outsource | 44 | 3 | 16.20 | Militate | 15 | 2 | 6.79 |
| Game | 13 | 1 | 14.36 | Preen | 13 | 2 | 5.89 |
| Reengage | 24 | 2 | 13.25 | Redound | 12 | 2 | 5.43 |

Table 13 Comparison of phrasal verbs with *up*, 2005–09 and 1990–94

| Word | 2005–09 | 1990–94 | % | Word | 1990–94 | 2005–09 | % |
|------------|---------|---------|------|------------|---------|---------|-------|
| Hit up | 21 | 3 | 7.70 | Foul up | 12 | 1 | 10.91 |
| Change up | 11 | 2 | 6.05 | Sew up | 12 | 3 | 3.64 |
| Switch up | 11 | 2 | 6.05 | Yield up | 11 | 3 | 3.33 |
| Listen up | 18 | 4 | 4.95 | Swell up | 22 | 8 | 2.50 |
| Rest up | 18 | 5 | 3.96 | Punch up | 16 | 6 | 2.42 |
| Snatch up | 21 | 6 | 3.85 | Bind up | 15 | 6 | 2.27 |
| Fatten up | 16 | 6 | 2.93 | Point up | 15 | 6 | 2.27 |
| Queue up | 13 | 5 | 2.86 | Hunt up | 10 | 4 | 2.27 |
| Board up | 10 | 4 | 2.75 | Button up | 10 | 4 | 2.27 |
| Ratchet up | 43 | 19 | 2.49 | Carve up | 22 | 9 | 2.22 |
| Scale up | 20 | 9 | 2.44 | Snuggle up | 14 | 6 | 2.12 |

Table 14 Comparison of *-friendly* words, 2005–09 and 1990–94

| Word | 2005–09 | 1990–94 | Ratio |
|-----------------|---------|---------|-------|
| Eco-friendly | 182 | 15 | 13.35 |
| Budget-friendly | 26 | 3 | 9.53 |
| Dog-friendly | 11 | 2 | 6.05 |
| Child-friendly | 55 | 11 | 5.50 |
| Market-friendly | 15 | 3 | 5.50 |
| Earth-friendly | 41 | 9 | 5.01 |
| Gay-friendly | 12 | 3 | 4.40 |
| Family-friendly | 185 | 71 | 2.87 |
| Fan-friendly | 15 | 7 | 2.36 |
| Kid-friendly | 85 | 40 | 2.34 |

etc), and many of those on the right already have a very slight “old-fashioned” feel after only 10 to 15 years (*they really fouled up, sew up the nomination, yield up its secrets*, etc).

Beyond these comparisons between blocks of the corpus, it is also possible to do comparisons even at the level of individual years, such as adjectives that are much more common in 2009 than earlier, or nouns that increased greatly in 2001 or 2002, but which have decreased since then. No other corpus, corpus architecture, or corpus interface provides this type of ability to compare and contrast the frequency of words between different historical sections of the corpus.

4.2 Morphological change

The COCA can also be used to look at changes in word formation, including the ‘productivity’ of given morphemes—where productivity refers to how freely a given morpheme can be used to create new words.

Let us consider as the first example a phenomenon that straddles the line between lexical change (in the section above) and morphology. This deals with the rise in hyphenated *-friendly* words, as shown in Table 14. The table shows the words ending in *-friendly* that are at least twice as common in 2005–09 as in 1990–94. (There is only one word—*customer-friendly*—that was more common in 1990–94, and *user-friendly* has about the same normalized frequency in both periods.)

Table 15 Comparison of *-gate* words, 1990s and 2000s

| WORD | 1990s | 2000s | RATIO |
|----------------|-------|-------|-------|
| Whitewatergate | 27 | 1 | 25.68 |
| Iraqgate | 43 | 0 | 20.85 |
| Zippergate | 14 | 1 | 13.31 |
| Filegate | 79 | 6 | 12.52 |
| Travelgate | 143 | 16 | 8.50 |
| Cattlegate | 7 | 1 | 6.66 |
| Chinagate | 7 | 1 | 6.66 |
| Irangate | 23 | 6 | 3.65 |
| Watergate | 1768 | 624 | 2.69 |
| Rubbergate | 5 | 0 | 2.42 |
| Spygate | 11 | 0 | 5.61 |
| Memogate | 9 | 0 | 4.59 |

As one can see, there has been a real increase in the use of this form since the early 1990s.

