Expanding horizons in historical linguistics with the 400-million word Corpus of Historical American English

Mark Davies[1]

Abstract

The Corpus of Historical American English (COHA) contains 400 million words in more than 100,000 texts which date from the 1810s to the 2000s. The corpus contains texts from fiction, popular magazines, newspapers and non-fiction books, and is balanced by genre from decade to decade. It has been carefully lemmatised and tagged for part-of-speech, and uses the same architecture as the Corpus of Contemporary American English (COCA), BYU-BNC, the TIME Corpus and other corpora. COHA allows for a wide range of research on changes in lexis, morphology, syntax, semantics, and American culture and society (as viewed through language change), in ways that are probably not possible with any text archive (e.g., Google Books) or any other corpus of historical American English.

1. Introduction

Some languages have large, robust historical corpora, which allow for research on a wide range of topics. In Spanish, for example, there is the 200-million word Corpus Diacrónico del Español and the 100-million word Corpus del Español, while for Portuguese there is the 45-million word Corpus do Português, which is carefully annotated for part-of-speech and lemma (see Davies, 2008, 2010b).

English historical corpora, on the other hand, have tended to be somewhat smaller, and this is true of corpora of Late Modern English as well. The Brown family of corpora (Brown, LOB, FROWN, FLOB, which date from the 1960s to the 1990s) has a combined total of four million words (see Hundt and Leech, forthcoming; and Mair, 1997; also note the recent

[1] Department of Linguistics and English Language, Brigham Young University, 4064 JFSB, Provo, Utah 84602-6278, USA.
Correspondence to: Mark Davies, *e-mail*: mark_davies@byu.edu

Corpora 2012 Vol. 7 (2): 121–157
DOI: 10.3366/corp.2012.0024

www.eupjournals.com/cor

Be06 and Am06 additions by Baker), while the ARCHER Corpus (1700s to the 1900s) contains less than two million words (see Biber *et al*., 1994; and Yáñez-Bouza, forthcoming). The Corpus of Nineteenth-Century Texts, known as CONCE (UK, 1800s) contains about one million words (see Kytö *et al*., 2000), and the Diachronic Corpus of Present-day Spoken English or DCPSE (UK, 1950s and 1990s) contains less than one million words (see Aarts *et al*., forthcoming; and Davies, 2009b).

One reason why historical corpora of English tend to be small may be the notion that there is an inherent dichotomy between 'small and tidy' corpora and 'large and messy' corpora. Just as this is a false dichotomy for contemporary, synchronic corpora (see the British National Corpus, which is large, and also well-constructed and well-balanced), one could argue that it is a false and unhelpful dichotomy for historical corpora, too.

In this paper, I shall discuss the 400-million word Corpus of Historical American English (COHA), which was released in late-2010, and which is freely accessible online.[2] COHA differs from all other corpora of historical English in that it is quite large – 100 times larger than any other structured corpus. But it is also well balanced by genre and sub-genre in each decade, and it has been carefully lemmatised and tagged for part-of-speech. As we will see, the unique balance of size, genre and corpus architecture with COHA results in a resource that allows us to carry out research on many types of language change – lexical, morphological, syntax and semantic – that could not be studied otherwise. As a result, it significantly expands our horizons about what can be done with historical corpora, when we no longer operate within the artificial constraints of small one- to five-million word corpora.

In Section 2 of this paper, I discuss how COHA was designed, created and annotated. In Sections 3 to 8, I show how the corpus can be used to research a wide range of phenomena relating to lexical, morphological, phraseological, syntactic and semantic changes in American English. In Section 6.1, I compare COHA with other corpora in terms of size and data granularity, while in Section 7 I compare it with large text archives (such as Google Books) in terms of architecture.

2. Designing, creating and annotating the corpus

The Corpus of *Historical* American English contains 400 million words from 1810 to 2009 – or, in other words, the last 200 years. This historical depth complements the coverage of its companion corpus, the Corpus of *Contemporary* American English (COCA), which contains 410 million

[2] See: http://corpus.byu.edu/coha

Decade	Fiction	Magazines	Newspaper	NF Books	Total	Percent fiction
1810s	641,164	88,316	0	451,542	1,181,022	0.54
1820s	3,751,204	1,714,789	0	1,461,012	6,927,005	0.54
1830s	7,590,350	3,145,575	0	3,038,062	13,773,987	0.55
1840s	8,850,886	3,554,534	0	3,641,434	16,046,854	0.55
1850s	9,094,346	4,220,558	0	3,178,922	16,493,826	0.55
1860s	9,450,562	4,437,941	262,198	2,974,401	17,125,102	0.55
1870s	10,291,968	4,452,192	1,030,560	2,835,440	18,610,160	0.55
1880s	11,215,065	4,481,568	1,355,456	3,820,766	20,872,855	0.54
1890s	11,212,219	4,679,486	1,383,948	3,907,730	21,183,383	0.53
1900s	12,029,439	5,062,650	1,433,576	4,015,567	22,541,232	0.53
1910s	11,935,701	5,694,710	1,489,942	3,534,899	22,655,252	0.53
1920s	12,539,681	5,841,678	3,552,699	3,698,353	25,632,411	0.49
1930s	11,876,996	5,910,095	3,545,527	3,080,629	24,413,247	0.49
1940s	11,946,743	5,644,216	3,497,509	3,056,010	24,144,478	0.49
1950s	11,986,437	5,796,823	3,522,545	3,092,375	24,398,180	0.49
1960s	11,578,880	5,803,276	3,404,244	3,141,582	23,927,982	0.48
1970s	11,626,911	5,755,537	3,383,924	3,002,933	23,769,305	0.49
1980s	12,152,603	5,804,320	4,113,254	3,108,775	25,178,952	0.48
1990s	13,272,162	7,440,305	4,060,570	3,104,303	27,877,340	0.48
2000s	14,590,078	7,678,830	4,088,704	3,121,839	29,479,451	0.49
Total	207,633,395	97,207,399	40,124,656	61,266,574	**406,232,024**	0.51

Table 1: Composition of COHA by genre and decade

words of text from 1990 to 2010 – or, in other words, the last twenty years of American English (see Davies, 2009a, 2010a, 2011). The composition of the corpus is summarised under Table 1.

As Table 1 indicates, COHA is balanced by genre across the decades. For example, fiction accounts for 48 to 55 percent of the total in each decade from the 1810s to the 2000s, and the corpus is balanced across decades

for genres, and for sub-genres and domains as well.[3] We also ensured that, decade-by-decade, we have nearly the same balance for twenty-four different non-fiction book categories based on the Library of Congress classification, (e.g., history, religion and technology). The same holds for other genres, such as fiction, where we have the same balance decade-by-decade for sub-genres like prose, poetry and drama. This balance across genres and sub-genres allows researchers to be reasonably certain that they are examining 'real world' changes, and that any change they observe is not an artefact of differences in genre balance. Much more data on the composition of the corpus, including a downloadable file with metadata on all 100,000 texts, can be found on the corpus website.

Having designed the corpus, we then assembled over 100,000 texts in COHA.[4] As Table 2 shows, some were already available as part of existing text archives (e.g., Project Gutenberg and Making of America); many had to be converted from PDF images to text (e.g., all of the 40,000+ newspaper files dating from 1860 to 1989), and many of the texts (especially novels and non-fiction books) were scanned from printed sources, using OmniPage 15 for Optical Character Recognition (OCR).

Having acquired the texts, we undertook a detailed post-processing phase to clean up the texts. For example, for each of the 100,000 newspaper texts that were converted from PDF images, we calculated what percent of types in each specific file were also found in a completely 'clean' eighty-million word corpus of newspaper texts from the 1990s to the 2000s: this provided an 'accuracy score'.[5] We originally converted more than 100 million words of text from newspaper PDF files (100,000 articles). Since we only needed forty million words for the corpus, however, we had the freedom to 'throw away' the 60 percent of the texts with the lowest accuracy scores, and the 40 percent that remained were of very good quality. For example, none of the newspaper texts have less than 98 percent of the types from the clean, modern texts. Similar procedures (to eliminate problematic texts

[3] The one exception is the lack of newspapers for the 1810s to the 1850s. We have been unable to find large amounts of 'clean' newspaper text for those decades, including newspapers in PDF format. For these decades, however, we do have magazine articles, which are similar in style to newspapers. Starting in the 1860s, we have very good genre balance from one decade to the next.

[4] Nearly all of the design and creation of the corpus was undertaken by myself alone. However, since some students did help with scanning books and with error correction, I use 'we' to discuss the corpus creation under Section 2.

[5] Since these are historical texts (with lexis that is now archaic and older spellings), clearly not all types would be found in the modern 'control' corpus. In addition, most texts had some types (for example proper names) which were correct, but which were not found in the clean, comparison corpus. Therefore, many of the texts had less than 100 percent 'recognition' in terms of comparison with COCA texts. Note also that we focussed just on single types, rather than on bigrams or trigrams. However, most texts with problematic bigrams or trigrams would also have problematic 1-grams (types) as well, and these texts were eliminated with the procedure that we have described.

Genre	Sources
Fiction	Project Gutenberg (1810–1930), Making of America (1810–1900), scanned books (1930–1990), movie and play scripts, COCA (1990–2010).
Magazine	Making of America (1810–1900), scanned and PDF (1900–1990), COCA (1990–2010) – In each decade, the magazines are balanced across at least ten magazines (with equivalent sub-genres in each decade of the 1900s).
Newspaper	PDF > TXT of at least five newspapers (1850–1980), COCA, *etc*. (1990–2010).
Non-fiction	Project Gutenberg (1810–1900), www.archive.org (1810–1900), scanned books (1900–1990), COCA (1990–2010) – In each decade, the non-fiction is balanced across the Library of Congress classification system.

