Observing recent change through electronic corpora

CHAPTER 12

SOME METHODOLOGICAL ISSUES RELATED TO CORPUS-BASED INVESTIGATIONS OF RECENT SYNTACTIC CHANGES IN ENGLISH

MARK DAVIES

1. Introduction

This group of chapters deals with "observing recent change" in English, which is a topic that has received renewed attention of late (cf. Leech et al. 2009). Some of the chapters deal with changes that have taken place primarily in the past 50 years or so, whereas others extend back to the 1800s. In all cases, the changes are still underway and they exhibit interesting genre-based and stylistically based variation in contemporary English. All of the chapters in the section focus primarily on syntactic change. This is in part a function of the resources used to look at recent change; lexical and semantic changes are somewhat more difficult to study with small corpora such as the BROWN family (see section 4 below).

This introduction focuses primarily on methodological issues—especially what types of corpora and text archives can provide us with the most useful data for looking at recent change in English. I first consider large text collections like Google, Google Books, and large text archives. I then consider "structured" corpora and focus in some detail on the question of size—especially the way in which large corpora like the Corpus of Historical American English provide insight that might not be available otherwise. Finally, I conclude with a discussion of "monitor corpora" like the Corpus of Contemporary American English, which are continually added to in order to monitor ongoing changes in the language.

2. Google, Google News Archive, Google Books, and similar tools

Although Google (and other search engines) are useful for finding web pages on a particular topic, they are much more limited in terms of linguistically oriented searches.[1] Indeed, it is challenging to look for anything beyond simple words and phrases. In their native format, it is not possible to search for substrings, thus making it hard to look at morphological change. These queries cannot extract collocates or use semantic information from thesauruses as part of the native queries, which makes it very difficult to look at semantic change. In terms of syntax, we cannot search for lemma or for part of speech. This means that we are reduced to looking at exact strings of words, which serve as proxy for entire constructions. In addition, as is fairly widely known (e.g. Kilgarriff 2007), Google counts for anything other than single words are often widely inaccurate, since Google simply "estimates" the frequency of multi-word strings. For example, it estimates the frequency of the phrase *would be taken for a* to be more than 2,000,000 hits, when in fact there are less than 500 total, once one actually starts paging through the hits.

When we turn to looking at evidence for recent linguistic changes, we also find that—while it is possible to limit queries to the last week, month, or year—there is really no way to limit searches of the Web to particular time periods like 1996–2000 or the year 2007. Thus, it is difficult or impossible to compare one period to another and see whether a linguistic feature is increasing or decreasing. Typically, we search the Web for a string of words X at time T and we can say that we found N examples of X, or (perhaps more advanced) string X is more common than Y. But again, it is nearly impossible to show how much X is increasing over time.

Although Google searches of the Web are problematic for looking at language change, some might be encouraged by Google Books. Via the new Google Books

1 See Mair (19) for a good overview of issues relating to Web-based data.

interface announced in late 2010,[2] users can search *hundreds of billions* of words of text in several historical "corpora" of books and periodicals. While this is an incredibly useful tool for looking at the frequency of exact words and phrases, it cannot do much beyond that. As with Google searches of the Web, research on diachronic syntax with Google Books is very difficult, since the corpus is not lemmatized or tagged for part of speech. For example, to look at changes with the [into V-ing] construction (*Bill talked Sue into staying*), the only part of the string that Google Books can find is the word *into*, which of course massively over-generates the results. In addition, as with Google searches of the Web, Google Books cannot efficiently find collocates, to look at semantic change.[3]

In summary, as promising as Google and Google Books might seem at first glance, there are fundamental weaknesses with these tools, which makes them of marginal usefulness for looking at recent changes in English.

3. Text archives

If the Google resources are inadequate, perhaps what we need is a somewhat more structured "text archive". Linguists can now search billions of words of text in text archives like Project Gutenberg or the Internet Archive (for books), Making of America (for magazines), or the increasing number of archives of historical newspapers.[4] Via their native interfaces, most of these text archives suffer from the same limitations as we saw with Google. Users can do simple searches for exact words or phrases, but not substrings, part of speech, lemmas, or collocates, which seriously limits their usefulness for looking at morphological, syntactic, or semantic change. But some of these words or phrases can be accurately searched by time period, to compare the frequency across different years or decades. In order to do this, however, users need to create their own "normalization" scheme by finding the frequency of a common word like *is*, *put*, or *if*—which supposedly would not change much over (recent) time. They could then compare the frequency of a given word or phrase in different periods against this "base frequency" item to get quasi-"normalized" results.

A useful alternative to using the text archive materials via their native interface is to download millions or tens of millions of words of text from these archives and then analyze them on their own computers. This is precisely the approach used by De Smet (e.g. 2008) and Cuyckens and De Smet (e.g. 2007) with the Corpus of

2 See http://ngrams.googlelabs.com/.

3 For an extended comparison of Google Books and structured corpora, see http://corpus.byu.edu/coha/compare-googleBooks.asp.