As a second example, consider the Table 15, which shows words that are formed with the novel suffix ‘*-gate*’ (referring to ‘scandal’) in the 1990s and 2000s. The words in the table to the left are from the 1990s, which occur at least five times in the corpus and which occur (per million words) at least twice as much as in the 2000s. Those in the table to the right are the *-gate* words that are at least twice as frequent in the 2000s. Notice that these *-gate* ‘scandal words’ appear to have been quite a bit more common in the 1990s, which may relate to current events from that time. Or it may simply indicate that the ‘trendy’ suffix from the 1990s is now seen as being a little bit passé.

4.3 Syntactic change

In looking at syntax, we will consider two very salient recent changes (‘quotative *like*’ and ‘*so not* ADJ’), changes in two prescriptively-focused constructions (*can/may* for permission, and split infinitives) and then three much less salient constructions: [*end up* V-ing], the ‘*get* passive’, and [*help* (to) V]. With all three types of constructions, the COCA provides very useful data, which would likely not be available from any other source.

First, let us consider the rise in two fairly salient grammatical constructions that have increased in frequency during the past two decades: the ‘quotative *like*’ construction (*and he’s like, I’m not going with her*) and the ‘*so not*’ construction (*I’m so not*

Table 16 Frequency of quotative *like*, by genre and time period



















| SECTION | SPOK | FICT | MAG | NEWS | ACAD | 1990–94 | 1995–99 | 2000–04 | 2005–09 |
|----------------|---|---|---|---|---|---|---|---|---|
| |  |  |  |  |  |  |  |  |  |
| PER MIL | 12.5 | 0.9 | 3.3 | 2.3 | 0.4 | 1.3 | 3.4 | 4.5 | 6.9 |
| SIZE (MW) | 81.7 | 78.8 | 83.3 | 79.4 | 79.3 | 103.3 | 102.9 | 102.6 | 93.6 |
| FREQ | 1025 | 72 | 271 | 179 | 29 | 130 | 347 | 462 | 645 |

Table 17 Frequency of [*so not* ADJ], by genre and time period

| SECTION | SPOK | FICT | MAG | NEWS | ACAD | 1990–94 | 1995–99 | 2000–04 | 2005–09 |
|----------------|---|---|---|---|---|---|---|---|---|
| |  |  |  |  |  |  |  |  |  |
| PER MIL | 0.17 | 0.13 | 0.14 | 0.0 | 0.0 | 0.02 | 0.06 | 0.11 | 0.18 |
| SIZE (MW) | 81.7 | 78.8 | 83.3 | 79.4 | 79.3 | 103.3 | 102.9 | 102.6 | 93.6 |
| FREQ | 14 | 10 | 12 | 0 | 0 | 2 | 6 | 11 | 17 |

interested in him). Turning first to the ‘quotative *like*’ (cf. Tagliamonte and D’Arcy, 2004, Buchstaller and D’Arcy, 2009; Barbieri, 2009), Table 16 shows the increase in the ‘quotative *like*’ construction during the past two decades.

As can be seen, its frequency has steadily increased in each five-year period since the early 1990s. Via the corpus interface, it is also possible to see the normalized frequency in each individual year, and this shows that for nearly every year during the past decade, the frequency is higher than the year before. Notice also the much higher frequency in Spoken, which is to be expected, and which also shows that COCA has a robust amount of informal spoken English. Finally, if we compare this to the data from the BoE and the OEC in Tables 3 and 7, we see that the frequency of the construction in COCA is between three and sixty-six times as common as in the BoE (per million words) and about four to five times as common as the OEC, depending on the particular time period.