Table 2: Sources

through comparison with clean, contemporary texts) were followed for the other genres as well.

After selecting just the most accurate texts and post-processing these texts, we then lemmatised the corpus and tagged it for part of speech, using the CLAWS tagger that has been employed on the British National Corpus, COCA and other corpora of English. Obviously, some older forms would not be correctly tagged or lemmatised by CLAWS, which was designed for contemporary English. Through placing all of the frequency data in a relational database, however, we could find those words whose frequency was much higher in COHA than in COCA. Some of these had been scanned correctly but were simply older or obsolete forms (e.g., *musick*, *common-place*, *academical* and *woful*), while others were in fact typos that resulted from bad scans of printed books or bad conversion from PDF files. Through a web interface, students examined each of the approximately 100,000 types that occurred more than two times in COHA, and which had a frequency (per million words) in COHA that was more than three times the rate in COCA (possibly indicating that it was a typo). They looked at the word in context, and corrected the word form, lemma and part of speech, when necessary. Thus, while it is obviously not perfect (indeed, this would be an almost impossible feat for a 400-million word corpus), the textual corpus in COHA is in fact very clean and accurate.

Having discussed the design and creation of the corpus, let us now examine in some detail the different types of research that can be carried out with the COHA data.

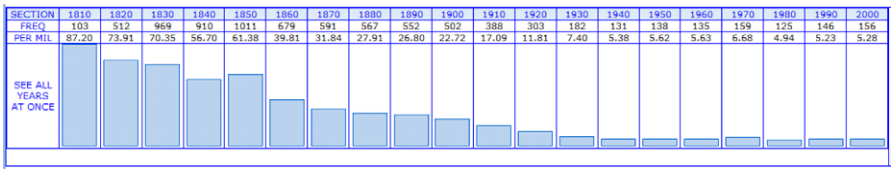

Figure 1: Frequency of *bestow**, 1810s to the 2000s

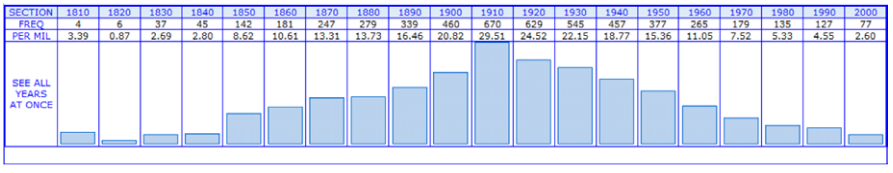

Figure 2: Frequency of *mustn't*, 1810s to the 2000s

3. Lexical change

3.1 Frequency of specific words and phrases

At the most basic level, the COHA architecture and interface[6] allow us to see the frequency of any word or phrase in each of the twenty decades in the corpus from the 1810s to the 2000s. This is, of course, much more useful than resources like the Oxford English Dictionary, which can show the first attestation of a word, but are then unable to show its changing frequency over time. Examples of the frequency charts are shown under Figures 1 to 3, where we see words that have been decreasing in frequency since the 1800s (forms of *bestow**, see Figure 1), a phrase which peaked about 100 years ago (*mustn't*, see Figure 2), and words that have been increasing over time (*teenager**, see Figure 3). As shown under Figure 3, the frequency is often a function of historical, cultural or societal changes, which impact on the language – in this case, different views of adolescents in the US in the post-war years of the 1940s and 1950s.[7]

Of course, the corpus interface does not simply show the frequency of words, phrases and grammatical constructions, but it also shows the Keyword in Context entries for any data shown in the frequency display. For example, users can click on the 1910s bar shown under Figure 1 to see all 388 tokens of forms of *bestow*, as shown under Table 3.[8]

---

[6] The COHA architecture and interface are the same as those used for the other corpora from http://corpus.byu.edu, such as COCA, TIME and BYU-BNC.

[7] Users can select a decade and see more detailed frequency data to the right of the bar chart (as under Figure 4). Also, the frequency charts are, of course, 'normalised', which means that they are based on the token frequency per million words in each decade.

[8] In view of the limitations of space in this paper, the format is different from what is seen in the web interface, where there is just one line for each entry and the word or phrase appears

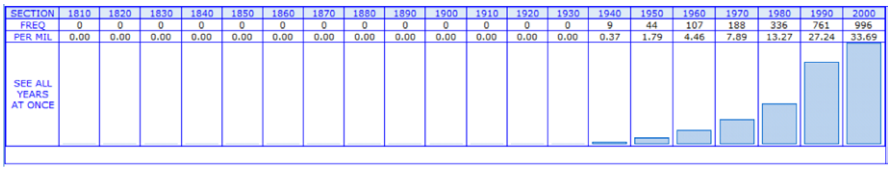

Figure 3: Frequency of *teenager**, 1810s to the 2000s

1910	FIC	AlchemistsSecret	and sorrow that were poured into his ears. Hour after hour he had spent **bestowing** the priestly absolution on the repentant sinner, giving fatherly advice and consolation to the
1914	MAG	Nation	Prince, if the Kaiser himself, can pause in the midst of conflict to **bestow** praise upon the high qualities exhibited by the French, Americans need not feel it
1914	NF	AmateurGarden	to receive two prizes consecutive in the list. The second prize can not be **bestowed** in the same district in which the first is being awarded, though the third
1914	FIC	Ponteach	And bring about as long a lasting Peace As tho' the Whole were lavishly **bestow**'d? CATCHUM. I'm clear upon' t they will, if we
1918	NEWS	NYT-Reg	the graves, showing the tender recent care which the French have not failed to **bestow** also on American graves. The shifting of soldiers brought to light many interesting stories

Table 3: Keyword in Context (KWIC) entries: *bestow**

For more detailed investigations of word- and phrase-frequency, users can also see the frequency in each individual year from 1810 to 2009. For example, Figure 4 shows that the word *Reds* is the most frequent in the 1950s. Users can click on the '1950s' heading to see the frequency in each year of the 1950s. In this case, as Figure 4 shows, they would see that its frequency is highest in 1953, and this again corresponds with changes in American history and society (for example, the year in which the 'anti-Red' congressional hearings of Senator Joseph McCarthy were most prominent).

in the centre of the line. In the web interface it is also possible to click on any KWIC entry and reveal up to 120 words of context.

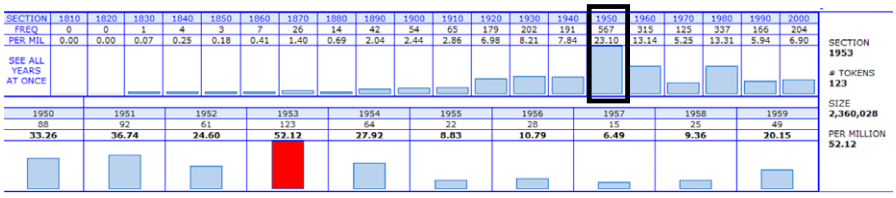

Figure 4: Frequency of *Reds* by decade and year

3.2 Comparing all words in different time periods

In the examples above, we found the frequency of a particular word or phrase. However, since the corpus architecture has stored the frequency of each matching string in each decade, COHA can also show us all words and phrases that are more frequent in one decade than another, even when we do not have any idea what these words might be. For example, Table 4 shows adjectives that are more frequent in the 1870s to the 1910s than in the 1970s to the 2000s (left side) and those which are more frequent in the 1970s to the 2000s (right side).[9, 10]

As we can see, with one simple search, COHA allows us to compare, quickly and easily, the frequency of *all* words in different periods. This is a powerful tool for finding neologisms and for seeing interesting cultural and historical shifts over time – such as the rise of adjectives like *global*, *electronic*, *online*, *sexy* and *innovative* in Table 4, which relate to cultural or technological advances in the late-1900s.

4. Morphological change

COHA also allows us to search the 400 million words to see changing patterns in terms of word formation. For example, Table 5 shows changes

[9] We should briefly explain the organisation of the data under Table 4, since similar tables are found in other sections of this paper. Taking the example of *unfitted* (the sixth word on the left), we see that it occurs 217 times from 1870 to 1919 (see Section 1) and just nine times between 1970 and 2009 (see Section 2). The next two columns show the frequency per million words (PM1 and PM2) in these two sections. Finally, we find the ratio of the normalised figures in the two sections and see that *unfitted* is about twenty-five times more frequent between 1870 and 1919 than it is between 1970 and 2009. The results are ranked in terms of this ratio between the two sections of the corpus.

[10] In this list we see adjectives that appear simply because they are spelt differently in the two periods (e.g., *mediaeval*, *every-day* and *especial*). If the results were lemmatised (which it is possible to do through the web interface), these forms would be grouped with *medieval*, *everyday*, *special*, etc., and would probably not appear in the list. I have also 'smoothed' the data to allow for division by zero, when a word does not occur in the other section. Finally, not all entries that appear in the online corpus are shown here (note the skipped numbers in the far left column of the entries to the left).