4 See http://www.gutenberg.org/, http://www.archive.org/, and http://quod.lib.umich.edu/m/moagrp/.

Late Modern English Texts (CLMET) which has been created at the University of Leuven; Rohdenburg (e.g. 2009) and Vosberg (e.g. 2003) of the University of Paderborn for their texts from the 1800s to 1900s; Rudanko in his several studies (e.g. 2006); and Mair (e.g. 2006a), who uses a wide range of text archives to look at changes in the 1900s. Davies has created a 100 million word corpus of articles from TIME Magazine, which has been used for a number of studies, including Rudanko (17). As we will see in sections 5–6—where I discuss two particular types of recent changes in verbal complementation in English—these text archives have yielded extremely useful data that has led to insightful analyses.

There are four potential problems with these text archives, however. The first is the accuracy of the data, especially with data that has been extracted from "messy" PDF or other image files. The second is accurately annotating the texts for date and genre, which is sometimes more difficult than one would imagine. For example, Project Gutenberg does not list the date for any of the more than 20,000 books in English from the 1800s to 1900s; researchers would have to find these dates themselves (and deal with issues like which version of the book is in Project Gutenberg). Third, researchers would need to develop their own architectures and interfaces to allow for a wide range of searches, rather than just simple words and phrases. Fourth, a fundamental problem with using text archives is that each researcher or group of researchers tends to create their own corpus. These corpora are typically not publicly available, and therefore their data cannot be reviewed or analyzed by others, which impacts on the issue of "falsifiability"—a fundamental tenet of the scientific method.

The final issue with collections based on text archives is genre balance. Most collections are "unbalanced" between genres, especially over time, and this can cause serious problems. For example, if most of the texts from the 1950s to 1970s are from newspapers, and then the percentage of fiction in the corpus from the 1980s to 2000s is much higher, it is difficult to know if any changes we find between the two time periods are due to changes in the corpus composition (and therefore relatively meaningless artifacts), or whether they reflect actual changes in the language. Ideally, a corpus would maintain roughly the same genre (and subgenre) balance from one period to another.

4. Structured corpora

4.1 Previous corpora

We saw in section 2 that search tools like Google, Google News Archive, and Google Books initially appear attractive for looking at language change, in large part because of their size. However, it is only with text archives (section 3) that the data becomes structured enough that we can make sense of the data and begin to

compare linguistic features in different time periods. However, with text archives we are still faced with the issues raised in the previous two paragraphs: accuracy, annotation, architecture, availability, and balance.

The third approach, then, has been to create carefully constructed "structured" corpora. Until recently, all of these structured corpora were relatively small. Perhaps the most famous structured historical corpus that has been created to look at recent change is the BROWN family of corpora—one million words each in Brown (US 1960s), LOB (UK 1960s), Frown (US 1990s), and FLOB (UK 1990s) (see e.g. Mair 1997 and Hundt and Leech, 13). Another corpus is the 1.8 million word ARCHER Corpus (A Representative Corpus of Historical English Registers), which in Version 3.1 now contains a little more than 500,000 words from 1950 to 1990 (see Biber, Finegan, and Atkinson 1994, as well as Curzan, 16, and Hundt and Leech, 13). Going back a bit more in time, the CONCE Corpus (Corpus of Nineteenth Century Texts) contains about 1,000,000 words from the United Kingdom in the 1800s (see Kytö, Rudanko, and Smitterberg 2000). The Diachronic Corpus of Present-Day Spoken English (DCPSE) compares two small corpora of spoken British English from the 1950s and 1990s. (On its use as a historical corpus, see Bowie and Aarts, 15, and for a contrasting view, see Davies 2009b.) Finally, there have been insightful studies based on the corpora from different countries in the International Corpus of English (ICE; cf. Mukherjee and Schilk, 14). Although the ICE corpora are not diachronic per se, researchers can still extrapolate from this data and use it to "triangulate" recent changes in English.

In most of these cases, the small corpora maintain almost perfect genre balance between the different time periods in the corpus, and they are therefore very useful for looking at change over time. In addition, because they are small and manageable, they have a high degree of textual accuracy, and they are very well annotated for date and genre. With some of these corpora, however, there are issues relating to availability (as with ARCHER, its use is limited to just a handful of universities) and architecture (i.e. they are just collections of simple text files).

4.2 The Corpus of Historical American English (COHA)

In 2010, a new historical corpus of English came online, which I would argue has nearly all of the advantages of these smaller corpora, but which is also much larger and which also has a more advanced architecture and interface. The Corpus of Historical American English (COHA) contains more than 400 million words from the 1810s to 2000s. In terms of *balance* (one of our five main objectives), its genre and subgenre balance stays almost identical from decade to decade. Regarding *accuracy*, great care has been taken as the PDF and scanned materials have been converted to text. We have run complex algorithms to find and eliminate texts from low-quality scans, and students have spent thousands of hours looking for and correcting words whose frequency was suspiciously high (compared to 100

percent clean texts in other corpora). Regarding *annotation*, the 100,000+ texts are very well annotated for year and genre, and the corpus is also well annotated at the word level for lemma and part of speech. In terms of *availability*, it is freely available, so the studies done by one researcher can be easily checked by others.

