The 400 million word Corpus of Historical American English (1810–2009)

Mark Davies
Brigham Young University

The 400 million word Corpus of Historical American English (1810–2009) provides researchers with an extremely robust set of data for Late Modern English. The corpus is composed of fiction, magazines, newspapers, and non-fiction books, and its genre balance stays roughly the same from decade to decade. Because of its size and its advanced architecture and interface, it allows researchers to look at an extremely wide range of changes – many of which would not be possible with a small 2–4 million word corpus. These include the frequency of any word or phrase by decade and mass comparison of all words in different periods (to examine lexical changes), morphological shifts (via wildcards and pattern matching), syntactic shifts (due to very accurate lemmatization and part of speech tagging), and semantic change (by comparing collocates over time, as well as searches that use data from the integrated thesaurus and customized word lists).

1. Introduction

The 450 million word Corpus of Contemporary American English (COCA), which was released in 2008, allows in-depth research on changes in American English since 2009 (see Davies 2009, 2010). Nevertheless, there was no large corpus of American English that extended back further than the early 1990s. Recently, however, we released the Corpus of *Historical* American English (COHA), which is now freely available online (http://corpus.byu.edu/coha). COHA is composed of 400 million words of text of American English in more than 100,000 texts, comprising fiction, popular magazines, newspapers, and non-fiction books from 1810–2009:

Table 1. Composition of COHA by genre and decade

Decade	Fiction	Magazines	Newspapers	Nonfic books	Total	% 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

COHA is balanced by genre across the decades. For example, fiction accounts for 48–55% of the total in each decade (1810s–2000s), and the corpus is balanced across decades for sub-genres and domains as well (e.g. by Library of Congress classification for non-fiction; and by sub-genre for fiction – prose, poetry, drama, etc). This balance across genres and sub-genres allows researchers to examine changes and be reasonably certain that the data reflects actual changes in the 'real world', rather than just being artifacts of a changing genre balance. Much more data on the composition of the corpus can be found at the corpus website.

In this paper, we will show how COHA can be used to research a wide range of phenomena relating to lexical, morphological, phraseological, syntactic, and semantic changes in American English. In the concluding section, we will compare COHA to other corpora and to unstructured corpora and text archives. We will

compare COHA to these other resources in terms of size, textual granularity, and architecture, and we will see that COHA allows us to obtain data on historical changes in American English in ways that would not be possible with any other tool.

2. Lexical change

At the most basic level, COHA allows us to see the frequency of any word or phrase in each of the twenty decades in the corpus (1810s–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 frequency over time. Examples of the frequency charts are shown in Figures 1–3, where we see words that have been decreasing in frequency since the 1800s (words starting with *bestow*), a phrase that peaked about 100 years ago (*must n't*), and words that have been increasing over time (*teenager* and *teenagers*). As in shown in Figure 3, the frequency is often a function of historical, cultural, or societal changes, which impact on the language (in this case, changing societal perceptions and explicit labeling of those in this age group).[1]

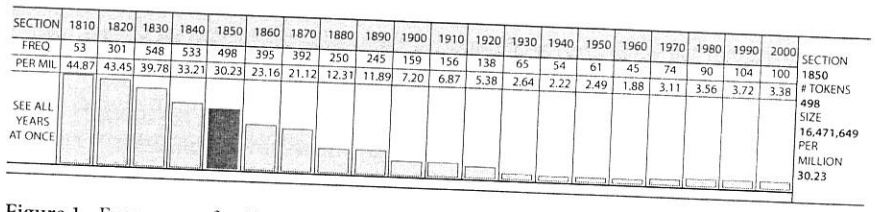

Figure 1. Frequency of *sublime*, 1810s–2000s

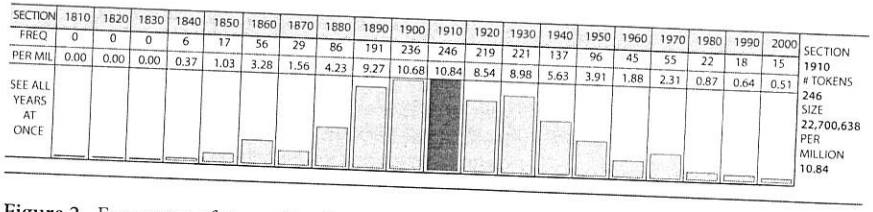

Figure 2. Frequency of *steamship*, 1810s–2000s

1. Note that in each case, one of the bars (representing a particular decade) is highlighted. Users can select a decade and see more detailed frequency data to the right. Note also that of course the frequency charts are 'normalized', which means that they are based on the token frequency per million words in each decade.

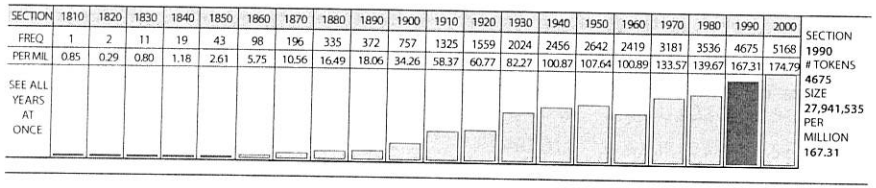

Figure 3. Frequency of *a lot of*, 1810s–2000s

Of course the corpus interface shows not just 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 click on the 1850s bar in Figure 1 above to see all 490 tokens of *sublime*, as in Table 2.[2]

Table 2. Keyword in Context (KWIC) entries

Date	Genre	Source	Keyword in context
1854	FIC	Eventide ASeries	poor, deluded victims of a false religion, and bring them all under his **sublime** sway and holy dominion. At length, Miss Gaddie was called on to sing
1854	FIC	RolloInSwitzerland	The attention of Mr. George, however, was attracted by the more grand and **sublime** features of the view which were to be seen in the other direction – the lake,
1854	FIC	Nugæ	that deep along whose sandy shore Are strewed bright hopes, gay visions, schemes **sublime**, Brilliant imaginings from fancy's store, Wild aspirations, follies, ghastly crimes
1854	FIC	RhymesWithReason	, replete with bitter sadness, Heard the sweet note that filled the air, **sublime**, And felt a thrill run through his frame of gladness; The fevered pulse
1855	FIC	WorksEdgarAllan	was as wide as the great hall of audience in your palace, O most **sublime** and munificent of the Caliphs. Its body, which was unlike that of ordinary

2. Note that because of space limitations 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 in the center of the line. In the web interface it is also possible to click on any KWIC entry and get up to 120 words of context.

For more detailed investigations of word and phrase frequency, users can also see the frequency in each individual year from 1810–2009. For example, the following chart shows that the word *depression* is the most frequent in the 1930s. Users can click on the [1930s] heading to see the frequency in each year of the 1930s. In this case, as Figure 4 shows, they would see that its frequency is highest in 1931–1933, which again corresponds to external changes in American history and society.