	1870s–1910s	1	2	PM2	PM1	Ratio
1	*anti-slavery*	268	9	2.6	0.1	30.5
2	*mediaeval*	575	21	5.5	0.2	28.0
3	*every-day*	537	20	5.2	0.2	27.5
4	*unwonted*	500	19	4.8	0.2	26.9
5	*especial*	1,296	53	12.4	0.5	25.0
6	*unfitted*	217	9	2.1	0.1	24.7
8	*pecuniary*	775	33	7.4	0.3	24.0
9	*impracticable*	659	30	6.3	0.3	22.5
11	*effectual*	478	23	4.6	0.2	21.3
12	*illimitable*	207	10	2.0	0.1	21.2
15	*sagacious*	388	20	3.7	0.2	19.8
16	*poetical*	964	50	9.2	0.5	19.7
17	*kind-hearted*	393	21	3.8	0.2	19.1
20	*toilsome*	198	11	1.9	0.1	18.4
21	*wonted*	197	11	1.9	0.1	18.3
22	*pleasantest*	268	15	2.6	0.1	18.3

	1970s–2000s	2	1	PM2	PM1	Ratio
1	*global*	4,748	1	44.5	0.0	4642.8
2	*israeli*	3,942	0	37.0	0.0	3696.6
3	*okay*	3,708	1	34.8	0.0	3625.8
4	*electronic*	2,970	1	27.9	0.0	2904.2
7	*goddamn*	1,853	1	17.4	0.0	1811.9
8	*iraqi*	1,912	0	17.9	0.0	1793.0
9	*online*	1,574	1	14.8	0.0	1539.1
10	*sexy*	1,396	0	13.1	0.0	1309.1
11	*teenage*	1,298	0	12.2	0.0	1217.2
13	*nonprofit*	945	0	8.9	0.0	886.2
14	*postwar*	906	1	8.5	0.0	885.9
15	*innovative*	914	0	8.6	0.0	857.1
16	*high-tech*	901	0	8.5	0.0	844.9
19	*prestigious*	798	1	7.5	0.0	780.3
20	*operational*	815	0	7.6	0.0	764.3
21	*ecological*	782	0	7.3	0.0	733.3

Table 4: Comparison of adjectives: 1870s to the 1910s, and 1970s to the 2000s

	–*ism* word	Total	1810	1830	1850	1870	1890	1910	1930	1950	1970	1990
2	*patriotism*	4,916	25	437	331	256	356	482	221	114	155	123
3	*communism*	4,778			14	57	24	32	437	1,449	294	320
5	*socialism*	3,526			54	145	181	446	397	277	310	135
8	*optimism*	2,502		12		22	48	163	223	235	227	194
9	*capitalism*	2,501					6	65	269	215	258	435
10	*despotism*	2,249	23	203	386	199	120	86	54	46	26	22
14	*nationalism*	1,838				4	23	97	172	195	141	169
15	*terrorism*	1,813		1	3	9	7	18	55	30	222	146
19	*skepticism*	1,599		53	62	58	42	79	84	90	126	174
20	*imperialism*	1,548				13	28	62	90	197	102	45
21	*barbarism*	1,541	5	79	114	128	146	80	54	40	22	23

Table 5: Frequency of –*ism* words[11]

during the last 200 years in the frequency of words ending in **ism*.[12] Note the decrease with a few words since the 1800s (*despotism*, *patriotism* and *barbarism*), but also those words that have increased much more in the mid- to late-1900s (e.g., *communism*, *capitalism*, *terrorism* and *skepticism*), which may provide interesting insights into cultural and societal changes in the United States.

As with simple words, COHA allows us to compare word forms across different time periods. For example, Table 6 compares **ism* words in the 1860s to the 1910s, and 1970 to 2009. Again, we see interesting shifts in American English, and American culture and society in general, with a decrease in words like *Romanism* and *heathenism*, and an increase in words like *racism* and *activism*.

While the preceding tables relate to a morphological subset of lexical items (in this case, words ending in **ism*), with COHA it is also possible to compare morphological alternates, such as the relative frequency of *lighted* / *lit*. Table 7 is based on 2,403 tokens, and it shows the relative frequency in each decade from the 1810s to the 2000s (e.g., *he lighted* / *lit the fire*), where *lighted* / *lit* is immediately preceded by a pronoun. As Figure 5 indicates, there is a clear increase in *lit* as the simple past form of *light* since the 1810s, and it is more than twice as common as it was eighty to ninety years ago.

5. Phraseological change

In this section, I expand the scope somewhat and look at localised patterns of words (phraseologies), and I will expand this even more in the following

[11] In this and other similar tables in this paper, we show the raw frequency in each decade, but users can see the normalised frequency as well.

[12] The raw frequency (number of tokens) is shown here, but it is also possible to see the normalised frequency by tokens per million, which is indicated here by shading (where a darker shade indicates a higher frequency). And, finally, as with other tables, for reasons of space, only every other decade is shown here, while all are shown in the web interface.

	1860s–1910s	1	2	PM2	PM1	Ratio
1	*pauperism*	217	1	1.8	0.0	190.7
2	*fetichism*	113	0	0.9	0.0	93.1
3	*bimetallism*	62	0	0.5	0.0	51.1
4	*romanism*	62	0	0.5	0.0	51.1
5	*heathenism*	94	2	0.8	0.0	41.3
6	*mohammedanism*	117	3	1.0	0.0	34.3
7	*monopolism*	40	0	0.3	0.0	33.0
8	*spiritism*	38	0	0.3	0.0	31.3
9	*invalidism*	70	2	0.6	0.0	30.8
10	*trade-unionism*	34	1	0.3	0.0	29.9

	1970s–2000s	2	1	PM2	PM1	Ratio
1	*racism*	994	0	9.3	0.0	932.1
2	*tourism*	759	0	7.1	0.0	711.7
3	*activism*	405	1	3.8	0.0	460.8
4	*marxism*	342	2	3.2	0.0	194.6
5	*fundamentalism*	176	0	1.7	0.0	165.0
6	*multiculturalism*	156	0	1.5	0.0	146.3
7	*counterterrorism*	134	0	1.3	0.0	125.7
8	*authoritarianism*	101	1	1.0	0.0	114.9
9	*consumerism*	121	0	1.1	0.0	113.5
10	*sexism*	116	0	1.1	0.0	108.8

Table 6: Frequency of *–ism* words in the 1860s to the 1910s *versus* the 1970s to the 2000s

PRON+	1810	1830	1850	1870	1890	1910	1930	1950	1970	1990
lighted	1	26	41	53	64	71	75	41	25	6
lit	0	11	13	19	50	68	133	162	132	150
% *lit*	0	30	24	26	44	49	64	80	84	96

Table 7: 'PRON + *lit*' versus 'PRON + *lighted*'

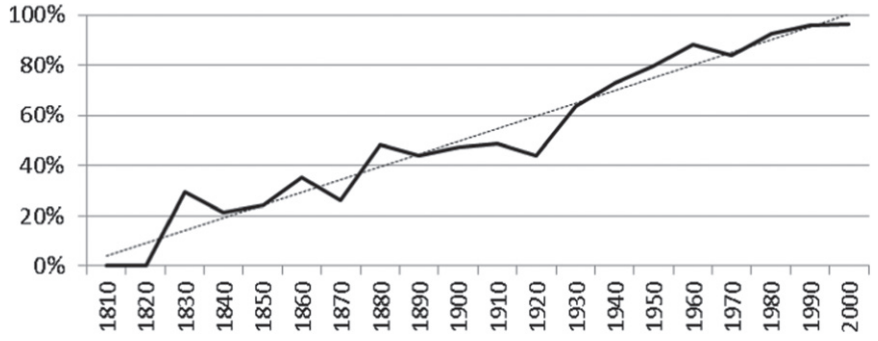

Figure 5: 'PRON + *lit*' versus 'PRON + *lighted*'

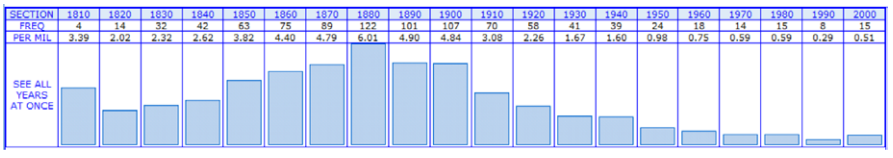

Figure 6: '[*have*] *quite* [V-*ed*]'

section when I consider syntactic change. As an introduction to this topic, consider first the frequency over time of the phrase *have quite* V-*ed* (*have quite forgotten*, *had quite gone*). As Figure 6 shows, the use of this phrase has decreased markedly since the 1800s.

In addition to seeing a display in the form of a chart, users can also see the frequency of each matching string in each decade (see Table 8). They can click on any number to see a particular word or phrase in a particular decade, or select multiple entries in multiple decades.

As another example of phraseological change, we might consider phrasal verbs. Table 9 shows the top ten phrasal verbs with *up* (with the results grouped by lemma) and with the frequency in each decade.[13]

Table 10 contains a comparison of phrasal verbs with the particle *up* in the 1910s to the 1940s and the 1960s to the 2000s. COHA quickly finds

[13] As noted above, for reasons of space in this paper, only every other decade is shown, but all may be viewed in the online corpus.