In terms of the fifth objective—*architecture and interface*—the Web-based COHA interface (which is the same as the other corpora from http://corpus.byu.edu) allows for research on a much wider range of phenomena than most other historical corpora. It allows complex searching by part of speech and lemma (for syntactic change), substrings (for morphological change), and collocates, synonyms, and customized word lists (for semantic change). Users can see the frequency of any word, phrase, grammatical construction, or collocate across each decade from the 1810s to 2000s (either by individual matching string or charts showing overall frequency). In addition, the architecture allows users to quickly and easily compare features in two different time periods, such as collocates of *woman* or *catch* in the 1850s to 1890s compared to the 1950s to 2000s, words with the suffix *-ism* in the 1970s to 1980s vs. the 1990s to 2000s, or all adjectives that are much more common in the 1980s to 1990s than in the 1920s to 1930s (and vice versa).

4.3 Comparing larger and smaller corpora

As noted above, COHA is quite different from the previous corpora of recent English in terms of size. Over the past 15–20 years, the smaller corpora have been used to look primarily at high frequency syntactic constructions, such as auxiliaries, modals, relative pronouns, or prepositions (see e.g. the studies in Leech et al. (2009), Hundt and Leech, 13, and Bowie and Aarts, 15). These small corpora have led to many highly insightful studies of these constructions, which have greatly enhanced our view of underlying patterns and shifts in Late Modern English.

Large corpora like COHA, however, allow us to expand our scope, in terms of the phenomena that we can look at. First, they allow us to look at lexical and semantic change, where there may only be 5–10 tokens per million words of text. Second, large corpora allow us to look at medium and low frequency constructions, such as verbal complementation or other lexically driven phenomena. I provide examples of two such phenomena in sections 5–6 below, and Hilpert (18) and Curzan (16) look at two other lower frequency constructions as well.

The number of tokens is also related to the "life cycle" of a linguistic change. As a change is first occurring in a language, there may only be a small handful of tokens—perhaps one or two tokens per 100 million words of text. For example, the still-awkward construction [*have been being V-ed*] (this issue had been being discussed for some time) occurs only 2 times in the 100 million word British National Corpus (BNC) and only 14 times in the 410 million word Corpus of Contemporary American English (COCA). Similarly, a feature may occur with very low frequency at the very end of its "life cycle". For example, consider post-verbal negation with certain modals. The forms *mightn't, oughtn't, mayn't, and daren't* occur only 64, 56,

10, and 9 times (respectively) in the 410 million word COCA corpus (and somewhat more frequently in the BNC).

With a small 4 million word corpus (e.g. the size of the BROWN family of corpora)[5] there would be (all other things being equal) one hundredth the number of tokens—or in the two cases just shown, virtually none for the incipient [*have been being V-ed*] or the recessive post-verbal negation. Thus, small corpora may at times limit us to looking primarily at high frequency constructions in the middle of their life cycle (or toward the end of the cycle, for high frequency forms). Large corpora, on the other hand, allow us to look at high, medium, and low frequency phenomena, in all stages of the "life cycle" of a linguistic change.

4.4 Challenges with large corpora

While large corpora provide certain advantages, as we have seen, some researchers (cf. Hundt and Leech, 13) suggest that the same is true for small corpora, which "are not only a valuable but arguably an indispensable part of the corpus linguist's toolbox" (p. 187). They suggest that small corpora such as the BROWN family of corpora have the following advantages over large corpora: (1) more careful genre sampling; (2) more accurate part of speech tagging; (3) ability to manually examine results (since there are fewer to look at); and (4) full-text access.

Although these advantages of smaller corpora may relate to some large corpora, I would suggest that they may not be as relevant for COHA, for the following reasons: (1) the genre balance is almost exactly the same from decade to decade; (2) the corpus is tagged with CLAWS (the same tagger used for the BNC) and tagging for nearly 100,000 "problematic" types from the 1800s has been manually reviewed and corrected; (3) results can be "thinned" to a series of 100–1,000 random hits; and (4) users can see context of up to 180 words, which should be sufficient for nearly all queries.

As evidence for the advantage of small corpora, Hundt and Leech look at the decrease in frequency with the relativizer *which*, as well as the decrease with *for* as a conjunction. We note that while BROWN certainly provides convincing data for these shifts, COHA shows essentially the same changes as well.[6] In addition, COHA has the advantage of showing the frequency in each decade (not just every 30 years), and it also covers a much wider time period (200 years, compared to just 30–60 years in BROWN).

4.5 Large corpora and verbal complementation

In the two sections that follow, I provide extended discussion of how COHA can be used to look at recent syntactic changes in American English, in ways that are not

5 Upper-case "BROWN" is used to refer to the four corpora: Brown, LOB, Frown, and FLOB.

6 See http://corpus.byu.edu/coha/compare-smallCorpora.asp.

possible with smaller corpora. In section 5, I look in some detail at the spread of the [V NP into V-ing] construction (e.g. *she talked me into buying it*), and in section 6, I consider the [to-V] to [V-ing] shift with several verbs (e.g. *he started to walk > he started walking*).

In both cases, I will compare COHA to BROWN. There are at least three reasons for doing so. First, neither ARCHER nor CONCE nor the DCPSE is freely available, and so it would be difficult for most researchers to replicate these searches. Second, over the past 10–15 years, there has been comparatively more research done with BROWN than with the other three corpora, and so this is the "prototypical" contrast between small and large corpora. Third, because these are relatively recent changes, there is much more data in BROWN (mainly the second half of the 1900s) than in ARCHER or CONCE (which include texts from earlier centuries, when the constructions were nonexistent or quite infrequent).