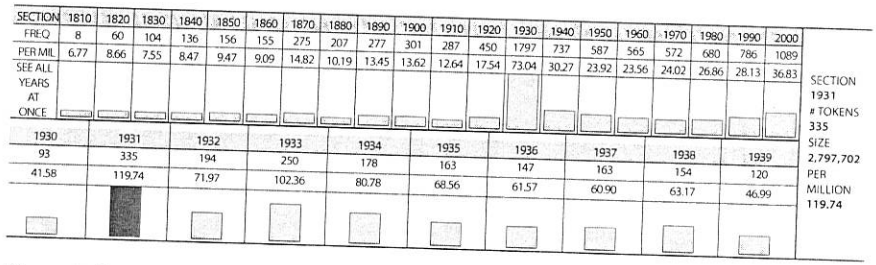

Figure 4. Frequency of *depression* by decade and year

In the examples above, we found the frequency of a particular word or phrase. But because 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 don't have any idea what these words might be. For example, Table 3 shows verbs that are more frequent in the 1870s–1900s (i.e. the decade from 1900–1909) than in the 1970s–2000s (left side) and those which are more frequent in the 1970s–2000s (right side).[3, 4]

3. We should briefly explain the organization of the data, since similar are found in other sections of the paper. Taking the example of *betake* (word #9 at the left), we see that it occurs 58 times from 1870–1909 (Section 1) and just one time 1970–2009 (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 normalized figures in the two sections and see that *betake* is about 76 times more frequent 1870–1909 than it is 1970–2009. The results are ranked in terms of this ratio between the two sections of the corpus

4. Note that we have looked for just the base form of the verb (e.g. *trigger*, but not *triggered* or *triggering*). In addition, in this list we see verbs that appear simply because they are spelled differently in the two periods (e.g. *ketch* as an informal, dialectal form of *catch*, or *intrust* as an earlier form of *entrust*). If the results were lemmatized (which it is possible to do), the forms *ketch* and *intrust* would be grouped with *catch* and *entrust*. We have also 'smoothed' the data to allow for division by zero, when a word does not occur in the other section. Finally, note that not all entries that appear in the online corpus are shown here (note the skipped numbers in the leftmost column of the entries to the left).

Table 3. Comparison of verbs, 1870s–1920s and 1970s–2000s

	1870s–1900s	1	2	PM1	PM2	RATIO
1	KETCH	139	1	1.70	0.01	181.7
5	INTRUST	110	0	1.35	0.00	134.8
6	SUBSERVE	67	1	0.82	0.01	87.6
8	UNDECEIVE	60	1	0.74	0.01	78.4
9	BETAKE	58	1	0.71	0.01	75.8
10	RECONNOITRE	99	2	1.21	0.02	64.7
16	CONTROVERT	38	1	0.47	0.01	49.7
18	PERSONATE	40	0	0.49	0.00	49.0
20	CHAFF	33	1	0.40	0.01	43.1
21	DOGMATIZE	33	1	0.40	0.01	43.1
22	ENTREAT	97	3	1.19	0.03	42.3
30	PREDICATE	27	0	0.33	0.00	33.1
31	PREMISE	25	1	0.31	0.01	32.7
34	SIGNALIZE	24	1	0.29	0.01	31.4
35	SOLEMNIZE	24	1	0.29	0.01	31.4
36	REPINE	47	2	0.58	0.02	30.7
37	DISFRANCHISE	23	1	0.3	0.0	30.1
38	SUPERINTEND	154	7	1.9	0.1	28.8

	1970s–2000s	2	1	PM2	PM1	RATIO
1	MONITOR	731	0	6.85	0.00	685.5
3	TRIGGER	384	0	3.60	0.00	360.1
6	MAXIMIZE	311	0	2.92	0.01	291.6
7	ACCESS	310	0	2.91	0.01	290.7
4	STABILIZE	278	0	2.61	0.00	260.7
5	SKI	265	0	2.48	0.00	248.5
8	UPGRADE	247	0	2.32	0.00	231.6
9	PINPOINT	241	0	2.26	0.00	226.0
10	BROADCAST	220	0	2.06	0.00	206.3
14	OPT	207	0	1.94	0.00	194.1
15	RETHINK	202	0	1.89	0.00	189.4
16	BYPASS	200	0	1.88	0.00	187.6
18	MOTIVATE	190	0	1.78	0.00	178.2
20	DIAL	187	0	1.75	0.00	175.4
21	PROGRAM	163	0	1.53	0.00	152.9
22	REPLICATE	162	0	1.52	0.00	151.9
30	PARK	327	2	3.1	0.0	125.1
32	DOWNLOAD	127	0	1.2	0.0	119.1

As we can see, COHA allows us to quickly and easily compare 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 verbs like *access, broadcast, program,* or *download* in the table above, which relate to scientific or technological advances in the late 1900s.

3. Morphological change

COHA also allows us to search the 400 million words to see changing patterns in terms of word formation. For example, Table 4 shows changes during the last 200 years in the frequency of words ending in **ist*, and relating to occupations or other categories of individuals[5]. Note the decrease with a few words since the 1800s (*philanthropist, capitalist,* and *geologist*), but also those occupations or categories that have increased much more in the mid to late 1900s (e.g. *psychiatrist, activist,* and *therapist*), which may provide interesting insight into cultural and societal changes in the United States.

As with simple words, COHA also allows us to compare word forms across different time periods. For example, Table 5 compares **ist* words (for categories of people) in the period 1850–1909 and 1970–2009. Again, we see interesting shifts in American English and American culture and society generally.

While the preceding tables relate to a morphological subset of lexical items (terms for people, ending in **ist*), with COHA it is also possible to compare alternate word forms themselves, such as different verb forms. For example, Table 6 compares the relative frequency of *have proved* and *have proven* by decade (all forms of have: *have, has, had,* etc), and is based on 6,477 tokens. Figure 5 shows that while there was a fair amount of variation through the 1950s, there has been a clear increase in the strong form *proven* since that time, and *proven* is now 6–7 times as frequent as it was 50–60 years ago.

5. Note that not all entries are shown, since some do not relate to the individuals (e.g. *list, waist, fist*). Also note that the raw frequency (number of tokens) is shown here, but it is also possible to see the normalized frequency by tokens per million, which is indicated here by color (darker color = 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.