		Total	1820	1840	1860	1880	1900	1920	1940	1960	1980	2000
1	*had quite forgotten*	149	5	10	11	17	14	9	7	1	1	1
2	*had quite recovered*	28	1	1	1	1	4		1	2	2	1
3	*had quite lost*	23	1	2	4	4	4		2			
4	*have quite forgotten*	23			2	1	3	3	1			
5	*had quite disappeared*	16			1	2	3	1	2			
6	*had quite finished*	16				5	1	1	2			
7	*have quite lost*	15			2	1	1	1				2
8	*had quite gone*	14			2	1	3	3				1
9	*had quite made*	13		1	1	2	3					
10	*has quite forgotten*	11		1		2	1	2				1

Table 8: '[*have*] *quite* [V-*ed*]'

		Total	1820	1840	1860	1880	1900	1920	1940	1960	1980	2000
1	*pick up*	32,670	18	216	413	734	998	1,716	2,944	3,092	3,095	3,850
2	*look up*	31,752	175	613	1,089	1,572	1,529	1,986	2,565	2,149	2,092	2,836
3	*come up*	29,743	93	709	1,111	1,322	1,286	1,834	2,286	2,150	2,118	2,398
4	*make up*	24,049	198	759	1,094	1,514	1,549	1,719	1,578	1,390	1,281	1,409
5	*get up*	22,689	61	366	668	748	888	1,375	2,345	1,827	1,614	1,676
6	*give up*	21,725	328	758	1,045	1,119	1,306	1,269	1,218	1,332	1,405	1,588
7	*take up*	20,116	369	880	1,020	1,493	1,666	1,528	1,130	843	758	773
8	*set up*	16,542	62	304	352	498	477	781	1,653	1,591	1,447	1,521
9	*stand up*	14,248	76	169	238	371	527	795	1,392	1,299	1,150	1,405
10	*grow up*	12,734	79	273	444	478	395	548	610	657	1,054	2,245

Table 9: Frequency of phrasal verbs with *up*

	1910s–1940s	1	2	PM2	PM1	Ratio
2	*bolster up*	94	6	0.97	0.05	21.0
3	*fit up*	14	1	0.14	0.01	18.8
4	*shin up*	14	1	0.14	0.01	18.8
5	*foot up*	13	1	0.13	0.01	17.5
6	*plaster up*	12	1	0.12	0.01	16.1
7	*whack up*	12	1	0.12	0.01	16.1
8	*prowl up*	11	1	0.11	0.01	14.8
9	*brace up*	104	10	1.07	0.08	13.9
10	*purse up*	31	3	0.32	0.02	13.9

	1960s–2000s	2	1	PM2	PM1	Ratio
2	*suit up*	119	1	0.91	0.01	88.65
3	*listen up*	91	0	0.70	0.00	69.67
4	*zip up*	144	2	1.10	0.02	53.64
5	*free up*	117	2	0.90	0.02	43.58
6	*beef up*	106	2	0.81	0.02	39.48
7	*chat up*	49	0	0.38	0.00	37.51
8	*ratchet up*	40	1	0.31	0.01	29.80
9	*rack up*	247	7	1.89	0.07	26.29
10	*boot up*	30	0	0.23	0.00	22.97

Table 10: Phrasal verbs with *up*

	1830s–1910s	1	2	PM1	PM2	Ratio
1	. *Latterly*,	32	1	0.10	0.01	24.9
2	. *Fifthly*,	30	1	0.18	0.01	23.4
3	. *Verily*,	148	5	0.88	0.04	23.1
4	. *Scarcely*,	27	1	0.16	0.01	21.0
5	. *Assuredly*,	41	2	0.24	0.02	16.0
6	. *Decidedly*,	18	1	0.11	0.01	14.0
7	. *Positively*,	18	0	0.11	0.01	10.7
8	. *Singly*,	11	1	0.07	0.00	8.6
9	. *Practically*,	106	1	0.63	0.08	8.3
10	. *Unluckily*,	46	5	0.27	0.04	7.2
11	. *Directly*,	23	3	0.14	0.02	6.0
12	. *Reciprocally*,	15	2	0.09	0.02	5.8

	1960s–2000s	2	1	PM2	PM1	Ratio
1	. *Ironically*,	444	1	3.40	0.01	569.8
2	. *Surprisingly*,	142	1	1.94	0.01	182.2
3	. *Alternatively*,	140	1	1.09	0.01	179.7
4	. *Basically*,	229	0	1.75	0.00	175.3
5	. *Additionally*,	220	0	1.68	0.00	168.4
6	. *Typically*,	206	0	1.58	0.00	157.7
7	. *Initially*,	189	0	1.45	0.00	144.7
8	. *Admittedly*,	112	1	0.86	0.01	143.7
9	. *Increasingly*,	160	0	1.22	0.00	122.5
10	. *Interestingly*,	121	0	0.93	0.00	92.6
11	. *Ideally*,	141	2	1.08	0.01	90.5
12	. *Hopefully*,	70	1	0.54	0.01	89.8

Table 11: 'period (full stop) + *–ly* adverb + comma'

	1940	1950	1960	1970	1980	1990	2000
to [v*] *ly.[r*]	4,022	4,088	4,065	3,897	3,965	3,943	3,896
to *ly.[r*] [v*]	159	236	369	604	784	1,505	2,101
% to *ly.[r*] [v*]	4.0	5.8	9.1	15.5	19.8	38.2	53.9

Table 12: '*to* ADV-*ly* VERB' versus '*to* VERB ADV-*ly*'

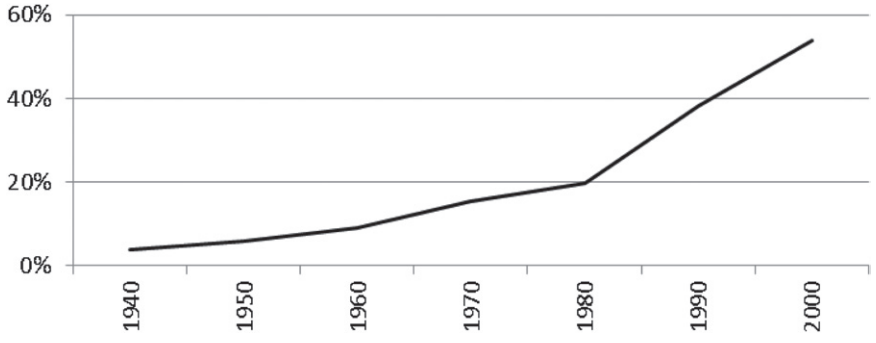

Figure 7: 'to ADV-*ly* VERB' versus '*to* VERB ADV-*ly*'

more recent phrasal verbs like *listen up*, *free up* and *ratchet up*, and now-obsolete and somewhat strange-sounding verbs like *bolster up* ('in a sincere desire to bolster up that foreign tyranny'), *fit up* ('he had fitted up his half of the building as an hotel') and *shin up* ('in simple boy-fashion by shinning up the tree').

Finally, COHA can provide an insight into changes in phraseological 'frames' (see Hunston and Francis, 2000). In these cases, we are looking neither at individual words nor at regular syntactic constructions, but, rather, the frames in which lexical items may appear. For example, consider Table 11, which compares words occurring in the frame '. *ly.[r*],' (i.e., full stop + *–ly* adverb + comma) in the 1830s to the 1910s and the 1960s to the 2000s.

6. Syntactic change

6.1 Prescriptive

Since COHA is lemmatised and tagged for part-of-speech, we are able to carry out in-depth research on syntactic change. Let us first consider changes in terms of two prescriptive rules. The first rule concerns shifts in terms of the split infinitive from the 1940s to the current time, using the search strings '*to* **ly*.[*r**] [*v**]' ('to boldly go') and '*to* [*v**] **ly*.[*r**]' ('to go boldly'), and is based on more than 33,000 tokens (see Table 12). As Figure 7 indicates, there

was a gradual increase in the split infinitive from the 1940s to the 1980s, and this increase has accelerated since that time, so that the relative percentage of split infinitives (*versus* non-split structures) is more than ten times higher than it was sixty to seventy years ago.

The second change in terms of a prescriptive rule is the shift from *whom* to *who* (see Schneider, 1992), as measured here by the ratio of the two phrases '*whom* [do] [p*]' ('whom/who did they') and '*whom* [do] [p*]' ('whom/who does she'). Table 13 contains the data from 2,415 tokens from the 1890s to the 2000s, and Figure 8 shows that the primary increase was from the late-1800s to about the 1930s, with only a slight increase since then (perhaps since the use of *who* is already so high, about 90 percent).

6.2 Descriptive

Turning to descriptive grammar, Figures 9 and 10 show the increase in '*have to* V' ('we have to leave' and 'John had to work') and the decrease with post-verbal negation with *need* ('you need not mention' and 'the people needn't worry'). In terms of extracting the data, it is just a matter of inputting the correct search string ('[*have*] *to* [*v**]' and '*need* [*x**] [*v**]') and COHA quickly finds all of the tokens (215,116 tokens for '[*have*] *to* [*v**]' and 12,998 tokens for '*need* [*x**] [*v**]') and creates a chart, with links to the KWIC entries.

Even more complicated studies of diachronic syntax can be carried out quite easily with COHA. For example, Table 14 and Figure 11 show the contrast between the '*be* passive' ('John was fired last week') and the '*get* passive' ('John got fired last week'; see Hundt, 2001; and Mair, 2006). In this case we simply submit the two competing strings (for a total of 2,726,936 tokens), copy the data from the two charts into a spreadsheet, and create a ratio of the two frequencies. In just a couple of minutes, we can clearly see the shift towards the *get* passive, and we can see that it is (compared to the *be* passive) more than four times as common as it was just eighty to ninety years ago.