5. V NP into [V-ing]

The construction [V NP into V-ing] (*we talked Bill into staying*) has received a fair amount of attention over the past decade or so, although most of the studies have been strictly synchronic. Rudanko (2005, 2006), however, briefly looks at the construction in the BROWN family of corpora and suggests that the construction is expanding its scope in English. But other than a short table with frequencies in the four corpora in the BROWN family, we have little sense of what has happened with the construction throughout the rest of the 1800s to 1900s.

The BROWN family of corpora provide a total of 29 tokens (6 in US-1960s and 3 in UK-1960s, compared to 11 in US-1990s and 9 in UK-1990s), which hints at an increase from at least the 1960s to 1990s. With 400 million words, the COHA corpus confirms this hypothesis—[V NP into V-ing] is increasing and has been since the early 1800s (and especially during the past 100 years). As Table 1 shows, it has more than doubled in frequency ("freq"; per million words) from about 1900 to the current time.[7]

In addition to what is motivating the overall increase with this construction, an equally interesting question is how the construction is spreading from one verb to another via analogy. With the sparse data from BROWN, there are just 21 matrix verbs, and none of them are overly surprising. Most are fairly basic verbs of influence or force: *talk, fool, goad, gull, terrify, entice, nag, shame, harass, coax, seduce, force, charm, coerce, deceive, hound, persuade, pressure, bluff, spur, cow.*

7 Note that "size" = size of COHA for that decade, in millions of words. Note also that for reasons of space here, only every other decade is displayed, but all are viewable in the online corpus.

Table 1. Overall frequency of *V NP into V-ing*, 1810s–2000s (COHA)

	1820	1840	1860	1880	1900	1920	1940	1960	1980	2000	TOTAL
size	7.3	17.3	18.2	20.7	22.3	25.9	24.5	24.2	25.6	31.5	406.7
tokens	6	30	57	88	108	197	207	207	266	438	3,012
freq	0.8	1.7	3.1	4.3	4.8	7.6	8.5	8.6	10.4	13.9	

In COHA, on the other hand, there are 322 different matrix verbs that take [into V-ing]. The following are some of the more interesting ones, and they are listed by the decade in which they first occur.[8]

> 1840s *cheat, quicken, wake* 1850s *irritate, natter, starve* 1860s *dazzle, flog, materialize* 1870s *daunt, plague, whip, worry* 1880s *gull, reason, stimulate* 1890s *bulldoze, hoodwink, thrash* 1900s *astonish, charm, cozen, jolly, storm* 1910s *freeze, bamboozle, chivy, jar, prick* 1920s *stampede, cheer, devil, nerve, strangle* 1930s *egg, jockey, laugh, tickle, vamp* 1940s *engineer, horsewhip, massage, wangle, wheel* 1950s *brainstorm, damn, outwit, stare, stiffen* 1960s *bait, blackjack, fox, pervert, scowl, service* 1970s *anesthetize, filibuster, knee, rassle, sting* 1980s *dog, hammer, jog, kiss, subsidize, witch* 1990s *boink, discombobulate, euchre, jiggle, romance, sphroxify* 2000s *annoy, chitchat, dose, jade, scam, suck*.

Due to space constraints I will not examine here the details of the semantic extension of the matrix verbs, such as:

- whether certain types of control have become more common (compare the idea in Wulff, Stefanowitsch, and Gries (2007) that the matrix verbs in British English tend to represent physical force more, while those in American English relate more to persuasion)
- when the romance-related uses arose (e.g. *charm/smooch someone into doing something*), or
- whether the metaphorical extension of physical force has increased or decreased in American English (e.g. *drive, push, pound, elbow, drill, move, nudge, budge, jar*).

However, with 322 types and 3,012 tokens over a 200-year period (compared to 29 tokens with 21 types in the 1960s and 1990s with the BROWN corpora), it would, of course, be possible to look at questions like these, and the data would likely give us valuable insight into the role that prototypes have played in the development of the construction (cf. Gries and Stefanowitsch 2003).

In terms of semantics, we might briefly consider an issue that may relate to the origin and initial extension of the [V NP into V-ing] construction. In the early 1800s there are many cases of [V NP into N]: *he bullied himself into power, you have driven him into exile, he was carrying it into effect*, etc. In the earliest stages of [V NP into V-ing], a high

8 Due to space constraints here, I cannot provide examples for these, but all have been manually checked and can be found online in the COHA corpus.

percentage of all tokens occur with the subordinate clause verb *being* (*called them into being, start others into being, brought this banquet into being, quickened such intention into being*, etc.). Note that *being* is a semantically rather simple verb and—although it can be analyzed as a verb in these case—it also has a strong nominal feel to it. Rather than have the construction created "ex nihilo", it apparently started where [V NP into V-ing] would be least noticeable—where [into V-ing] could also be analyzed as a noun, as with the pre-existing [into N]. And then once the [into V-ing] construction was firmly "established" in about the 1850s, the percentage of *being* decreased markedly.