Consider now the ‘*so not*’ construction (*I’m so not interested in him*). As shown in Table 17, although the tokens for this construction are relatively

sparse (but still five to ten times higher than in the BoE), we still see the higher relative frequency of the construction in the more informal genres (spoken, fiction, and popular magazines), as well as the clear increase in the construction over time.

Let us now briefly consider two ‘prescriptive’ issues –*can/may* for permission, and the split infinitive. First, consider the data for *can* vs. *may* (cf. Facchinetti, 2000; Leech, 2003; Millar 2009), as measured by the frequency of the two strings *can I* and *may I*. As the data show, there is a steady shift away from the prescriptive rule (i.e. from *may I* to *can I*) during the past two decades (Table 18).

Consider as another prescriptive rule the split infinitive (to [verb] [Adv] > to [Adv] [Verb], e.g. *to go boldly* > *to boldly go*) (cf. Close, 1987). This is measured by the percentage of –ly adverbs (e.g. *boldly*, *quickly*) either before or after the infinitive following *to*. As can be seen, there is an increase in each five year block during the past two decades (Table 19).

To this point, we have looked at two salient, recent grammatical constructions and two fairly salient prescriptive rules. For these phenomena,





Table 18 Frequency of *may / can* (I), by time period

| | 1990–94 | 1995–99 | 2000–04 | 2005–09 | 1990–94 to 2005–09 |
|---------|---------------------|---------|---------|---------|--------------------|
| may I | 1223 | 855 | 768 | 674 | |
| can I | Top of Form 2976 | 3541 | 3027 | 3055 | Increase |
| % can I | 70.9% | 80.6% | 79.8% | 81.9% | 16% |

Table 19 Frequency of split infinitive (*to go boldly* > *to boldly go*) by time period

| | Search string | 1990–94 | 1995–99 | 2000–04 | 2005–09 | 1990–94 to 2005–09 |
|---------|------------------|---------|---------|---------|---------|--------------------|
| – split | to [v*] *ly,[r*] | 17675 | 15981 | 16124 | 13859 | |
| + split | to *ly,[r*] [v*] | 8068 | 9349 | 10419 | 10363 | Increase |
| % split | | 31.3% | 36.9% | 39.3% | 42.8% | 37% |

Table 20 Frequency of (end up V-ing) by time period, 1990s–2000s

| SECTION | 1990–94 | 1995–99 | 2000–04 | 2005–09 |
|-----------|---|---|---|--|
| |  |  |  |  |
| PER MIL | 20.1 | 24.3 | 25.3 | 30.0 |
| SIZE (MW) | 103.3 | 102.9 | 102.6 | 93.6 |
| FREQ | 2075 | 2505 | 2600 | 2803 |

however, sociolinguistic surveys or other means of gathering data might also be sufficient, since the speakers are quite aware of the phenomena. Where corpora really shine, however, is for the ‘lower level’ constructions where speakers themselves seem quite unaware of what is going on. To conclude this section on syntax, consider three more syntactic shifts in contemporary American English (from among many that we could choose): the rise in the ‘end up V-ing’ construction (*we’ll end up paying too much*), the increase in the ‘get passive’, and the shift from (help to V) to (help V).

First, Table 20 shows the increase in the ‘end up V-ing’ construction over the past two decades.

Notice that the normalized frequency increases in each five-year period since the early 1990s. In fact, this continues a trend that has been in progress for the last 80 to 90 years, as shown in data from

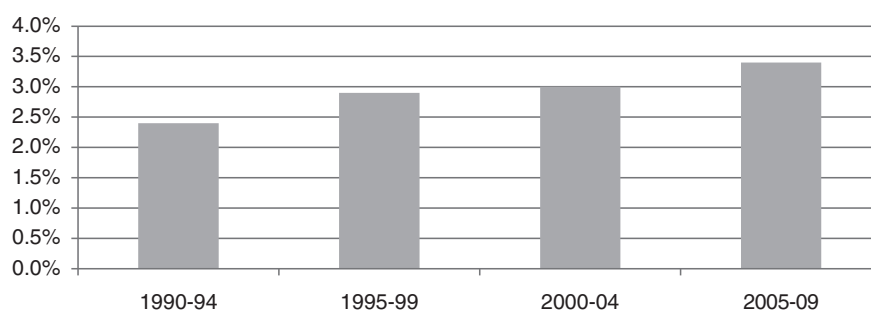
the 100 million word TIME Corpus of Historical American English (<http://corpus.byu.edu/time>) (Table 21).