Table 4. *-ist* nouns referring to people

		TOTAL	1810	1830	1850	1870	1890	1910	1930	1950	1970	1990
7	SCIENTIST	4622				69	68	178	386	452	432	431
9	TOURIST	3362		26	21	41	107	83	225	343	354	371
10	JOURNALIST	2917		10	47	70	79	142	116	150	307	360
12	SPECIALIST	2830				12	34	118	154	240	256	267
14	PSYCHIATRIST	1962						2	49	268	318	375
15	PSYCHOLOGIST	1961				14	62	101	105	203	112	197
16	DENTIST	1956	2	8	21	20	35	85	137	219	171	76
17	CHEMIST	1851	1	20	41	57	71	142	132	146	74	11
20	COLUMNIST	1601							97	208	168	151
21	PHYSICIST	1553				10	25	45	138	168	110	43
23	NATIONALIST	1480					31	31	96	402	122	48
24	CAPITALIST	1313		35	66	102	76	87	109	56	31	55
26	ACTIVIST	1030								6	143	379
28	THERAPIST	982							10	19	68	18
29	PHILANTHROPIST	887		75	75	46	52	47	38	21	11	73
32	GEOLOGIST	763		10	54	62	26	29	26	39	25	163
34	RECEPTIONIST	762							4	65	98	525

Table 5. *-ist* nouns referring to people

	1850s–1900s	1	2	PM1 1	PM2 2	RATIO
1	DIPLOMATIST	217	2	1.9	0.0	100.5
2	AUTOMOBILIST	84	0	0.7	0.0	73.0
3	ALCHYMIST	56	0	0.5	0.0	48.7
4	AGRICULTURIST	103	2	0.9	0.0	47.7
5	DUELLIST	38	1	0.3	0.0	35.2
6	PHYSIOGNOMIST	35	1	0.3	0.0	32.4
7	LYRIST	31	1	0.3	0.0	28.7
8	ROMANIST	30	0	0.3	0.0	26.1
11	CASUIST	28	0	0.2	0.0	24.3
13	PANTHEIST	36	2	0.3	0.0	16.7
15	DOGMATIST	17	1	0.2	0.0	15.8
16	DAGUERREOTYPIST	17	0	0.2	0.0	14.8
17	ANNALIST	30	2	0.3	0.0	13.9
18	ARTILLERIST	15	1	0.1	0.0	13.9
20	SECESSIONIST	73	5	0.6	0.1	13.5

	1970s–2000s	2	1	PM2	PM1 1	RATIO
1	PSYCHIATRIST	1119	1	10.5	0.0	1207.8
2	THERAPIST	888	1	8.3	0.0	958.5
3	ACTIVIST	971	0	9.1	0.0	910.5
4	COLUMNIST	878	0	8.2	0.0	823.3
5	LEFTIST	523	1	4.9	0.0	564.5
6	RECEPTIONIST	579	0	5.4	0.0	543.0
7	FEMINIST	226	1	2.1	0.0	243.9
9	RAPIST	195	0	1.8	0.0	182.9
11	CARDIOLOGIST	157	0	1.5	0.0	147.2
12	INDUSTRIALIST	156	0	1.5	0.0	146.3
13	PACIFIST	150	0	1.4	0.0	140.7
14	ENVIRONMENTALIST	148	0	1.4	0.0	138.8
16	DERMATOLOGIST	126	0	1.2	0.0	118.2
19	GYNECOLOGIST	121	0	1.1	0.0	113.5
20	BASSIST	120	0	1.1	0.0	112.5

Table 6. *Have proven vs. have proved*

have +	1810	1820	1830	1840	1850	1860	1870	1880	1890	1900
proved	22	134	245	303	330	324	332	323	345	403
proven	0	1	1	1	4	8	9	21	20	35
% proven	0.00	0.01	0.00	0.00	0.01	0.02	0.03	0.06	0.05	0.08
have +	**1910**	**1920**	**1930**	**1940**	**1950**	**1960**	**1970**	**1980**	**1990**	**2000**
proved	386	388	325	349	359	314	246	269	199	202
proven	39	45	22	23	22	35	57	71	110	155
% proven	0.09	0.10	0.06	0.06	0.06	0.10	0.19	0.21	0.36	0.43

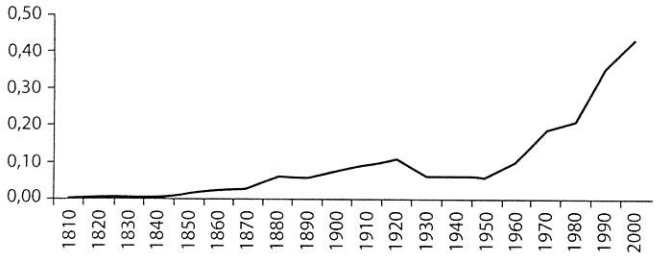

Figure 5. *Have proven vs. have proved*

Another example of morphological change is the relative frequency of [pron] + *dived/dove* in each decade from the 1920s–2000s (e.g. *he dived/dove into the pool*), and this is based on 602 tokens. Table 7 indicates that there is some variation between the two forms from the 1920s–1940s (and especially the 1810s–1910s; not shown here). But as Figure 6 shows, there is a clear increase in *dove* as the simple past form of *dive* since the 1930s, to nearly three times what it was 60–70 years ago.

Table 7. *[pron] + dove vs. [pron] + dived*

[pron] +	1920	1930	1940	1950	1960	1970	1980	1990	2000
dived	33	36	52	27	24	23	25	26	25
dove	6	13	14	10	16	19	33	41	39
% dove	0.15	0.27	0.21	0.27	0.40	0.45	0.57	0.61	0.61

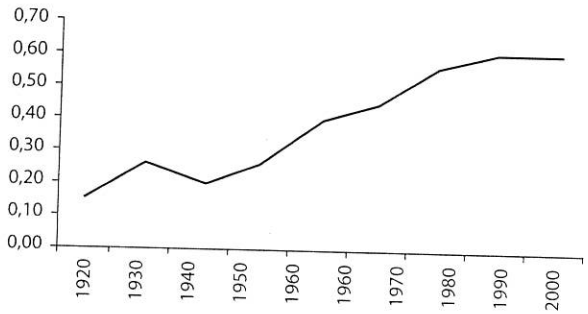

Figure 6. *[pron] + dove vs. [pron] + dived*

4. Phraseological change

In this section we expand our scope somewhat and look at localized patterns of words (phraseology), and we will expand that in the following section when we consider syntactic change. Consider first the phrase *a most* ADJ NOUN (*a most delicate operation, a most unruly child*). As Figure 7 shows, this has decreased markedly since the 1800s, and this signals an interesting stylistic shift in the language.