6. [TO-V] VS. [V-ING]

The synchronic alternation between [to-V] vs. [V-ing]: (*he started* [*to walk/walking*] *down the street*) has been discussed at length in a number of articles and books during the past decade or two. What has been studied somewhat less is the historical development of these two constructions. Previous studies include Rohdenburg (2009), Vosberg (2003), Mair (e.g. 2002, 2006a), and Rudanko (2000). One problem, however, is that because each study looks at different verbs using different corpora, it is very difficult to get an overall picture of the changes in the 1800s to 1900s.

The general suggestion in these studies, however, is that there has been an overall shift from [to-V] to [V-ing] over time, relating to what Rohdenburg (2007, 2009) has called the "Great Complement Shift" in English. The COHA data allow us to look at this change as a *series* of shifts from [to-V] to [V-ing] with different matrix verbs, one after another. The sequencing of these micro-level shifts in the overall Great Complement Shift can provide us with important clues about what may have been driving the overall shift, as we consider why some verbs changed before others (e.g. *start, begin, continue, try, love, prefer, bother*). Did higher frequency verbs shift before lower frequency verbs? Did certain semantic classes (e.g. aspectual verbs, or verbs of emotion) lead the way?

Let us now examine what type of data we get from something small like the BROWN family of corpora. For each of the four corpora (Brown, LOB, Frown, FLOB), Table 2 shows the number of tokens of [to-V]:[V-ing].[9]

Two important things stand out in the BROWN data. First, the data is too sparse to provide statistically significant values. For example, it does seem that

9 For example, there are 50 tokens of [start to V] in the Brown corpus, and 52 tokens of [start V-ing]. The column labeled "American" shows the overall percentage of [V-ing] with that verb over time, e.g. 51 percent of the tokens with *start* are [V-ing] in the 1960s (Brown), and this increases to .61 in the 1990s (Frown), and the data in the "British" column works the same way. Finally, the "X^2 (p)" column gives the p value from the chi-square test for the American shift. For example, for the shift with *start* in Brown and Frown, the p value is .23, which is not statistically significant. (I have only calculated the p value and chi-square for the two American corpora, since that is what I will compare to the (American) COHA Corpus).

Table 2. *to-V vs. V-ing* by verb in the BROWN corpora

	Brown	Frown	American	X^2 (p)	LOB	FLOB	British
start	50:52	59:94	.51 > .61	.23	36:48	47:51	.57 > .52
begin	252:47	203:85	.16 > .30	**<.001**	249:22	212:23	.08 > .10
like	125:43	115:53	.26 > .32	.23	126:37	109:55	.23 > .34
love	10:2	19:5	.17 > .21	.77	7:0	15:7	.00 > .32
hate	8:2	6:6	.20 > .50	.15	1:2	4:7	.67 > .64
bother	13:0	8:1	.00 > .11	.22	9:1	18:0	.10 > .00
propose	12:1	12:3	.08 > .20	.35	34:1	15:1	.03 > .06
try	344:6	371:6	.02 > .02	.90	350:6	322:14	.02 > .04

there is an increase in [V-ing] with *hate* (20 percent > 50 percent [V-ing] from the 1960s to 1990s), but since there are just 22 tokens, the *p* value is .15, which is not below the statistically significant value of $p < .05$. In fact, there is only one of the eight verbs where it is statistically significant (*begin*). The second important fact with the BROWN data is that there is really no way to show how these shifts are related, to see the sequencing in terms of the Great Complement Shift. We only have two time periods (1960s and 1990s) and so even if one shift occurred mainly between 1950 and 1970 and the other was between 1980 and 1990, there would be no way to know this.

The data from COHA is, of course, much more robust. As we will see, nearly all of the shifts are statistically significant, and we can also sequence the shifts with the different matrix verbs. Table 3 shows the data for the same eight verbs as in Table 2. For each verb, we see the number of tokens overall (the combined total of [to V] and [V-ing]) in each decade and below that the percentage of tokens that are [V-ing].

As with BROWN, all of the verbs in COHA show a shift toward [V-ing] over time. Note, however, that whereas there are almost no shifts in the BROWN corpora that are statistically significant, all of the verbs in COHA show a statistically significant shift. The verb *try* is statistically significant at $p < .0001$, while the other six verbs are significant at $p < .0000001$. As we see, while the data in the BROWN corpora are *suggestive* of change, the data in COHA confirm this, and they show that the shift toward [V-ing] has in fact been occurring during at least the past 100 years.

Having established that the shifts are statistically significant, we can examine the relative chronology of the shifts toward [V-ing] with different verbs. As the table shows (pay special attention to the bolded cells), the high frequency verbs *start* increased more than almost any other verbs from about 1900 to 1930s, and *begin* followed just a bit later and its increase has been somewhat more attenuated. While these two verbs have continued to move toward [V-ing] since the 1930s, this shift has been at a slower speed. Since about the 1960s, the largest percent increases have been with *propose* (and