The second low-level shift is the rise in the ‘get passive’ (*Bill got hired last week*, vs. *Bill was hired last week*) (cf. Hundt, 2001; Mair, 2006; Ruhlemann, 2007). Figure 1 was not produced directly by the COCA interface, but it is based on two searches in COCA (the *be* passive: [be] [vvn*] and the *get* passive: [get] [vvn*]). It shows the percentage of all passives (*be* or *get*) that occur with *get*. As one can see, the *get* passive steadily increases from one time period to the next, and the overall effect since the early 1990s is that the *get* passive has increased (compared with the *be* passive) more than 40% during this time.

The final low-level syntactic change is the slow but consistent shift from [help to V] to [help

Table 21 Frequency of (end up V-ing) by decade, 1920s–2000s

| SECTION | 1920s | 1930s | 1940s | 1950s | 1960s | 1970s | 1980s | 1990s | 2000s |
|----------------|------------|------------|------------|------------|------------|------------|-------------|-------------|-------------|
| | | | | | | | | | |
| PER MIL | 0.0 | 0.5 | 1.2 | 1.8 | 5.3 | 7.6 | 10.0 | 18.1 | 32.1 |
| SIZE (MW) | 7.6 | 12.7 | 15.5 | 16.8 | 16.1 | 13.6 | 11.4 | 9.7 | 6.4 |
| FREQ | 0 | 6 | 18 | 31 | 86 | 103 | 114 | 176 | 206 |

**Fig. 1.** Frequency of 'get passive' versus 'be passive' by time period, 1990s–2000s**Table 22** Frequency of [help to V / help V] by period, 1990s–2000s

| search string | | 1990–94 | 1995–99 | 2000–04 | 2005–09 | 1990–94 to 2005–09 |
|---------------|------------------------|---------|---------|---------|---------|--------------------|
| + | to [help] [p*] to [v*] | 5586 | 6501 | 7164 | 7202 | |
| - | to [help] [p*] [v*] | 841 | 809 | 728 | 634 | Increase |
| % | -to | 86.9% | 88.9% | 90.8% | 91.9% | 5.7% |

V] (*I'll help Mary to clean the room* > *I'll help Mary clean the room*), which is a change that has been commented on from a corpus-based approach by Kjellmer (1985), McEnery *et al.* (2005), among others (Table 22).

This data from COCA complements the data from the TIME Corpus, which also shows a slow but steady evolution towards the bare infinitive (*help him clean the room*) from the 1920s to the 2000s (Table 23).

4.4 Semantic change

Measuring semantic change is somewhat more difficult than measuring syntactic change. One way to

look for semantic change is to simply go through hundreds or thousands of lines of text, looking for new meanings and uses. However, it is also possible to simplify this and make it much quicker by looking for changes in collocates—the idea being that as a word changes meanings, the collocates (nearby words) may change as well.

To take an example that is simple to understand (though perhaps a bit trivial), consider the collocates of *web* in 1990–94 compared with 2005–09. The collocates for the earlier period are found to the left in Table 24, while those for the latter period are found to the right. This table clearly shows the emergence of the meaning of *web*

Table 23 Frequency of [help to V / help V] by decade, 1920s–2000s

| | 1920s | 1930s | 1940s | 1950s | 1960s | 1970s | 1980s | 1990s | 2000s |
|--------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| + to | 15 | 33 | 47 | 54 | 53 | 54 | 24 | 11 | 8 |
| – to | 73 | 214 | 316 | 369 | 287 | 303 | 270 | 391 | 363 |
| % – to | 0.83 | 0.87 | 0.87 | 0.87 | 0.84 | 0.85 | 0.92 | 0.97 | 0.98 |