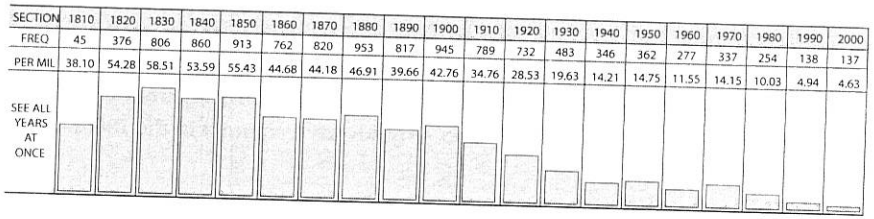

Figure 7. *a most* [ADJ] [NOUN]

In addition to seeing a chart display, users can also see the frequency of each matching string in each decade (see Table 8). They can see 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 compares phrasal verbs with the particle *out* in the 1910s–1940s and the 1970s–2000s. In just 2–3 seconds, COHA finds more recent phrasal verbs like *freak out* or *phase out*, and now-obsolete and strange-sounding verbs like *bar out* (*the fences barred out the hurrying figure*), *crop out* (*lest eagerness should crop out in spite of her*), and *gleam out* (*the lights of the city gleamed out*).

Finally, COHA can provide insight into changes into the types of phraseological 'frames' (see Hunston & Francis 2000). In these cases, we are not looking either at

Table 8. *a most* [ADJ] [NOUN]

	TOTAL	1820	1840	1860	1880	1900	1920	1940	1960	1980	2000
A MOST IMPORTANT PART	52	2	2	1	4	8	6	4	2		
A MOST UNUSUAL THING	20			1	4	2		1	2		1
A MOST CORDIAL WELCOME	17			1	2	2	1	1			
A MOST IMPORTANT ELEMENT	15		1		2	1	1				
A MOST DIFFICULT TASK	13			1	1	3				1	
A MOST EXTRAORDINARY THING	13	3		2	1		2				
A MOST IMPORTANT FACTOR	13			1	3	2	4				

individual words or regular syntactic constructions, but rather at 'frames' in which lexical items may appear. For example, consider Table 10, which compares words occurring in the frame [. *ly.[r*],] (i.e. full stop + *-ly* adverb + comma) in the 1840s–1910s and the 1950s–2000s.

5. Syntactic change

Because COHA is lemmatized and because it is 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 prescriptive rules. The first rule we will consider here is the shift from *may* to *can* for permission (as measured by the ratio of the two phrases *may I* and *can I*). Table 11 contains the data from 13,346 tokens from 1900 to 2009, and Figure 8 shows perhaps more clearly the shift from *may* to *can* during this time. Notice that although there are some increases and decreases in terms of the percentage of *can* (perhaps due to the varying effect of the prescriptive rule at times), the gray trendline shows the overall increase in *can*, and we see that it is now 50% more common than it was 100 years ago.

Table 9. Phrasal verbs with *out*

	1910s–1940s	1	2	PM1	PM2	RATIO
1	BAR OUT	22	0	0.23	0.00	22.6
2	CROP OUT	19	0	0.20	0.00	19.5
3	GLEAM OUT	15	1	0.15	0.01	16.4
4	HEW OUT	25	2	0.26	0.02	13.7
5	PRICK OUT	12	0	0.12	0.00	12.3
6	TOLL OUT	12	0	0.12	0.00	12.3
7	JAR OUT	11	1	0.11	0.01	12.1
8	SAVE OUT	38	4	0.39	0.04	10.4
9	FALTER OUT	10	0	0.10	0.00	10.3
10	FOLLOW OUT	83	9	0.85	0.08	10.1

	1970s–2000s	2	1	PM2	PM1	RATIO
1	FREAK OUT	189	0	1.77	0.00	177.2
2	PHASE OUT	92	0	0.86	0.00	86.3
3	CHURN OUT	152	2	1.43	0.02	69.4
4	BOTTOM OUT	66	0	0.62	0.00	61.9
5	OPT OUT	125	2	1.17	0.02	57.0
6	TOP OUT	61	1	0.57	0.01	55.7
7	STRESS OUT	56	0	0.53	0.00	52.5
8	CHILL OUT	51	1	0.48	0.01	46.5
9	CASH OUT	48	0	0.45	0.00	45.0
10	CHICKEN OUT	41	1	0.38	0.01	37.4

Table 10. [period + *-ly* adverb + comma]

	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	0.00	10.7
8	. SINGLY ,	11	1	0.07	0.01	8.6
9	. PRACTICALLY ,	106	10	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. *Can I* vs. *may I*

	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000
may I	488	485	498	460	451	550	456	390	473	327	348
can I	559	577	543	572	731	675	833	813	887	1135	1095
% can I	0.53	0.54	0.52	0.55	0.62	0.55	0.65	0.68	0.65	0.78	0.76

Figure 8. *Can I* vs. *may I*

The second prescriptive rule shows the shift from *different from* to *different than* from the 1870s to the current time (*Bill is quite different from/than the others*), and is based on 9,636 tokens (see Table 12). As Figure 9 shows, the increase in *different than* is perhaps more noticeable in the following chart, where we see that although there was still some tentativeness in the 1940s–1950s, the increase in *different than* has been quite pronounced since that time, and *different than* is about four times as common now as it was 60 years ago.

Table 12. *Different than* vs. *different from*

different +	1870	1880	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000
from	537	535	513	627	683	663	631	641	668	686	664	692	796	747
than	0	2	2	6	10	13	20	37	20	40	51	69	133	150
% than	0.00	0.00	0.00	0.01	0.01	0.02	0.03	0.06	0.03	0.06	0.08	0.10	0.17	0.20

Turning to descriptive grammar, Figure 10 and Figure 11 show the increase in the *need to V* (*we need to leave*) and the *end up V-ing* (*we'll end up getting there late*) constructions. Notice the nice S-curve increase in both constructions in the last 40–50 years. In terms of extracting the data, it is just a matter of inputting the correct search string (*[need] to [v*]* and *[end] up [v?g*]*) and COHA will find all of the tokens (1827 tokens for *end up V-ing* and 37,503 tokens for *need to V*) and create the chart in less than two seconds.

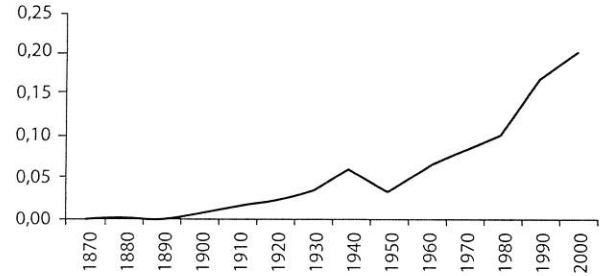

Figure 9. *Different than* vs. *different from*

SECTION	1810	1820	1830	1840	1850	1860	1870	1880	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000
FREQ	4	94	177	283	487	530	770	797	824	1089	1160	1176	1404	1728	1940	2277	2734	3948	6966	9115
PER MIL	3.39	13.57	12.85	17.63	29.57	31.08	41.48	39.23	40.00	49.28	51.10	45.84	57.07	70.97	79.04	94.97	114.80	155.95	249.31	308.28
SEE ALL YEARS AT ONCE																				