Table 3. *to-V vs. V-ing* by verb in COHA

	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000	TOTAL
start	571	993	1,978	2,717	3,378	3,816	3,798	4,174	4,775	7,040	8,299	41,839
	0.06	**0.17**	**0.22**	**0.41**	0.45	0.47	0.52	0.54	0.51	0.56	0.58	
begin	7,623	7,900	8,498	9,338	9,722	8,973	9,970	9,313	10,549	10,389	10,154	10,9090
	0.08	**0.10**	**0.14**	**0.17**	0.18	0.19	0.20	0.23	0.25	0.27	0.32	
like	2,549	2,924	3,209	3,615	3,703	4,125	3,905	4,027	3,755	4,576	4,993	43,923
	0.02	0.03	0.03	0.06	0.04	0.05	0.05	**0.07**	**0.09**	**0.12**	**0.15**	
love	340	446	428	336	321	406	419	463	534	809	1,291	6,112
	0.04	0.06	0.07	0.11	0.08	0.11	0.15	**0.17**	**0.21**	**0.25**	**0.33**	
hate	303	413	470	468	395	426	383	415	397	438	645	4,943
	0.07	0.07	0.07	0.10	0.12	0.13	0.16	**0.19**	**0.27**	**0.32**	**0.38**	
try	6,066	7,879	9,215	10,120	10,848	11,713	11,941	12,976	14,337	16,415	18,272	135,122
	0.00	0.01	0.01	0.02	0.02	0.02	0.01	0.02	0.02	0.03	0.03	
bother	24	43	107	196	277	341	403	398	361	497	684	3,338
	0.08	0.12	0.03	0.07	0.08	0.11	0.14	**0.12**	**0.17**	**0.18**	**0.23**	
propose	543	586	712	529	399	339	266	274	264	214	169	5,017
	0.04	0.06	0.05	0.05	0.07	0.09	**0.12**	**0.26**	**0.28**	**0.40**	**0.49**	

to a much lesser extent *bother*), and with the "emotion" verbs *love, like*, and *hate*—with largest increase occurring with the most strongly emotional verbs *love* and *hate*.

In this chapter I will not comment further on possible motivations for the early, sharp increase with *start* and the later increase with verbs of emotion (especially the most semantically charged ones). The point here is that whatever motivations we propose for these changes are dependent on having robust, unambiguous data, and such data can only come from a large, well-balanced corpus such as COHA.

7. Monitor corpora

In the previous two sections I looked at recent changes in English verbal complementation, which have been underway for the past 200 years. In this section, I will conclude by looking at changes that are much more recent, and which are still underway. We will see that in order to look at ongoing changes, we need a very

specially constructed corpus—a monitor corpus—and that there are very few corpora of contemporary English that can fulfill this role.

7.1 Previous corpora

There are several large corpora of contemporary English, such as the British National Corpus (BNC). But as Burnard (2002) mentions, neither the BNC nor most other large, contemporary corpora were designed to look at language *change*. 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, or Word Banks Online). The corpus was started in the 1980s and texts continued to be added to the corpus each year for over 20 years, resulting in a corpus of about 455 million words by 2005.

Although (as far as I am aware) the creators of the BoE have never claimed themselves that the BoE could be used as a monitor corpus, this claim has sometimes been made on its behalf.[10] However, the Bank of English 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 diachronic work with the corpus has actually been done.

The flaw that prevents the BoE from being a truly reliable monitor corpus is the fact that its genre balance varies wildly from one year to the next. This means that we can never be quite sure whether the changes that we see from one period to the next are truly indicative of changes in the "real world", or whether they simply result from the changing genre balance. For example, if phenomena that we know are more related to the fiction genre (e.g. the pluperfect, or certain descriptive vocabulary) decrease 40–50 percent overall in the corpus from 1990–94 to 1995–99, it may just be because there is less fiction in the corpus in 1995–99 than in 1990–94. But it would tell us little or nothing about real changes in the language.

7.2 The Corpus of Contemporary American English (COCA)

The only other structured corpus of contemporary English that has been suggested as a valid monitor corpus is the 410 million word Corpus of Contemporary American English (COCA)—the companion corpus to the 400 million word

10 To give just two examples, see Hunston (2002: 30–31) and McEnery, Xiao, and Tono (2006: 67–70). See Davies (2011) for a more complete list. Davies (2011) also provides expanded discussion of much of the material found in section 6.

Corpus of Historical American English (COHA, 1810s–2000s), which was discussed above[11] (for an overview of COCA, see Davies 2009a).

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

As a result, the strange variations over time that are seen in the BoE do not occur in COCA—the frequency of words, phrases, and constructions that are not expected to change over time in fact stay very stable over time. On the other hand, COCA provides clear evidence for ongoing changes from 1990 to the current time. COCA can be used to look at lexical change (e.g. the frequency of any word or phrase in each year since 1990, or easily searching to find words that are much more common in 2005–10 than 1990–94), morphological change (e.g. the relative frequency of the suffix *-gate* (*Iraqgate, Monicagate*) over time), or semantic change (i.e. using changes in collocates to measure change in usage and meaning, such as *web* or *green*).

For reasons of space, however, I will focus here on just a handful of examples of how COCA serves as a monitor corpus to provide evidence for recent syntactic shifts in English. As mentioned, more examples can be found in Davies (2011).

First, consider the salient "quotative like" construction (*and he's like, "I'm not going with her"*) (cf. Barbieri 2009). COCA shows that this is much more common in spoken English, and that there is a clear increase over time as the frequency (per million words) has increased more than fivefold from the early 1990s through the late 2000s (Table 4).