As I will discuss more fully under Section 6.1, one of the important advantages of using large corpora is that there are enough tokens to focus on constructions such a verbal subcategorisation, where there would be far too few tokens with a small one- to five-million word corpus. For example, Table 15 and Figure 12 show the shift from '*to*-V' to '*to* V-*ing*' with *accustomed*, as is seen in this data from 3,548 tokens (see also Rudanko, 2010, which is based on our 100-million word TIME Corpus).[14]

Let us consider one more syntactic search that might be quite complex with other corpora, but which can be done quite easily with COHA. This deals with the placement of negation and the use of the 'dummy do'

[14] See: http://corpus.byu.edu/time

_ [do] [p*]	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000
whom	58	42	31	29	29	26	25	29	18	19	28	21
who	93	118	120	135	195	167	172	189	208	239	209	215
% *who*	62	74	79	82	87	87	87	87	92	93	88	91

Table 13: '*who* [*do*] PRON' versus '*whom* [*do*] PRON'

_ + [v?n*]	1820	1840	1860	1880	1900	1920	1940	1960	1980	2000
be	63,529	136,789	129,620	154,447	162,092	179,112	161,397	145,959	136,972	121,851
get	98	370	614	915	1,181	1,546	2,367	2,665	2,965	4,941
% *get*	0.2	0.3	0.5	0.6	0.7	0.9	1.4	1.8	2.1	3.9

Table 14: *get* versus *be* + passive

	1850	1860	1870	1880	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980
V	422	302	340	331	281	209	247	175	111	86	46	42	21	15
V-*ing*	16	21	30	39	38	78	97	90	76	80	85	80	88	102
% V-*ing*	4.0	7.0	8.0	11.0	12.0	27.0	28.0	34.0	41.0	48.0	65.0	66.0	81.0	87.0

Table 15: '*accustomed to* V-*ing* / V'

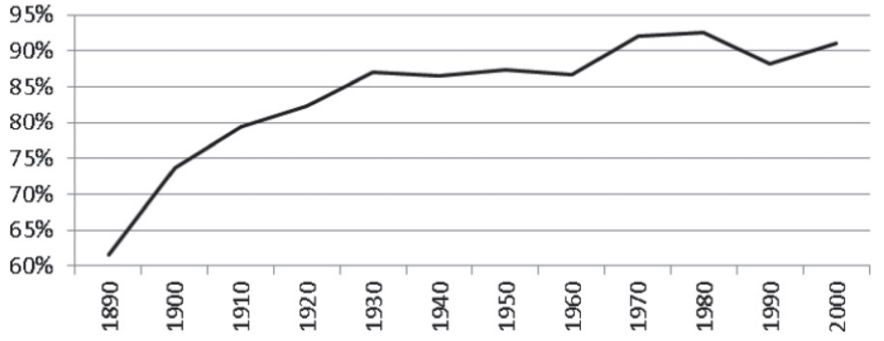

Figure 8: '*who* [*do*] PRON' versus '*whom* [*do*] PRON'

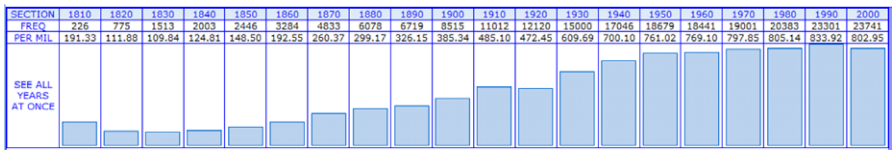

Figure 9: '[*have*] *to* VERB'

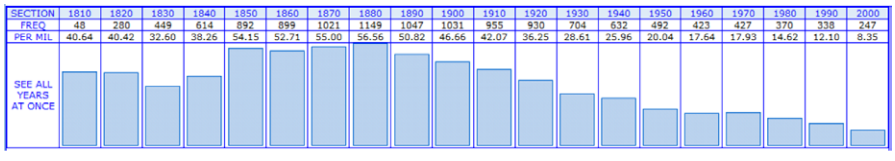

Figure 10: '*need* [NEG] [VERB]'

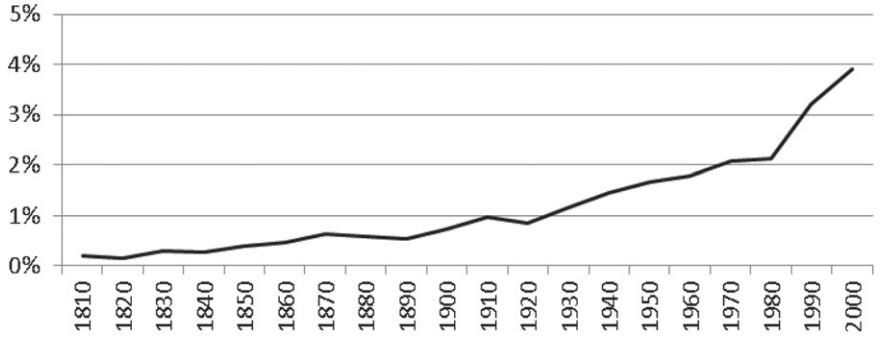

Figure 11: Passive with *get* and *be* (percent *get*)

with 'possessive have'. Under Table 16, [A] represents the older post-verbal placement ('[p*] [have] [x*] [a*]|[d*] [nn*]': 'she hasn't the time') whilst [B] represents the pre-verbal placement with dummy *do* ('[p*] [do] [x*] [have] [a*]|[d*] [nn*]': 'she doesn't have the time'). As before, we simply copy the data from the two charts (13,827 tokens) and do a simple ratio in a

	1820	1840	1860	1880	1900	1920	1940	1960	1980	2000
haven't (A)	155	304	353	536	541	645	368	192	115	81
don't have (B)	1	23	38	65	130	242	407	706	852	1,381
% B	0.6	7.0	9.7	10.8	19.4	27.3	52.5	78.6	88.1	94.5

Table 16: Negation with possessive *have*: (e.g., '*don't have* NP' versus '*haven't* NP')

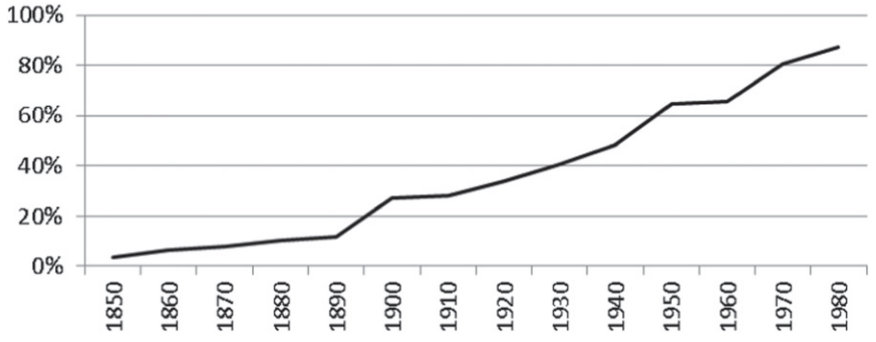

Figure 12: '*accustomed to* [V-*ing* / V]' (percent V-*ing*)

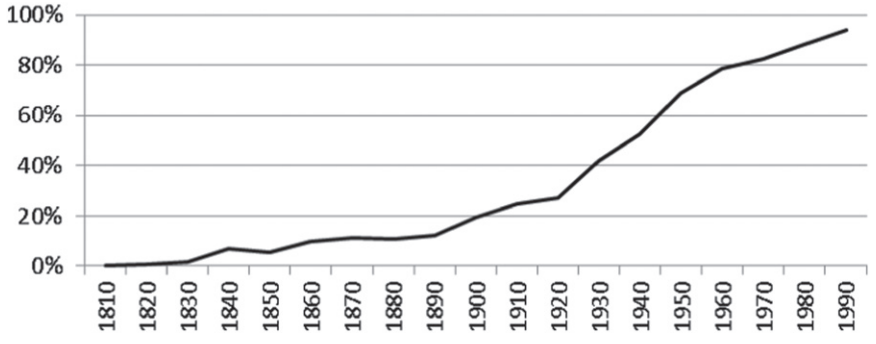

Figure 13: '[*do*] *not*/*n't have* NP' versus '[*have*] *not*/*n't* NP'

spreadsheet. With COHA, we can do even relatively complex searches such as this – resulting in clear and unambiguous data like that in Table 16 and Figure 13 – in just a minute or two.

Finally, note that all of the examples above deal with changes in the complete corpus – all genres. However, language change often spreads through genres, perhaps starting in the more informal genres and then spreading to the more formal genres over time. We can easily map this out with COHA. For example, Table 17 and Figure 14 show the frequency per million words for *must* + lexical verb ('*must* [vv*]'): 'he must know the answer', 'we must leave immediately' (see the Modals chapter in Leech *et al*., 2009). We run the query four times, selecting each of the different genres. We then copy the data into a spreadsheet (as in Table 17) and we can then see

Genre	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000
Fiction	232	191	174	159	163	148	141	136	83	71
Magazine	128	90	90	81	75	71	66	53	57	38
Newspaper	23	42	46	52	44	41	40	33	26	22
NF book	71	54	46	49	43	50	50	37	36	21

Table 17: '*must* [VV*]' by genre (tokens per million words)

	Date	Genre	Source	KWIC
1	1880	NF	RoyalEdinburgh	a prodigal son of that **gay**, brilliant, attractive, and impracticable kind
2	1886	FIC	PoemsStory	all are kindly, some of them, indeed, **Gay**, jolly, joking;
3	1983	MAG	Time	I'm as **gay** as I am heterosexual. O.K., I've experimented with both sexes
4	1988	MAG	GoodHouse	"high risk" groups (**gay** and bisexual men and intravenous drug users),

Table 18: Keyword in Context entries for *gay*

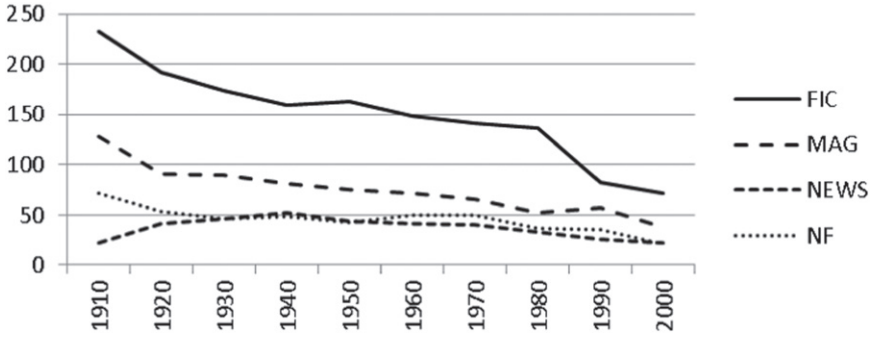

Figure 14: '*must* [VV*]' by genre

(as in Figure 14) how in every decade since the early 1910s the construction has decreased, but that it has decreased the most in the more informal genres.

7. Semantic change

How can we use corpora to see whether words have changed meaning over time? One option would be simply to look up all tokens (or a randomised subset of tokens) and investigate the use of the word. For example, *gay* tokens in the 1880s might look like those shown in rows one and two of Table 18, while those from the 1980s might look like those in rows three and four. As we laboriously examine hundreds or thousands of tokens, one by one, we can begin to see changes in meaning.