Figure 10. *Need to* [VERB]

SECTION	1810	1820	1830	1840	1850	1860	1870	1880	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000
FREQ	0	0	0	0	0	0	0	0	0	0	0	0	2	13	39	90	156	236	442	562
PER MIL	0.00	0.00	0.00	0.00	29.57	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.08	0.53	1.59	3.75	6.55	9.32	15.82	19.01
SEE ALL YEARS AT ONCE																				

Figure 11. *End up* [V-ing]

Table 13. [modal] *always|never* [VERB] (B) vs. *always|never* [modal] [VERB] (A)

	1860	1870	1880	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000
A	490	523	437	389	435	423	405	281	280	241	157	147	122	135	82
B	2301	2547	2772	2608	2864	3128	3180	3051	2922	3143	2815	2755	3137	3665	3876
% B	0.82	0.83	0.86	0.87	0.87	0.88	0.89	0.92	0.91	0.93	0.95	0.95	0.96	0.96	0.98

Even more complicated studies of diachronic syntax can be carried out quite easily with COHA. For example, Table 13 and Figure 12 consider adverb placement with modals. [A] represents pre-modal placement (*never|always* [vm*] [v*] : *he* never would *answer his mail*) while [B] is post-modal placement: ([vm*] *never|always* [v*]: *he* would never *answer his mail*). In this case we simply submit the two competing strings (for a total of 49,311 tokens), copy the data from the two charts into Excel, and create a ratio of B/(A+B). In less than one minute total, we can clearly see the shift towards post-verbal placement: *he would* never *answer his mail.*

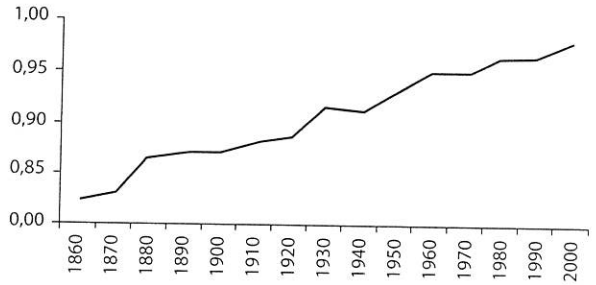

Figure 12. [modal] *always|never* [VERB] vs. *always|never* [modal] [VERB]

To conclude, 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 increase in null relative pronouns at the expense of overt relative pronouns. [A] in Table 14 represents overt relative pronouns with *he* as relative clause subject ([nn*] that|which|who|whom he [vv*]: *the woman* that *he married*) while [B] is the zero relative pronoun: ([nn*] – he [vv*]: *the woman – he married*). As before, we simply copy the data from the two charts and do a simple ratio in Excel. Of course we might want to change the relative clause subject, experiment with different type of antecedents, and so on. But the point is that with COHA, we can do even relatively complex searches such as this – resulting in clear and unambiguous data like that in Table 14 and Figure 13 – in just a minute or so.

Table 14. zero vs. explicit relative pronoun

	1840	1850	1860	1870	1880	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990
A	1835	1668	1683	1758	2052	1911	2067	1995	2039	1740	1463	1516	1392	1291	1124	910
B	4871	4939	5155	6139	7841	8586	8972	9693	10983	9964	9098	9106	9089	8273	8697	7739
% B	0.73	0.75	0.75	0.78	0.79	0.82	0.81	0.83	0.84	0.85	0.86	0.86	0.87	0.87	0.89	0.89

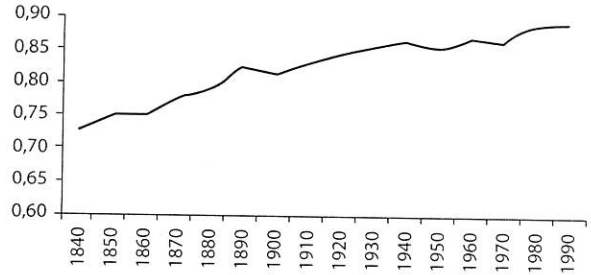

Figure 13. zero (B) vs. explicit (A) relative pronoun

Finally, note that all of the examples above deal with changes in the complete corpus – all genres. As we know, 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 15 and Figure 14 show the frequency per million words for the *end up* constructions (+ADJ: *he ended up dead*, and also +V-ing: *he ended up buying the tickets*). We run the query four times, selecting each of the different genres. We then copy the data into Excel (as in Table 15) and we can then see (as in Figure 14) how in every decade since the early 1900s, the construction has been most common in the more informal genres.

Table 15. *End up* [V-ing] by genre

GENRE	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990
fiction	0.63	0.66	1.09	1.75	4.02	5.34	9.38	12.18	13.27	20.36
magazine	0	0.13	0.55	1.38	1.93	3.38	5.34	7.35	10.19	15.46
newspaper	0.05	0	0.12	0.08	0.21	0.86	1.33	2.86	6.04	8.66
non-fiction book	0.09	0.04	0.12	0.28	0.53	0.73	0.71	1.51	2.37	3.61

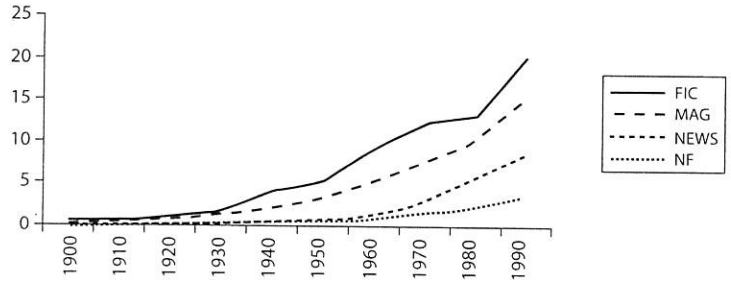

Figure 14. *End up* [V-ing] by genre

6. Semantic change

How can we use corpora to see whether words have changed meaning over time? One option would be to simply look up all tokens (or a randomized subset of tokens) and investigate the use of the word. For example, tokens of *gay* in the 1880s might look like 1–2 in Table 16, while those from the 1980s might look like 3–4. As we laboriously examine hundreds or thousands of tokens – one by one – we can begin to see changes in meaning.

Table 16. Keyword in Context entries for *gay*

	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),

With the right corpus architecture, however, we can both simplify this an make it much quicker. A central concept in corpus linguistics is the idea that "yo can tell a lot about a word by the other words that it hangs out with". If we find tha the collocates of a word are changing over time, this may indicate semantic chang For example, in the examples above, we see that the collocates of *gay* in the 188(are *brilliant*, *attractive*, *jolly*, and *joking*, while in the 1980s they are *heterosexua sexes*, *groups*, and *bisexual*. The goal, then, is to have a corpus architecture that ca quickly find and summarize the data from collocates, to help look for semanti change.