Consider also the "so not" construction (*I'm so not interested in him*; see Hoffmann 2007). Although the number of tokens for this construction are relatively sparse (but still 5–10 times higher than in the Bank of English), it is most common in the more informal genres, and there is a clear increase in the construc-

Table 4. Frequency of quotative *like* in COCA, by genre and time period

	SPOK	FICT	MAG	NEWS	ACAD	1990–94	1995–99	2000–4	2005–9
freq	1,025	72	271	179	29	130	347	462	645
size	81.7	78.8	83.3	79.4	79.3	103.3	102.9	102.6	93.6
per mil	12.5	0.9	3.3	2.3	0.4	1.3	3.4	4.5	6.9

11 See http://corpus.byu.edu/publications.asp for a list of publications that are based on COCA. Curzan (16), Rudanko (17), and Mair (19) are examples of such COCA-based studies.

Table 5. Frequency of [*so not* ADJ] in COCA, by genre and time period

	SPOK	FICT	MAG	NEWS	ACAD	1990–94	1995–99	2000–4	2005–9
freq	14	10	12	0	0	2	6	11	17
size (MW)	81.7	78.8	83.3	79.4	79.3	103.3	102.9	102.6	93.6
per mil	0.17	0.13	0.14	0.0	0.0	0.02	0.06	0.11	0.18

tion over time—it is nearly nine times as common 2005–9 as it was 15–20 years ago (Table 5).

Of course, COCA provides data for less salient syntactic shifts as well. For example, COCA provides clear evidence for the rise in the "*get* passive" (*Bill got hired last week* vs. *Bill was hired last week*) (cf. Hundt 2001; Mair 2006a). It is most common in the more informal genres, and it has increased in each five-year period since the early 1990s, and it is now about 50 percent more common (compared to the *be* passive) than it was 15–20 years ago (Table 6).

Another low-level syntactic change is the slow but consistent shift from [help to V] to [help V] (*I'll help Mary to clean the room* > *I'll help Mary Ø clean the room*), which is a change that has been commented on from a corpus-based approach by Mair (2002), and others (Table 7).

8. Conclusion

As we have seen, Google (and Google News Archive and Google Books)—while initially promising—have serious limitations for use in mapping out recent changes in English. A somewhat more structured approach is to use text archives, where we have more assurance about the date of the texts and more ability to create advanced architectures and interfaces to search the data. As we have seen in section 3, several researchers have been using text archives from the 1800s to 1900s to provide insightful analyses on recent changes in English. Problems with text archives do

Table 6. Percentage of "*get* passive" vs. "*be* passive" by time period, 1990s–2000s

	SPOK	FIC	MAG	NEWS	ACAD	1990–94	1995–99	2000–4	2005–9
[be] [vvn*]	349,128	256,864	420,038	443,669	734,273	584,851	547,651	529,224	542,246
[get] [vvn*]	20,829	14,780	12,860	12,176	2,889	13,966	15,663	15,628	18,277
% [get]	2.9%	2.8%	1.5%	1.4%	0.2%	1.2%	1.4%	1.5%	1.7%

Table 7. Frequency of *help to V/help V* by period, 1990s–2000s

	search string	1990–94	1995–99	2000–4	2005–9
+ to	[help] [p*] to [v*]	5586	6501	7164	7202
- to	[help] [p*] [v*]	841	809	728	634
	% –to	86.9%	88.9%	90.8%	91.9%

remain, however, including accuracy, annotation, architecture, availability, and genre balance between different time periods.

While text archives emphasize size, small structured corpora like BROWN, ARCHER, CONCE, and the DCPSE emphasize "attention to detail". As we have seen, researchers have used data from these small corpora to provide insightful analyses of high (and some medium) frequency phenomena like auxiliaries and modals. As I have attempted to show, however, there is probably a false dichotomy between the "small and tidy" approach advocated by some users of the small corpora and the supposed "large and [presumably] dirty (and unbalanced)" approach used in large corpora and text archives (for a similar analysis, see Mair 2006b).

The Corpus of Historical American English proves that it is possible to create corpora that are 100–400 times as large as these smaller corpora, but it is still possible to have corpora that are textually accurate, well-annotated, and genre-balanced. As we have seen with the two types of verbal complements shown in sections 5–6, a large corpus like COHA provides detail, robustness, granularity, and insight for these lower and medium frequency constructions that would likely never be possible with a small 1–4 million word corpus.

Finally, we have seen the difficulty in creating useful monitor corpora to look at ongoing changes. Unless we have the same genre balance from one time period to the next, we will never know if the changes in the corpus are merely artifacts of the corpus or whether they actually represent changes in the "real world". I have argued that the Corpus of Contemporary American English (COCA)—because of its size (410 million words, 1990–2010) and because it has almost exactly the same genre balance from year to year—is perhaps the only structured corpus of English that would allow us to accurately map out the precise increase or decrease in these constructions, for very recent changes in American English.

References

Barbieri, Federica. 2009. 'Quotative "Be Like" in American English: Ephemeral or Here to Stay?' *English World-Wide* 30: 68–90.

Biber, Douglas, Edward Finegan, and Dwight Atkinson. 1994. 'ARCHER and Its Challenges: Compiling and Exploring a Representative Corpus of Historical English

Registers'. In *Creating and Using English Language Corpora*, ed. Udo Fries, Gunnel Tottie, and Peter Schneider, 1–13. Amsterdam: Rodopi.