With the right corpus architecture, however, we can both simplify this and make it much quicker. A central concept in corpus linguistics is the idea that we 'shall know a word by the company it keeps' (Firth, 1957: 179). If we find that the collocates of a word are changing over time, this may indicate semantic change. For example, in the instances mentioned above, we can see that the collocates of *gay* in the 1880s are *brilliant*, *attractive*, *jolly* and *joking*, while in the 1980s they are *heterosexual*, *sexes*, *groups* and *bisexual*. The goal, then, is to have a corpus architecture that can quickly find and summarise the data from collocates, to help look for semantic change.

Some other corpus architectures look just for exact strings (e.g., *gay party* and *gay men*). With COHA, we are not limited to examining just immediately adjacent words, but, rather, we can look at the entire 'cloud of words' – up to ten words to the left and to the right of the indicated node word. For example, Table 19 shows the most frequent noun and adjective collocates near the noun *gay* in each of the different decades.[15]

Notice that in the 1800s, we find collocates like those in the lighter colour rows, such as *bright*, *flowers*, *laugh*, *lively*, *cheerful* and *attire*. In the late-1900s and in the 2000s, however, collocates such as *lesbian*(*s*), *rights* and *marriage* are more common.

As before, a direct comparison of the collocates in two contrasting periods provides even clearer evidence for the shift in meaning and usage, with collocates of the 'happy, cheerful' meaning (*gallant*, *attire* and *brilliant*) more common in the 1800s, and the 'sexual' meaning in the late-1900s (*lesbian*, *marriage* and *straight*):[16]

Again, the ability to compare quickly the collocates of a given word in two periods can provide insight into semantic change in ways that are perhaps not available with any other corpus or interface.

In addition to using collocates, the COHA architecture provides another tool for looking at change with entire semantic fields. Integrated into COHA is a thesaurus with entries for about 30,000 individual 'synsets'. By searching for '[=word]', we can see the frequency of each matching synonym in each decade. For example, the simple search [=*intelligent*] results in the data shown under Table 21.

This allows us to see that in the semantic field of 'intelligent', the words *wise*, *sensible* and *judicious* have decreased over time, while the words *smart*, *knowledge* and *brainy* have increased. Such data can be useful in seeing how different words are 'competing for semantic space'.[17]

---

[15] As noted above, the figures in this and some other tables show the raw frequency in each decade, but users can see the normalised frequency, too.

[16] Although the 'taboo' sexual meaning may have been present even in the 1800s, in colloquial, spoken language.

[17] Not every token for every word is synonymous with the search word, but this is a good start. For more precision, it would be possible to limit the search to a specific context, such as '[=intelligent] person'.

	Collocate	Total	1820	1840	1860	1880	1900	1920	1940	1960	1980	2000
1	*bright*	173	5	10	13	12	12	12	8	8	2	
2	*flowers*	158	5	11	10	17	13	11	5	1		
3	*lesbian*	155									6	79
4	*laugh*	143	3	5	13	14	13	9	2	7		
7	*rights*	134									19	58
9	*marriage*	99		1	1			1				85
12	*lesbians*	85								1	6	38
16	*lively*	67	3	7	2	5	5	4	1	5		1
17	*cheerful*	67	2	5	5	8	5	6	4		2	
18	*attire*	63	3	11	4	2	4		3	1	1	
32	*abortion*	40										15

Table 19: Noun collocates of *gay* ('N/ADJ near *gay*')

	1830s–1910s	1	2	PM1	PM2	Ratio
1	*grave*	121	1	0.7	0.0	77.0
2	*gallant*	82	0	0.5	0.0	48.9
3	*bird*	79	0	0.5	0.0	47.1
4	*throng*	72	0	0.4	0.0	43.0
5	*attire*	62	1	0.4	0.0	39.4
6	*brilliant*	61	0	0.4	0.0	36.4
7	*circle*	54	1	0.3	0.0	34.4
8	*lady*	104	2	0.6	0.0	33.1
9	*cavalier*	51	0	0.3	0.0	30.4
10	*glad*	50	0	0.3	0.0	29.0

	1970s–2000s	2	1	PM2	PM1	Ratio
1	*lesbian*	237	1	2.2	0.0	372.6
2	*bar*	69	1	0.7	0.0	108.5
3	*activist*	58	1	0.5	0.0	91.2
4	*straight*	72	2	0.7	0.0	56.6
5	*marriage*	106	3	1.0	0.0	55.5
6	*openly*	30	1	0.3	0.0	47.2
7	*right*	146	5	1.4	0.0	45.9
8	*male*	28	1	0.3	0.0	44.0
9	*abortion*	40	0	0.4	0.0	37.5
10	*bisexual*	37	0	0.4	0.0	34.7

Table 20: Adj/noun collocates near the noun *gay*, comparison

Synonym	Total	1820	1840	1860	1880	1900	1920	1940	1960	1980	2000
wise	23,467	110.3	85.9	86.5	89.0	78.9	66.8	44.0	34.8	33.1	26.8
intelligent	14,601	45.2	47.0	50.1	49.7	39.6	35.5	32.0	29.4	25.6	19.7
smart	11,927	10.8	15.3	21.1	18.4	17.6	29.8	39.5	30.7	35.1	57.6
clever	8,526	6.1	10.7	15.4	21.8	33.0	29.6	19.8	18.5	20.5	18.1
sensible	8,352	50.1	30.7	28.4	25.4	20.1	19.0	17.7	16.5	13.1	10.4
shrewd	4,857	6.4	10.2	13.3	16.0	14.6	18.6	15.6	10.2	7.5	3.6
gifted	3,924	19.1	13.5	16.1	16.0	7.7	6.6	5.8	7.8	7.9	11.1
judicious	2,630	31.2	19.8	13.3	9.1	6.1	2.7	2.0	2.1	1.6	1.2
scholarly	1,938		0.5	2.8	4.4	7.7	2.9	4.6	7.3	7.1	5.2
cerebral	1,100	1.0	0.6	2.6	1.3	1.8	1.4	2.8	2.5	2.9	2.7
knowledgeable	794					0.1	0.1	0.5	4.7	6.4	5.9
brainy	214				0.2	0.5	0.8	0.5	1.0	0.5	1.8

Table 21: Frequency of synonyms of *intelligent* (per million words)

In addition to the 30,000+ synonym sets, it is also possible for users to create their own customised lists of semantically related words, and to use them, then, as part of their queries. For example, users could create a list of forty to fifty words relating to the body (*hair*, *leg*, *shoulder*, *finger*, *mouth*, *ear*, *foot*, *knee*, *neck*, *lip*, etc.) and then input this list through the web interface. They could then find all cases where one of these words is near (one to ten words to the left and/or right) a synonym of the verb *stroke*. COHA quickly indicates that the most frequent pairings are *pat*|*head* (96 tokens), *pat*|*back* (94), *rub*|*back* (80), *stroke*|*hair* (74), *pat*|*shoulder* (49), *rub*|*nose* (49), *rub*|*head* (38), and so on. As we can see, this allows us to move far beyond the simple 'strings of exact words' search facilities of other corpora. Here, we can look for 'any semantic field near any other semantic field', and see how these concepts and relationships have changed over time.

8. Changes in language and culture

The same features of COHA that allow us to look at semantic change (such as changing collocates) can also allow us to move beyond purely linguistically orientated searches, to look at changes in American history, culture and society. For example, consider Table 22, which compares the collocates of *women* in the 1830s to the 1890s (left) and the 1960s to the 2000s (right).

Note the emphasis in the 1800s on the 'moral' or 'vulnerable' quality of women, with collocates such as *noble*, *cultivated*, *devoted*, *pious*, *fair* and *abandoned*. In the late-1900s, on the other hand, the collocates of *women* are somewhat more prosaic (*Catholic*, *middle-class* and *working-class*) and they also relate to topics that might have been somewhat more taboo in the 1800s (e.g., *pregnant*, *battered* and *naked*).[18]

[18] Obviously, the frequency of a word as collocate is related to the overall frequency of the word itself in the corpus. For example, *African–American* is much more frequent as a collocate of *women* in the 1900s, simply because the word *African–American* is more

In Table 22, we have searched just for the exact string '[adjective] *women*', but in Table 23 we look for collocates of *women* – up to four words to the left and four to the right (and, of course, we could search up to ten left and ten right, using the corpus interface).

This time we compare the 1930s to the 1950s and the 1960s to the 1980s – two very different historical periods in terms of how women were viewed by society. In the 1930s to the 1950s period, note the emphasis on appearance (e.g., *wear* ('*women's wear*'), *fabrics* and *hips*) or women entering the workforce in World War II (e.g., *factories*, *coast* and *wartime*). In the 1960s to the 1980s period, on the other hand, there are references to the feminist movement and other related social movements (e.g., *liberation*, *minorities*, *abortion*, *AIDS* and *activists*).

Together with the comparisons of lexis (adjectives: *global*, *electronic*, *online* or *sexy* in the late-1900s) and even morphology (*–ism* nouns: *terrorism*, *communism* and *skepticism* in the mid- to late-1900s) seen above, the ability to compare collocates across time provides insights not only into semantic change, but also into cultural and societal changes in the United States over the past 200 years.

9. Comparisons with small corpora

9.1 Corpus size

With a large, robust corpus, we can greatly expand our horizons in terms of the types of language change that we can study. Consider for example Table 24, which reviews some of the phenomena that have been discussed previously in this paper.