Fortunately, the corpus architecture for COHA allows us to quickly and easil see and compare the collocates of a word or phrase in different periods. For example Table 17 shows us the most frequent nouns occurring immediately after *fast* in eac of the decades. We see that *fast friend* (= 'firm, solid') has decreased, *fast horses* (= 'th horses run fast') is found in nearly all periods, and *fast food*, *fast track*, and *fast lan* (where it has a more figurative meaning) are more recent.[6]

As with lexis, morphology, and phraseology, we can also compare the collocate in different periods. For example, Table 18 shows the nouns occurring after *fast* in th 1830s–1890s and the 1960s–2000s, and we find that the semantic shifts hinted at abov are even more apparent here.

With COHA, we are not limited to examining just immediately adjacent word (such as *fast* + noun), but rather we can look at the entire 'cloud of words' – up to 1(words to the left and to the right of the indicated 'node word'. For example, Table 1!

6. Note that the formalized frequency per million words is shown here. Users can choose t see raw frequency, normalized figures, or a combination of these.

Table 17. Noun collocates of *fast* (*fast* + N), all strings by decade

		TOTAL	1810	1820	1830	1840	1850	1860	1870	1880	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000
2	FAST FOOD	172																	0.55	0.95	2.08	2.60
3	FAST BALL	106					0.06						0.40	0.04	0.24	0.53	1.79	0.50	0.25	0.40	0.04	0.10
4	FAST TRACK	99															0.04	0.13	0.25	1.15	0.97	1.12
5	FAST PACE	93				0.06	0.06	0.06			0.05	0.27	0.13	0.27	0.33	0.12	0.53	0.42	0.63	0.36	0.25	0.27
6	FAST HORSES	79			0.07		0.36	0.29	0.86	0.49	0.49	0.18	0.18	0.23	0.33	0.04	0.04	0.08	0.04	0.04	0.04	0.07
8	FAST FRIEND	64		0.14	0.58	0.69	0.43	0.41	0.32	0.59		0.09	0.13	0.04	0.04	0.08	0.04	0.04			0.04	
10	FAST CARS	53									0.05				0.04	0.16	0.08	0.25	0.50	0.12	0.39	0.44
12	FAST LANE	47																0.04	0.04	0.51	0.50	0.61
13	FAST HOLD	46		0.43	0.36	0.19	0.30	0.35	0.27	0.59	0.15		0.04	0.08	0.04							

Table 18. Noun collocates of *fast* (*fast* + N), comparison

	1830s–1890s	1	2	PM2	PM1	RATIO
1	FAST HOLD	39	0	0.32	0.00	31.8
2	FAST FRIEND	51	2	0.42	0.02	27.1
3	FAST MEN	21	1	0.17	0.01	22.3
4	FAST STEAMER	23	0	0.19	0.00	18.7
5	FAST SAILER	18	0	0.15	0.00	14.7
6	FAST SET	13	1	0.11	0.01	13.8
7	FAST MAIL	10	1	0.08	0.01	10.6
8	FAST MAN	9	1	0.07	0.01	9.6
9	FAST COLORS	11	0	0.09	0.00	9.0
10	FAST STEAMERS	11	0	0.09	0.00	9.0
11	FAST TRAINS	7	1	0.06	0.01	7.4
12	FAST LIVING	7	1	0.06	0.01	7.4

	1960s–2000s	2	1	PM2	PM1	RATIO
1	FAST FOOD	172	0	1.32	0.00	131.7
2	FAST TRACK	98	0	0.75	0.00	75.0
3	FAST CARS	45	1	0.34	0.01	42.3
4	FAST LANE	47	0	0.36	0.00	36.0
5	FAST COMPANY	33	1	0.25	0.01	31.0
6	FAST BALL	32	1	0.24	0.01	30.1
7	FAST ACTION	26	1	0.20	0.01	24.5
8	FAST BREAK	31	0	0.24	0.00	23.7
9	FAST START	30	0	0.23	0.00	23.0
10	FAST GROWTH	25	0	0.19	0.00	19.1
11	FAST BUCK	24	0	0.18	0.00	18.4
12	FAST FACTS	23	0	0.18	0.00	17.6

shows the most frequent noun and adjective collocates near the noun *care* in the different decades.[7]

Notice that in the 1800s, collocates such as *tender*, *utmost*, *anxiety*, *sorrow*, *toil*, and *watchful* were more common, which suggests that *care* was used primarily in the sense of (personal) concern and attention. In the late 1900s, however, collocates such as *health*, *medical*, *intensive*, *foster*, and *physician* are more common, suggesting that the more common use now relates to 'formal, institutional (medical) care'. As before, a direct comparison of the collocates in the two periods provides perhaps even clearer evidence for the shift in meaning and usage (Table 20).

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 for about 30,000 individual words. By searching for '[= word]', we can see the frequency of each matching synonym in each decade. For example, the simple search [= *beautiful*] results in the data in Table 21.

This allows us to see that in the semantic field of 'beautiful', the words *lovely*, *delightful*, *exquisite*, and *pleasing* have decreased over time, while the words *attractive*, *good-looking*, and *scenic* have increased. Such data can be useful in seeing how different words are 'competing for semantic space'. (Note that 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 '[= beautiful] woman'.)

In addition to the 30,000+ synonym sets, it is also possible for users to create their own 'customized lists' of semantically-related words, and to then use them as part of their queries. For example, users could create a list of 40–50 words relating to the body (*hair*, *leg*, *shoulder*, *finger*, *mouth*, *ear*, *foot*, *knee*, *neck*, *lip*, etc.) and then input this list via the web interface. They could then find all cases where one of these words is 'near' (1–10 words, left and/or right) a synonym of the verb *stroke*. In 2–3 seconds, COHA 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.

Finally, we should point out that the features of COHA that are related to semantic change also allow us to move beyond purely linguistically-oriented searches, to look at changes in American history, culture and society. For example, consider Table 22, which shows the most frequent noun collocates of *problem* in the 1830s–1890s and the 1960s–2000s.