Table 24 Comparison of collocates of *web*, 1990–94 and 2005–09

| Collocate | 1990–94 | 2005–09 | Ratio | Collocate | 2005–09 | 1990–94 | Ratio |
|-----------|---------|---------|-------|-----------|---------|---------|----------|
| Seamless | 18 | 0 | 17.42 | Site | 6453 | 1 | 7,126.42 |
| Strand | 15 | 2 | 6.79 | E-mail | 459 | 0 | 490.58 |
| Thread | 7 | 1 | 6.34 | Visit | 283 | 0 | 302.47 |
| Wrap | 6 | 1 | 5.43 | Page | 523 | 2 | 288.79 |
| Passage | 5 | 0 | 4.84 | Check | 172 | 1 | 189.95 |
| Silver | 5 | 1 | 4.53 | Address | 164 | 1 | 181.11 |
| Intrigue | 7 | 2 | 3.17 | Post | 149 | 0 | 159.25 |
| Surround | 6 | 2 | 2.72 | Search | 145 | 0 | 154.98 |
| Delicate | 5 | 2 | 2.26 | Surf | 116 | 1 | 128.11 |
| Economic | 5 | 2 | 2.26 | Browser | 114 | 1 | 125.9 |

Table 25 Changes in meaning with *green* = ‘environmentally friendly’, 1990s–2000s

| Collocate | 2005–09 | 1990–94 | Ratio |
|-------------|---------|---------|-------|
| Economy | 34 | 0 | 36.30 |
| Benefits | 29 | 1 | 32.03 |
| Jobs | 72 | 3 | 26.50 |
| Community | 19 | 1 | 20.98 |
| Sustainable | 18 | 1 | 19.88 |
| Investment | 18 | 0 | 19.24 |
| Successful | 15 | 1 | 16.57 |
| Solutions | 15 | 0 | 16.03 |
| Technology | 43 | 3 | 15.83 |
| Cities | 14 | 1 | 15.46 |

referring to the Internet, compared with the earlier meaning of *spider web* or ‘connection between parts’.

A somewhat less trivial example is found in Table 25, where we see the collocates of *green* that are much more common in 2005–09 than in 1990–94, and these relate to the relatively new meaning of ‘environmentally friendly’.

One of the unique features of COCA is that it has an integrated thesaurus with entries for more than 60,000 words. Using this semantic information,

users can easily compare the frequency of words in entire semantic fields across time. To do such a search, users simply enter [=word] in the search form (where *word* is the synonym set that they are interested in). If they choose to see the frequency by time period and genre, they will see something like in Table 26 (the color represents the normalized frequency—per million words—and users can also see that number displayed in the results).

By simply selecting to the two competing time periods, users can also see exactly which synonyms are more common in one time period (or genre) than another. For example, Table 27 shows the synonyms of *beautiful* that are at least 10% more common (per million words) in the period 2005–09 and also 1990–94. Even though only 10–15 years have passed since the early 1990s, native speakers probably do have at least a vague sense that the words on the right (the words from the early 1990s) do sound somewhat more ‘old-fashioned’ than those on the left.

4.5 Discourse analysis

Closely related to semantic change is the issue of discourse analysis, or what it is that we are saying

Table 26 Frequency of synonyms of *beautiful* by genre and time period

| Word(s) | SPOK | FIC | MAG | NEWS | ACAD | 1990–94 | 1995–99 | 2000–04 | 2005–07 |
|------------|-------|-------|------|------|------|---------|---------|---------|---------|
| Beautiful | 8297 | 13004 | 7809 | 4776 | 2518 | 8588 | 10552 | 10734 | 6530 |
| Wonderful | 10688 | 4848 | 4601 | 4160 | 1226 | 6821 | 8193 | 6934 | 3575 |
| Attractive | 1308 | 1838 | 2922 | 1939 | 2224 | 3083 | 2849 | 2697 | 1602 |
| Striking | 1071 | 1128 | 2274 | 1883 | 2600 | 2645 | 2557 | 2425 | 1329 |
| Lovely | 1584 | 3629 | 1642 | 999 | 278 | 2119 | 2450 | 2267 | 1296 |
| Handsome | 436 | 3452 | 1464 | 975 | 313 | 1830 | 1815 | 1853 | 1142 |
| Charming | 622 | 1619 | 1028 | 911 | 215 | 1158 | 1132 | 1199 | 906 |
| Gorgeous | 947 | 1075 | 1028 | 598 | 70 | 738 | 952 | 1132 | 896 |
| Superb | 332 | 197 | 1089 | 715 | 289 | 871 | 687 | 703 | 361 |
| Scenic | 71 | 123 | 941 | 661 | 238 | 596 | 587 | 544 | 307 |
| Exquisite | 106 | 564 | 708 | 352 | 241 | 527 | 553 | 550 | 341 |