As this table indicates, the data from COHA is quite robust. The features in rows one to five show the average number of tokens per decade for the word or construction in COHA. This is calculated by finding the total number of tokens and then dividing by twenty (the twenty decades from the 1810s to the 2000s). For example, there are 8,259 tokens of *bestow**, giving an average of 413 tokens (8259 / 20) in each of the twenty decades.[19]

I then assume a hypothetical four-million word corpus for the 1810s to the 2000s and calculate the average number of tokens that we would have in each decade (and which would obviously be one-hundredth the total from the 400-million word COHA corpus). In this small corpus, there would

---

frequent overall in the 1900s. A more sophisticated display and calculation (which may be available by the time this article is published) would take this into account, although many users already find displays like this fairly complicated, and there is a question about how much more complexity we want to add.

[19] As noted above, for reasons of space in this printed version, the tables in the earlier sections only show the token count for every other decade. The data under Table 24, on the other hand, show the total number of tokens in COHA.

	1830s–1890s	1	2	PM2	PM1	Ratio
2	*noble w.*	30	2	0.28	0.02	17.97
3	*true w.*	13	1	0.12	0.01	15.57
6	*cultivated w.*	11	0	0.10	0.00	10.09
7	*defenceless w.*	11	0	0.10	0.00	10.09
8	*loveliest w.*	8	1	0.07	0.01	9.58
9	*noblest w.*	10	0	0.09	0.00	9.17
11	*devoted w.*	10	0	0.09	0.00	9.17
12	*clever w.*	15	2	0.14	0.02	8.98
13	*excellent w.*	14	2	0.13	0.02	8.38
14	*pious w.*	7	1	0.06	0.01	8.38
16	*fair w.*	42	6	0.39	0.05	8.38
17	*abandoned w.*	9	0	0.08	0.00	8.25

	1960s–2000s	2	1	PM2	PM1	Ratio
1	*pregnant w.*	26	4	2.02	0.04	55.10
2	*battered w.*	71	0	0.54	0.00	54.36
4	*african-american w.*	61	0	0.47	0.00	46.70
5	*professional w.*	47	1	0.36	0.01	39.24
7	*black w.*	49	13	3.77	0.12	31.66
8	*naked w.*	69	2	0.53	0.02	28.80
10	*divorced w.*	28	0	0.21	0.00	21.44
12	*muslim w.*	27	0	0.21	0.00	20.67
13	*middle-class w.*	26	0	0.20	0.00	19.91
14	*catholic w.*	25	0	0.19	0.00	19.14
15	*working-class w.*	24	0	0.18	0.00	18.37
16	*homeless w.*	22	1	0.17	0.01	18.37

Table 22: Adjectival collocates of *women* ('ADJ + *women*'), 1830s–1890s *versus* 1960s–2000s

	1930s–1950s	1	2	PM2	PM1	Ratio
1	*wear*	31	1	0.42	0.01	30.8
2	*misses*	15	0	0.20	0.00	20.4
3	*fabrics*	13	0	0.18	0.00	17.6
6	*factories*	13	1	0.18	0.01	12.9
7	*coast*	9	0	0.12	0.00	12.2
8	*wartime*	9	0	0.12	0.00	12.2
12	*hips*	8	0	0.11	0.00	10.8
13	*leisure*	7	0	0.10	0.00	9.5

	1960s–1980s	2	1	PM2	PM1	Ratio
1	*blacks*	80	0	1.09	0.00	109.4
2	*liberation*	10	1	1.45	0.01	106.5
4	*minorities*	61	1	0.83	0.01	61.3
7	*abortions*	36	1	0.49	0.01	36.1
9	*risk*	19	0	0.26	0.00	25.9
10	*aids*	16	0	0.22	0.00	21.8
11	*liberationists*	14	0	0.19	0.00	19.1
13	*activists*	12	0	0.16	0.00	16.4

Table 23: Noun collocates of *women* ('NOUN near *women*'), 1930s–1950s *versus* 1960s–1980s

	Feature	Example	Table/ Figure	Tokens COHA	Tokens [small] = COHA / 100
SINGLE OR COMBINED FREQUENCY				Average tokens per decade	
1	Lexical: single form	*bestow** *mustn't*	F1 F2	413 260	4 3
2	Morphology: comparing two forms	*have burnt/burned* *he lighted/lit*	T7 T8	71 120	1 1
3	Syntax: high frequency	*have to* V *going to* V	F10 T16	37,700 6,201	**377** **62**
4	Syntax: medium-frequency	± split infinitive *get* passive	T13 T15	1,682 1,683	**17** **17**
5	Syntax: low frequency	*help* [p*] ±to [v*] *accustomed to* [V/V-*ing*]	T17 T18	675 177	7 0.5
MULTIPLE ENTRIES / FREQ. BY DECADE				# entries: frequency ≥ 20 (1 token per decade)	
6	Lexical/morphology	**ism* words	T5	395	13
7	Phraseology	[v*] *up*	T10	34	0.3
8	Semantics: collocates	Collocates of *gay*	T22	249	0
COMPARE HISTORICAL PERIODS				Average frequency of top 10 entries	
9	Lexical	All adjectives	T4	589	6
10	Semantics: collocates	Collocates of *gay*	T23	74	0.7
11	Discourse: collocates	ADJ + *women*	T26	13	0.1

Table 24: Number of tokens for different phenomena, in COHA and small four-million word corpus

be, typically, only a handful of tokens per decade. For example, a four-million word corpus would only have (on average) four *bestow** tokens per decade, three *mustn't* tokens, one *have burnt/burned* token, one '[PRON] + *lighted/lit*' token, seven '*help* [PRON] ±to [VERB] tokens, and 0.5 '*accustomed to* [V/V-*ing*]' tokens.

In terms of significance, the data from COHA often yields statistically significant results where those from a small four-million word corpus would not. To take just one example, if we calculate the chi square for the number of tokens for 'accustomed to [V/V-*ing*]' in each decade (see Table 15 and Figure 12), in COHA we obtain a chi square value of 412.157, which is significant at $p \leq .000001$. In the small four-million word

corpus, on the other hand, if we divide the number of tokens by 100 for each decade and then use these figures, we obtain a chi square value of 24.714, which is only significant at $p \leq .17$ (i.e., not statistically significant). The only phenomena where there are probably enough tokens to yield statistical significance (at $p \leq .05$) is for the high frequency syntactic constructions (e.g., '*have to* V' and '*going to* V' versus '*will* V'), where there would be 377 and 108 tokens respectively, and perhaps also some of the medium-frequency syntactic constructions (e.g., the split infinitive and the *get* passive), where there would be seventeen tokens per decade.

To interpret rows six to eight, recall that in addition to the overall frequency for all matching words or strings, it is also possible to see the frequency of each matching form, string or collocate in each decade. In Table 24, I count the number of entries that have at least twenty tokens overall (or an average of one token per decade). For example, row six shows that there are 395 **ism* words that occur at least twenty times in COHA. All things being equal, there would have to be about 2,000 tokens in COHA for the same word to occur twenty times in our small four-million word corpus (2000 / 100 = 20). As we see in row six, there are only thirteen **ism* words in COHA that occur at least 2,000 times, and a list with just thirteen entries in the small corpus would probably be too limited to be of much interest.

In rows nine to eleven, I compare the words or collocates in one section of the corpus (typically three to four decades) against another section. In each case, I have taken the average number of tokens for the first ten entries in the left side of the tables indicated. For example, in row nine (adjectives in the 1870s to the 1910s compared to the 1970s to the 2000s), the average frequency for the top ten entries in the left side of Table 4 is eighty-five. In a four-million word corpus, on the other hand, there would be less than one token per word, which would be far too small to reveal much that is of interest about the lexical and semantic changes.

It is no surprise that small corpora like the Brown family (Brown, LOB, FROWN, FLOB), ARCHER, CONCE and the DCPSE are used almost exclusively to research high-frequency (and select medium-frequency) syntactic constructions. While they have led to many highly insightful studies of these constructions (e.g., modals, auxiliary verbs and relative pronouns) during the past decade or two, I would argue that these small corpora are largely inadequate for research on lexical, morphological, semantic and low-frequency syntactic change.

9.2 Data granularity

As we have seen, there is often not enough data in a small one- to five-million word corpus to yield statistically significant results, if we compare tokens by decade. One way around this problem might be to group the number of tokens into thirty to forty year blocks (giving us larger numbers to work with), rather

than comparing the data by decade. The downside of this, of course, is that by looking at changes every thirty to forty years, we have less 'granularity' in terms of knowing when a change has occurred, and it is more difficult to see the sequencing of related changes.

For example, consider Table 25 and Figure 15, which look at the shift from '*to*-V' to 'V-*ing*' with *start* and *begin* ('*we* started / began *to walk away*' → '*we* started / began *walking away*'), and which is based on nearly 40,000 tokens with *start* and nearly 100,000 tokens with *begin* (for an overview of changes with V/V-*ing*, see Rohdenburg, 2006; and De Smet, 2008). We see that in one single decade, the 1920s, the percentage of 'V-*ing*' with *start* nearly doubled (23 percent to 41 percent). In a corpus with data from just every thirty years, we would not know if the change occurred in the 1920s, or perhaps the 1910s or the 1930s.

As mentioned, granularity is also important in terms of looking at related shifts. For example, the largest increase in 'V-*ing*' with *start* occurred in the 1920s, whereas with the emotion verbs *love*, *hate* and *like* occurred somewhat later (1950s to the 2000s), as we see with the data under Table 26 for the verb *hate*.

Only by tracking language change every decade would we notice that the one change occurred before the other, and then (hopefully) begin to consider possible motivations for this sequence of changes in terms of analogy, grammaticalisation, specific functional and stylistic motivations, and so on. If we sample the data just every thirty to forty years, we may not be able to compare and analyse related shifts in the language.