Burnard, Lou. 2002. 'Where Did We Go Wrong? A Retrospective Look at the British National Corpus'. In *Teaching and Learning by Doing Corpus Analysis*, ed. Bernhard Kettemann and Georg Marko, 51–71. Amsterdam: Rodopi.

Cuyckens, Hubert, and Hendrik De Smet. 2007. 'For . . . to-Infinitives from Early to Late Modern English'. *Linguistic Insights—Studies in Language and Communication* 28: 77–103.

Davies, Mark. 2009a. 'The 385+ Million Word Corpus of Contemporary American English 1990–2008+: Design, Architecture, and Linguistic Insights'. *International Journal of Corpus Linguistics* 14: 159–90.

———. 2009b. 'Review of The International Corpus of English—British Component ICE-GB, the Diachronic Corpus of Present-Day Spoken English DCPSE, and ICECUP 3.1'. *Language* 85: 443–45.

———. 2011. 'The Corpus of Contemporary American English as the First Reliable Monitor Corpus of English'. *Literary and Linguistic Computing* 25: 447–65.

De Smet, Hendrik. 2008. 'Diffusional Change in the English System of Complementation: Gerunds, Participles and For . . . To-Infinitives'. Ph.D. dissertation. University of Leuven.

Gries, Stefan Th., and Anatol Stefanowitsch. 2003. 'Co-Varying Collexemes in the *Into*-Causative'. In *Language, Culture, and Mind*, ed. Michel Achard and Suzanne Kemmer, 225–36. Stanford, CA: CSLI Publications.

Hoffmann, Sebastian. 2007. 'From Web-Page to Mega-Corpus: The CNN Transcripts'. In *Corpus Linguistics and the Web*, ed. Marianne Hundt, Nadja Nesselhauf, and Carolin Biewer, 69–85. Amsterdam: Rodopi.

Hundt, Marianne. 2001. 'What Corpora Tell Us about the Grammaticalisation of Voice in *Get*-Constructions'. *Studies in Language* 25: 49–87.

Hunston, Susan. 2002. *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.

Kilgarriff, Adam. 2007. 'Googleology is Bad Science'. *Computational Linguistics* 33: 147–51.

Kytö, Merja, Juhani Rudanko, and Erik Smitterberg. 2000. 'Building a Bridge between the Present and the Past: A Corpus of 19th-Century English'. *ICAME Journal* 24: 85–97.

Leech, Geoffrey, Marianne Hundt, Christian Mair, and Nicholas Smith, eds. 2009. *Change in Contemporary English: A Grammatical Study*. Cambridge: Cambridge University Press.

Mair, Christian. 1997. 'Parallel Corpora: A Real-Time Approach to the Study of Language Change in Progress'. In *Corpus-Based Studies in English*, ed. Magnus Ljung, 195–209. Amsterdam: Rodopi.

———. 2002. 'Three Changing Patterns of Verb Complementation in Late Modern English: A Real-Time Study Based on Matching Text Corpora'. *English Language and Linguistics* 6: 105–31.

———. 2006a. *Twentieth-Century English: History, Variation, and Standardization*. Cambridge: Cambridge University Press.

———. 2006b. 'Tracking Ongoing Grammatical Change and Recent Diversification in Present-Day Standard English: The Complementary Role of Small and Large Corpora'. In *The Changing Face of Corpus Linguistics*, ed. Antoinette Renouf and Andrew Kehoe, 355–76. Amsterdam: Rodopi.

McEnery, Tony, Richard Xiao, and Yukio Tono. 2006. *Corpus-Based Language Studies: An Advanced Resource Book*. London: Routledge.

Rohdenburg, Gunter. 2007. 'Functional Constraints in Syntactic Change: The Rise and Fall of Prepositional Constructions in Early and Late Modern English'. *English Studies* 88: 217–33.

———. 2009. 'Grammatical Divergence between British and American English in the Nineteenth and Early Twentieth Centuries'. *Linguistic Insights—Studies in Language and Communication* 77: 301–29.

Rudanko, Juhani. 2000. *Corpora and Complementation*. Lanham, MD: University Press of America.

———. 2005. 'Lexico-Grammatical Innovation in Current British and American English: A Case Study on the Transitive *into -ing* Pattern with Evidence from the Bank of English Corpus'. *Studia Neophilologica* 77: 171–87.

———. 2006. 'Watching English Grammar Change: A Case Study on Complement Selection in British and American English'. *English Language and Linguistics* 10: 31–48.

Vosberg, Uwe. 2003. 'Cognitive Complexity and the Establishment of *-ing* Constructions with Retrospective Verbs in Modern English'. *Linguistic Insights—Studies in Language and Communication* 7: 197–220.

Wulff, Stefanie, Anatol Stefanowitsch, and Stefan Th Gries. 2007. 'Brutal Brits and Persuasive Americans: Variety Specific Meaning Construction in the into-Causative'. In *Aspects of Meaning Construction*, ed. Gunter Radden, Klaus-Michael Köpke, Thomas Berg, and Peter Siemund, 265–81. Amsterdam: Benjamins.