Table 27 Frequency of synonyms of *beautiful*, 2005–09 and 1990–94

| Word(s) | 2005–09 | 1990–94 | RATIO | WORD | 1990–94 | 2005–09 | RATIO |
|--------------|---------|---------|-------|--------------|---------|---------|-------|
| Gorgeous | 1413 | 748 | 2.09 | Superb | 875 | 567 | 1.40 |
| Stunning | 1393 | 1047 | 1.47 | Striking | 2657 | 2009 | 1.20 |
| Beautiful | 9868 | 8700 | 1.25 | Wonderful | 6908 | 5341 | 1.17 |
| Charming | 1273 | 1173 | 1.20 | Attractive | 3121 | 2421 | 1.17 |
| Good-looking | 383 | 361 | 1.17 | Magnificent | 1113 | 870 | 1.16 |
| | | | | Scenic | 597 | 470 | 1.15 |
| | | | | Delightful | 479 | 383 | 1.13 |
| | | | | Fine-looking | 23 | 19 | 1.10 |

about Topic X at a given point in time. It may not be that a word itself has ‘changed meaning’, but rather that it is being used in conjunction with a new set of collocates (cf. Stubbs 1996). For example, the following are words that are used much more with *gay* or *gays* in 2005–09 compared to those from 1990 to 2004. In the period 2005–09, we see much more mention of issues related to gay marriage and civil unions, as well as collocates like *cowboys* (related to the movie *Brokeback Mountain*) and *gun* (related to conservative efforts on behalf of *gun control* and against *gay marriage*) (Table 28).

Perhaps even further away from strictly linguistic issues are queries that relate to current events, but where the corpus still gives interesting insight into changes in American history and culture. For example, the Table 29 shows collocates that are used more with *crisis* in the period 2005–09 than in 1990–94.

Table 28 Change in collocates with *gay/gays* between 1990–94 and 2005–09

| Collocate | 2005–09 | 1990–94 | % |
|----------------|---------|---------|-------|
| Marriage | 469 | 8 | 64.74 |
| Transgender | 36 | 0 | 38.48 |
| Unions | 22 | 0 | 23.51 |
| Constitutional | 38 | 2 | 20.98 |
| Amendment | 19 | 1 | 20.98 |
| Cowboys | 13 | 0 | 13.89 |
| Gun | 12 | 1 | 13.25 |
| Adoption | 12 | 0 | 12.83 |
| Same-sex | 11 | 0 | 11.76 |
| Marriages | 20 | 2 | 11.04 |

These collocates for 2005–09 compare with earlier (1990–94) collocates like *Gulf*, *Persian*, *Soviet*, *Kuwait*, *Saddam*, *Hussein*, *Gorbachev*, *Yugoslavia*, and *Bosnia*, which related to crises from 15 to 20

Table 29 Change in collocates with *crisis* between 1990–94 and 2005–09

| Collocate | 2005–09 | 1990–94 | % |
|-------------|---------|---------|-------|
| Mortgage | 80 | 0 | 85.50 |
| Credit | 138 | 3 | 50.80 |
| Subprime | 43 | 0 | 45.96 |
| Foreclosure | 42 | 0 | 44.89 |
| Darfur | 34 | 0 | 36.34 |
| Asian | 46 | 2 | 25.40 |
| Climate | 29 | 2 | 16.01 |
| Muslim | 14 | 0 | 14.96 |
| Lending | 11 | 1 | 12.15 |
| Obesity | 10 | 0 | 10.69 |

years ago. This should give just a small sense of how the data from COCA can be used to look at an extremely wide range of phenomena, some of them transcending purely linguistic issues.