10. Corpus architecture: Google Books

Under Section 6.1, we saw the important role that size and data granularity play in providing robust data. However, corpus size is obviously not everything: a text archive might be hundreds or thousands of times larger than COHA, and yet be much less useful than COHA for looking at language change.

As an example of this, let us consider Google Books,[20] including the interface that was introduced in late-2010.[21] Using this interface, linguists can look for changes in 500 billion words of American English from the 1800s to the 1900s, which is, of course, much larger than the 400-million word COHA corpus.

So why not use these larger resources instead of COHA? The answer lies with corpus architecture. With unstructured corpora like Google Books and with text archives, it would be difficult or even impossible to study

[20] See: http://books.google.com/
[21] See: http://books.google.com/ngrams/

	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000
start to V	514	792	1,493	1,576	1,845	1,983	1,784	1,853	2,256	2,984	3,186
start V-*ing*	34	159	439	1,110	1,499	1,780	1,926	2,150	2,363	3,792	4,340
% *start* V-*ing*	6.2	16.7	**22.7**	**41.3**	44.8	47.3	51.9	53.7	51.2	56.0	57.7
begin to V	6,587	6,644	6,856	7,399	7,612	6,995	7,527	6,797	7,657	7,198	6,244
begin V-*ing*	569	802	1,180	1,570	1,702	1,688	1,999	2,118	2,558	2,742	3,005
% *begin* V-*ing*	8.0	10.8	14.7	17.5	18.3	19.4	21.0	23.8	25.0	27.6	32.5

Table 25: 'V-*ing*' versus '*to*-V' with *start* and *begin*

	1860	1870	1880	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000
*to*_v	86	129	156	178	281	383	437	419	346	372	323	338	288	300	400
V-*ing*	1	8	13	12	22	30	33	49	49	54	60	77	109	138	245
% V-*ing*	0.01	0.06	0.08	0.06	0.07	0.07	0.07	0.10	0.12	**0.13**	**0.16**	**0.19**	**0.27**	**0.32**	**0.38**

Table 26: 'V-*ing*' versus '*to*-V' with *hate*

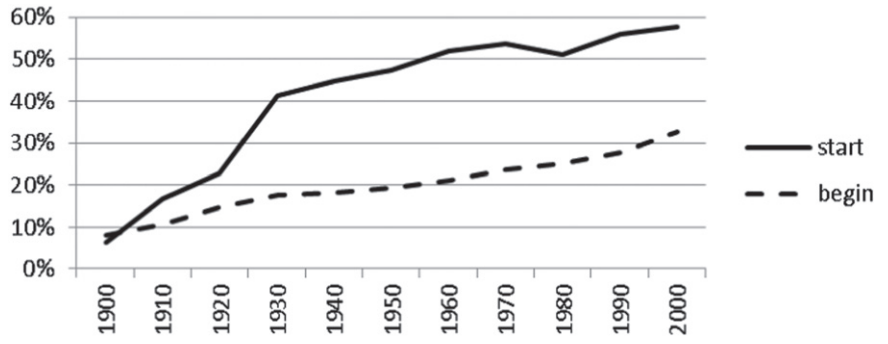

Figure 15: 'V-*ing*' versus '*to*-V' as complements of *start* and *begin*

the wide range of language changes that can be studied quickly and easily with COHA. With Google Books,[22] it certainly would be possible to find the frequency by decade for exact words and phrases (see row one of Table 24) and exact strings that compare morphology (see row two of Table 24); for example, *have burnt* and *have burned*.[23] All of the other searches, however, would either be impossible or extremely cumbersome.

Using Google Books, it would be impossible to conduct the searches shown in rows six to eleven of Table 24. Unlike COHA, Google Books allows users to find the frequency of words and phrases – but only once the user already knows the exact word or phrase that he or she is looking for. In rows six to eleven, we do not know what the words will be; the COHA architecture finds them for us. An additional problem is that research on phenomena shown on row eight and rows ten to eleven, which deal with collocates, would at best be extremely cumbersome. With Google Books, we would have to write a program to input the node word into the search interface, retrieve the hits (until we are blocked by Google or until we have gone through all ten pages of the search results), find and copy the four to five words on each side, eliminate high frequency words like *the*, *with* or *to*, import the collocates into a database or hash file, and then compare the data from the two periods. With COHA, all of this is done 'behind the scenes' in just a few seconds.

With Google Books, it is also difficult or impossible to carry out studies of morphological change, since Google Books does not allow users to search by wildcard, as in the *–ism* search (row six under Table 24). It is also difficult or impossible to carry out syntactic research, because the texts in Google Books are not lemmatised or tagged for part-of-speech (see rows three to five under Table 24). For example, if we are interested in the

[22] We focus here on Google Books, but most if not all of the limitations listed here would apply to other text archives of historical magazines, newspapers and books (e.g., Project Gutenberg) as well.

[23] Other forms like 'had burnt' and 'has burned' would have to be separate searches, since Google Books (like Google web search) only allows users to search for exact strings.

rise of the '*into* V-*ing*' construction ('we talked / tricked / persuaded him into staying') – which is composed of 'verb + NP + *into* + V-*ing*' – the only element that we can search for would be the word *into*, which would of course massively over-generate results. With COHA, we can quickly carry out this search ('[*vv**] [*p**] *into* [*v?g**]') to find all 1,669 tokens with an embedded clause subject that is a pronoun.

11. Conclusion

As we have seen, small one- to five-million word corpora of Late Modern English have been used almost exclusively to look for high-frequency syntactic constructions, but it is difficult or impossible to use them to look at lexical, morphological and semantic change, or low- (and some medium-) frequency syntactic constructions. Unstructured, unannotated corpora and text archives (like Google Books) may be extremely large in terms of their size, but their architecture and interface is too rudimentary to allow searches for anything beyond exact words and phrases. With the 400-million word Corpus of Historical American English, on the other hand, we can quickly and easily conduct a wide range of research on lexical, morphological, syntactic and semantic change, and this allows us to expand significantly our horizons in terms of what can be done with historical corpora.

References

Aarts, B., J. Bowie and S. Wallis. Forthcoming. 'Profiling the English verb phrase over time: modal patterns' in M. Kytö and I. Taavitsainen (eds) Corpus Linguistics and the History of English. Cambridge: Cambridge University Press.

Biber, D., E. Finegan, D. Atkinson, A. Beck, D. Burges and J. Burges. 1994. 'ARCHER and its challenges: compiling and exploring a representative corpus of historical English registers' in U. Fries, G. Tottie and P. Schneider (eds) Creating and Using English Language Corpora, pp. 1–13. Amsterdam: Rodopi.

Davies, M. 2008 'Spanish and Portuguese corpus linguistics', Studies in Hispanic and Lusophone Linguistics 1 (1), pp. 149–86.

Davies, M. 2009a. 'The 385+ million word Corpus of Contemporary American English (1990–2008+): design, architecture and linguistic insights', International Journal of Corpus Linguistics 14 (2), pp. 159–90.

Davies, M. 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 (2), pp. 443–5.

Davies, M. 2010a. 'More than a peephole: using large and diverse online corpora', International Journal of Corpus Linguistics 15 (3), pp. 405–11.

Davies, M. 2010b. 'Creating useful historical corpora: a comparison of CORDE, the Corpus del Español, and the Corpus do Português' in A. Enrique-Arias (ed.) Diacronía de las lenguas iberorromances: nuevas perspectivas desde la lingüística de corpus, pp. 137–66. Frankfurt and Madrid: Vervuert/Iberoamericana.

Davies, M. 2011. 'The Corpus of Contemporary American English as the first reliable monitor corpus of English', Literary and Linguistic Computing 25 (4), pp. 447–65.

De Smet, H. 2008. Diffusional Change in the English System of Complementation: Gerunds, Participles and for… to-infinitives. Unpublished dissertation. University Leuven.

Firth, J.R. 1957. Papers in Linguistics 1934–1951. Oxford: Oxford University Press.

Hundt, M. 2001. 'What corpora tell us about the grammaticalisation of voice in *get*-constructions', Studies in Language 25 (1), pp. 49–87.

Hundt, M and G. Leech. Forthcoming. "Small is beautiful" – on the value of standard reference corpora for observing recent grammatical change' in T. Nevalainen and E.C. Traugott (eds) Handbook on the History of English: Rethinking Approaches to the History of English. Oxford: Oxford University Press.

Hunston, S. and G. Francis. 2000. Pattern Grammar: A Corpus-driven Approach to the Lexical Grammar of English. Philadelphia: John Benjamins.

Kytö, M., J. Rudanko and E. Smitterberg. 2000. 'Building a bridge between the present and the past: a corpus of 19th-century English', ICAME Journal 24, pp. 85–97.

Leech, G, M. Hundt, C. Mair and N. Smith. 2009. Change in Contemporary English: A Grammatical Study. Cambridge: Cambridge University Press.

Mair, C. 2006. Twentieth-century English: History, Variation and Standardization. Cambridge: Cambridge University Press.

Mair, C. 1997. 'Parallel corpora: a real-time approach to the study of language change in progress' in M. Ljung (ed.) Corpus-based Studies in English, pp. 195–209. Amsterdam: Rodopi.

Rohdenburg, G. 2006. 'The role of functional constraints in the evolution of the English complementation system', Linguistic Insights – Studies in Language and Communication 39, pp. 143–66.

Rudanko, J. 2010. 'Explaining grammatical variation and change: a case study of complementation in American English over three decades', Journal of English Linguistics 38 (1), pp. 4–24.

Schneider, E.W. 1992. 'Who(m)? Case marking of wh-pronouns in written British and American English' in G. Leitner (ed.) New Directions in English Language Corpora: Methodology, Results, Software Developments, pp. 231–45. Berlin: Mouton de Gruyter.

Yáñez-Bouza, N. Forthcoming. 'ARCHER past and present (1990–2010)', ICAME Journal 35.