7. Note that the figures relate to normalized frequencies per million words.

Table 19. Noun collocates of *care* (N/ADJ near *care*)

	COLLOCATE	TOTAL	1810	1820	1830	1840	1850	1860	1870	1880	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000
1	HEALTH	2492		0.29	0.36	1.37	0.55	0.82	0.48	0.54	1.02	0.36	0.22	0.27	0.2	0.62	0.37	2.79	7.52	8.85	38.4	27.29
2	MEDICAL	1358			0.07	0.12	0.12	0.06	0.16	0.1	0.1	0.09	0.57	0.66	4.63	4.93	5.05	11.8	7.81	5.25	7.09	5.24
6	TENDER	275	3.39	0.87	0.8	1.37	1.76	2.29	1.45	0.98	1.02	0.91	0.35	0.31	0.08	0.16	0.41	0.46	0.21	0.47	0.32	0.24
7	UTMOST	267	1.69	1.59	0.94	1	1.09	0.88	1.35	1.87	1.31	1.04	0.84	0.27	0.49	0.58	0.33	0.13	0.25	0.12	0.04	0.2
8	INTENSIVE	249												0.04		0.08	0.04	0.38	1.64	1.9	2.83	2.37
9	FOSTER	234	0.85	0.58	0.87	1.06	0.91	0.82	0.48	0.34	0.63	0.23	0.26	0.08	0.04	0.04	0.12	0.25	0.55	0.24	1.54	1.89
10	PRIMARY	194				0.12						0.05	0.04				0.04	0.08	0.17	0.04	5.15	1.29
11	NURSING	182		0.14	0.22	0.25	0.3	0.18	0.11	0.15	0.1	0.05	0.13	0.16	0.12	0.58	0.45	0.83	0.88	0.71	1.29	0.95
12	ANXIETY	179	1.69	0.58	1.09	1.56	1.82	1.35	0.81	1.13	0.83	0.45	0.18	0.04	0.08			0.04	0.21	0.04	0.04	
13	PHYSICIAN	169		0.14	0.15	0.12	0.55	0.29	0.32	0.15	0.19	0.41	0.31	0.35	0.24	0.37	0.24	0.42	0.5	0.16	1.22	1.05
14	SORROW	163	1.69	0.72	1.16	1.68	1.52	1.47	0.75	0.74	0.87	0.23	0.09	0.04		0.08			0.04	0.12	0.04	0.03
15	TOIL	136	2.54	0.14	1.09	1.81	1.09	1.17	0.38	0.39	0.68	0.18	0.13	0.2	0.04		0.08			0.24		
17	WATCHFUL	120	0.85	1.01	0.8	0.75	1.09	0.53	1.02	0.64	0.39	0.36	0.04	0.08	0.08	0.08	0.12	0.04			0.04	0.07

Table 20. ADJ/NOUN collocates near the noun *care/cares*, comparison

	1850s–1910s	1	2	PM1	PM2	RATIO
2	JEALOUS	53	0	0.38	0.00	38.5
3	PRECIOUS	37	1	0.27	0.01	35.1
4	FAITHFUL	36	1	0.26	0.01	34.1
6	KINDNESS	31	1	0.22	0.01	29.4
7	ANXIOUS	59	2	0.43	0.02	28.0
8	TENDEREST	38	0	0.28	0.00	27.6
9	PAINS	29	1	0.21	0.01	27.5
10	SYMPATHY	28	1	0.20	0.01	26.5
11	WEIGHT	28	1	0.20	0.01	26.5
12	SORROWS	35	0	0.25	0.00	25.4
13	WEARY	26	1	0.19	0.01	24.6

	1960s–2000s	2	1	PM2	PM1	RATIO
1	PRIMARY	189	1	1.45	0.01	199.4
2	INTENSIVE	245	0	1.88	0.00	187.6
3	FOSTER	123	1	0.94	0.01	129.8
5	PRENATAL	90	1	0.69	0.01	95.0
6	CENTER	117	0	0.90	0.00	89.6
7	MANAGED	115	0	0.88	0.00	88.0
8	COSTS	113	0	0.87	0.00	86.5
9	UNIT	111	0	0.85	0.00	85.0
10	ACCESS	93	0	0.71	0.00	71.2
12	SERVICES	125	2	0.96	0.01	65.9
13	PROVIDERS	85	0	0.65	0.00	65.1

Table 21. Frequency of synonyms of *beautiful* by decade

	SYNONYM	TOTAL	1810	1820	1830	1840	1850	1860	1870	1880	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000
3	LOVELY	22211	116	61.6	71.9	75.5	60.3	53.8	81.8	69.0	65.1	54.9	60.0	58.0	57.6	52.2	53.3	41.5	45.1	42.2	36.8	35.7
6	ATTRACTIVE	11457	16.9	16.3	20.8	25.1	22.7	21.7	27.3	28.2	30.2	27.3	28.1	28.2	25.8	25.9	28.4	36.5	32.7	34.5	30.1	30.3
7	CHARMING	11382	38.1	19.6	20.6	26.2	27.9	39.8	49.0	47.1	43.3	42.6	35.1	34.0	27.4	20.8	21.3	17.7	19.2	17.5	14.8	18.7
9	DELIGHTFUL	8945	45.7	41.3	47.1	35.7	32.2	30.8	43.5	41.4	39.8	36.5	28.9	24.0	14.3	10.4	10.1	8.6	8.4	7.0	5.8	6.2
10	EXQUISITE	6574	20.3	23.2	24.7	24.7	28.7	23.1	28.6	27.1	25.1	26.1	20.5	18.3	9.9	8.6	7.9		8.2	8.7	7.5	8.1
11	PICTURESQUE	5456	12.7	12.9	22.2	30.9	22.5	22.5	26.5	26.8	23.1	22.8	18.4	15.9	9.8	7.2	4.4	3.0	2.9	3.4	3.2	3.8
12	PLEASING	5256	49.1	29.5	31.8	34.8	26.2	21.7	16.8	18.5	16.0	14.4	11.3	13.0	8.3	7.2	7.0	5.4	7.1	5.1	4.8	5.3
15	GOOD-LOOKING	2210		0.4	1.8	3.9	2.5	3.9	3.1	5.0	4.8	4.8	5.8	7.3	7.4	6.8	5.8	5.1	6.9	5.5	6.3	8.0
16	STUNNING	1971	0.9	2.7	2.0	2.6	2.7	2.5	2.4	1.8	1.6	2.7	2.3	3.2	3.2	2.3	3.6	4.4	6.4	10.6	11.1	14.6
17	SCENIC	993	0.9	1.0	0.8	1.0	0.7	0.7	0.9	1.2	2.4	1.8	2.6	2.3	1.8	1.6	2.0	2.6	5.5	2.5	5.8	4.6

Table 22. Noun collocates of *problem* (NOUN near *problem*), comparison

	1830s–1890s	1	2	PM2	PM1	RATIO
1	DESTINY	16	1	0.13	0.01	17.0
2	RAILWAY	23	2	0.19	0.02	12.2
3	SELF-GOVERNMENT	8	1	0.07	0.01	8.5
5	PAUPERISM	8	0	0.07	0.00	6.5
6	MYSTERIES	6	1	0.05	0.01	6.4
7	IMMORTALITY	5	1	0.04	0.01	5.3
9	STATESMEN	9	2	0.07	0.02	4.8
10	MORALS	5	0	0.04	0.00	4.1
11	ELEMENT	18	5	0.15	0.04	3.8
12	UNIVERSE	12	4	0.10	0.03	3.2
14	ORIGIN	21	7	0.17	0.05	3.2
15	INFLUENCE	5	2	0.04	0.02	2.7
17	SIN	7	3	0.06	0.02	2.5