5. Summary

As we have shown, the COCA is the first reliable ‘monitor corpus’ of English, and it provides data for ongoing changes in English that are not available from any other source. In fact, it is the first large-scale monitor corpus of any language, which is balanced between a number of different genres. Its usefulness as a monitor corpus is a function of at least two different aspects of the corpus.

The first aspect is of course the data itself. In terms of the textual corpus—the articles and books in the corpus—it was designed from the ground up as a monitor corpus. There are other corpora, such as the BNC, which are quite useful in their own right, but which do not have any real diachronic dimension. There are also corpora like the Brown family (Brown, LOB, Frown, FLOB), which attempt to show changes between recent decades in English, but which are far too small to provide useful data on many types of linguistic change. Recall also that COCA is the only corpus that continues to be updated. Work on the BoE ended in 2005 and work on the OEC stopped in 2006, while twenty million words have been added to COCA

every year since 1990, with the most recent additions being just one month ago.

As we have discussed, an important aspect of the corpus design is its genre balance. The fact that it does have data from a number of different genres also sets it apart from ‘text archives’, which are typically just one genre (online newspapers or magazines), or a collection of texts that are difficult to categorize and separate by genre (e.g. the Web). In addition, COCA has almost exactly the same genre balance from year to year, which means that we can compare data across different years and time periods and be quite certain that the corpus models accurately what is happening in the real world. As we have discussed at some length, this is quite different from the BoE or the OEC, which vary widely in genre composition from one period to the next, and which therefore provide some unreliable data in terms of frequency comparisons across time periods.

Beyond the textual corpus, however, an important aspect of the COCA is the corpus architecture and interface. Unlike text archives and search engines like Google, users of COCA can search by substring, lemma, part of speech, collocates, synonyms, and limit and compare by sections of the corpus. The COCA interface allows users to see the frequency of all matching strings in each section of the corpus (e.g. Table 26), as well as to directly compare the frequencies in different sections of the corpus (there are many examples of this in the paper, such as Tables 28 and 29). No other architecture for large corpora—such as Corpus Workbench or Sketch Engine—has an architecture and interface that can process the data and display the results in this way. In addition, the corpus architecture is very robust. Nearly all of the queries discussed here take just two or three seconds to search and display results from the 400+ million word corpus.

The end result of all of this, then, is that with the COCA we now have a unique tool that we can use to easily and quickly map out and study historical changes in contemporary English. Courses on historical linguistics typically focus on changes that occurred two hundred or two thousand years ago, but they often do a much poorer job talking about

ongoing change—because until now, they have not had reliable monitor corpora on which to base their data. With the COCA, however, teachers, students, and researchers can bring linguistic change right up to the present time, and thus study the way in which the language is changing in ways that are not possible with any other resource.

References

- Baker, D., Hardie, A., and McEnery, T.** (2006). *A Glossary of Corpus Linguistics*. Edinburgh: Edinburgh University Press.
- Barbieri, F.** (2009). Quotative 'be like' in American English: ephemeral or here to stay? *English World-Wide*, 30: 68–90.
- Biber, D., Johansson, S., Leech, G., Conrad, S., and Finnegan, E.** (1999). *Longman Grammar of Spoken and Written English*. London: Longman.
- Buchstaller, I. and D'Arcy, A.** (2009). Localized globalization: a multi-local, multivariate investigation of quotative 'be like'. *Journal of Sociolinguistics*, 13: 291–331.
- Burnard, L.** (2002). Where did we go wrong? A retrospective look at the British National Corpus. In Kettemann, B. and Markus, G. (eds), *Teaching and Learning by Doing Corpus Analysis*. Amsterdam: Rodopi, pp. 51–71.
- Close, R.** (1987). Notes on the split infinitive. *Journal of English Linguistics*, 20: 217–29.
- Davies, M.** (2005). The advantage of using relational databases for large corpora: speed, advanced queries, and unlimited annotation'. *International Journal of Corpus Linguistics*, 10: 301–28.
- Davies, M.** (2009a). Relational databases as a robust architecture for the analysis of word frequency. In Archer, D (ed.), *What's in a Wordlist?: Investigating Word Frequency and Keyword Extraction*. London: Ashgate, pp. 53–68.
- Davies, M.** (2009b). The 385+ million word corpus of contemporary American English (1990–2008+): design, architecture, and linguistic insights. *International Journal of Corpus Linguistics*, 14: 159–90.
- Facchinetti, R.** (2000). Be able to in present-day British English. In Mair, C. and Hundt, M. (eds), *Corpus Linguistics and Linguistic Theory*. Amsterdam: Rodopi, pp. 117–130.
- Gardner, D. and Davies, M.** (2007). Pointing out frequent phrasal verbs: a corpus-based analysis. *TESOL Quarterly*, 41: 339–59.
- Hundt, M.** (2001). What corpora tell us about the grammaticalisation of voice in get-constructions. *Studies in Language*, 25: 49–87.
- Hunston, S.** (2002). *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.
- Kjellmer, G.** (1985). Help to / Help Ø revisited. *English Studies*, 66: 156–61.
- Leech, G.** (2003). Modality on the move: the English modal auxiliaries 1961–1992. In Facchinetti, R., Krug, M., and Palmer, F. (eds), *Modality in Contemporary English*. Berlin: Mouton de Gruyter, pp. 224–40.
- Mair, C.** (2006). Tracking ongoing grammatical change and recent diversification in present-day standard English: the complementary role of small and large corpora. In Renouf, A. and Kehoe, A. (eds), *The Changing Face of Corpus Linguistics*. Amsterdam: Rodopi, pp. 355–76.
- Martin, P.** (1991). *The Phrasal Verb: Diachronic Development in British and American English*. Ph.D. thesis, Columbia University.
- McEnery, T. and Wilson, A.** (2001). *Corpus Linguistics: An Introduction*. 2nd edn. Edinburgh: Edinburgh University Press.
- McEnery, T. and Xiao, Z.** (2005). Help or help to: What do corpora have to say? *English Studies*, 86: 161–87.
- McEnery, T., Xiao, R., and Tono, Y.** (2006). *Corpus-Based Language Studies: An Advanced Resource Book*. London: Routledge.
- Meyer, C.** (2002). *English Corpus Linguistics: An Introduction*. Cambridge: Cambridge University Press.
- Millar, N.** (2009). Modal verbs in TIME: frequency changes 1923–2006. *International Journal of Corpus Linguistics*, 14: 191–220.
- Ruhlemann, C.** (2007). Lexical grammar: the GET-passive as a case in point. *ICAME Journal*, 31: 111–27.
- Sampson, G. and McCarthy, D.** (2004). *Corpus Linguistics: Readings In A Widening Discipline*. London: Continuum.
- Stubbs, M.** (1996). *Text and Corpus Analysis*. Oxford: Blackwell.
- Tagliamonte, S. and D'Arcy, A.** (2004). He's like, she's like: the quotative system in Canadian youth. *Journal of Sociolinguistics*, 8: 493–514.

Notes

1. The one-way around problem of uneven genre distribution in the BoE is to divide the corpus into ten to twenty different sections (e.g. US-Fiction, UK-Fiction, US-Newspapers, etc.), and then look at changes in each of these sections individually. However, this seems to be a rather cumbersome solution.
2. There are monitor corpora of some other languages, which are composed strictly of newspapers, such as the Norwegian Newspaper Corpus (University of Bergen),

but there are no monitor corpora that are composed of a wide range of genres. There are also corpora like the CREA Corpus of Contemporary Spanish (<http://corpus.rae.es/creanet.html>) which do contain a wide range of genres, but which have such a limited architecture and interface (no part of speech tagging, lemmatization, substrings, or collocates searching, for example) that they are much more like a text archive than a true corpus.