	1960s–2000s	2	1	PM2	PM1	RATIO
1	HEALTH	486	1	3.72	0.01	457.0
3	DRUG	222	0	1.70	0.00	170.0
5	SECURITY	125	1	0.96	0.01	117.6
6	AREAS	121	1	0.93	0.01	113.8
8	POLLUTION	140	0	1.07	0.00	107.2
9	MONEY	226	2	1.73	0.02	106.3
11	RESPONSE	93	1	0.71	0.01	87.5
12	POLICY	86	1	0.66	0.01	80.9
13	TRAFFIC	85	1	0.65	0.01	79.9
14	ENERGY	104	0	0.80	0.00	79.6
15	BEHAVIOR	102	0	0.78	0.00	78.1
16	DRINKING	94	0	0.72	0.00	72.0
17	UNEMPLOYMENT	91	0	0.70	0.00	69.7

Note the emphasis in the 1800s on philosophical concepts like *destiny*, *mysteries*, *immortality*, *morals*, *influence*, and *sin*, whereas in the late 1900s the collocates of *problem* relate to more contemporary concerns like *health (care)*, *drugs*, *(national) security*, *pollution*, *traffic*, and *energy*.[8] Together with the comparisons of lexis (verbs: *access*, *broadcast*, *program*, or *download* in the late 1900s) and even morphology (*-ist* nouns: *psychiatrist*, *activist*, and *therapist* in the late 1900s) seen above, the ability to compare collocates across time provides insight not only into semantic change, but also cultural and societal changes in the United State during the past 200 years.

7. Conclusion: Size and architecture

The Corpus of Historical American English allows researchers to study many different types of changes in English for the last 200 years, in ways that are not possible with other corpus. This is due in large part to corpus size, corpus granularity, and the architecture of the corpus.

7.1 Corpus size

First, the importance of size cannot be ignored. COHA contains 400 million words from the 1810s–2000s. Other than COHA, there are very few historical corpora of English for this period, and the other corpora are quite small in comparison. The Brown family of corpora (Brown, LOB, FROWN, and FLOB) contains four million words of text from the 1960s–1990s (two million for the US) (see Mair 1997), and work is proceeding on small one million word extensions backwards in time as well. The ARCHER corpus (see Biber 1994) – even with recent expansions – will have less than three million words (with less than 1.5 million words for the US). In other words, COHA is about 200 times as large as the American components of either BROWN+ or ARCHER.

And size matters – a great deal. Imagine that we take any of the 'comparison' tables shown above (comparing lexis, morphology, phraseology, or semantics) and divide

8. Obviously, the frequency of a word as collocate is related to the overall frequency of the word itself in the corpus. For example, *pauperism* is much more frequent as a collocate of *problem* in the 1800s, simply because the word *pauperism* is more frequent overall in the 1800s. 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 Tables 3, 5, 9, 10, 18, 20, and 22 fairly complicated, and there is a question about how much more complexity we want to add.

the number of tokens by 200 – in other words, the size of BROWN+ or ARCHER. Now the 60–100 tokens for a given word form, collocate, or lexical item becomes 2 or 3 tokens with the small corpora – far too small to say anything meaningful. It is no surprise then that the vast majority of studies that have been done on recent changes in English have focused on syntax. Often there are simply not enough tokens with smaller 2–3 million word corpora to carry out insightful studies of lexis, morphology, and semantics.

Even in the area of syntax, size matters a great deal. Consider Table 23, which shows the frequency of [to-V] and [V-ing] complements of the verb *hate*: *I hate to write papers/I hate writing papers*. (This is part of a much larger data set, which shows a general shift from [to-V] to [V-ing] complements with a number of verbs; see Rohdenburg 2006.)

Table 23. [V-ing] vs. [to-V] with *hate*

	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

With a small two million word corpus (one two-hundredths the size of COHA), rather than 300–400 tokens per decade, we would have only 2 or 3 tokens. At this point, it would be very difficult to show statistical significance in terms of changes – the numbers are just too small. As a result, it is not surprising that the vast majority of studies on recent changes in English syntax have focused on just the highest-frequency constructions – modals, auxiliaries, relative pronouns, and the like. These are the only constructions that have enough tokens to carry out insightful analyses.

Our view, however, is that we should not be artificially forced into looking at just a small subset of all linguistic changes, simply because those are the only ones that *can* be studied with small corpora. A truly useful corpus will allow us to look at all types of changes – lexical, morphological, semantic, and syntactic (and high and low frequency syntactic constructions as well).

7.2 Corpus granularity

With some other historical corpora of English, the corpus is purposely limited to data from every thirty years – e.g. the 1930s, 1960s, and 1990s. COHA, however, contains texts from a continuous range of years – every decade from the last 200 years (and in

most cases, every year in each of those decades). As a result, COHA allows us to see interesting changes that other corpora might miss.

For example, consider Table 24 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*), based on nearly 40,000 tokens with *start* and nearly 100,000 tokens with *begin*. We see that in one single decade – the 1920s – the percentage of [V-ing] with *start* nearly doubled (23% to 41%). 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.

Table 24. [V-ing] vs. [to-V] with *start* and *begin*

construction	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000
start to V	514	792	1493	1576	1845	1983	1784	1853	2256	2984	3186
start V-ing	34	159	439	1110	1499	1780	1926	2150	2363	3792	4340
% start V-ing	0.06	0.17	0.23	0.41	0.45	0.47	0.52	0.54	0.51	0.56	0.58
begin to V	6587	6644	6856	7399	7612	6995	7527	6797	7657	7198	6244
begin V-ing	569	802	1180	1570	1702	1688	1999	2118	2558	2742	3005
% begin V-ing	0.08	0.11	0.15	0.18	0.18	0.19	0.21	0.24	0.25	0.28	0.32

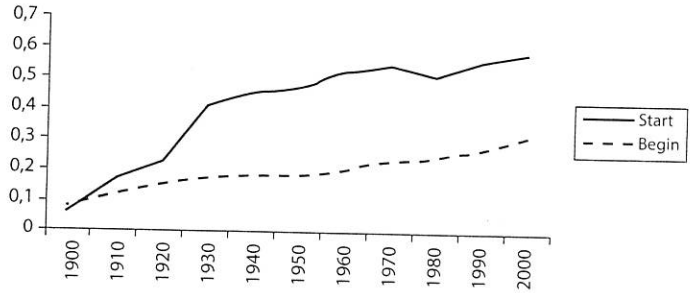

Figure 15. [V-ing] vs [to-V] as complements of *start* and *begin*

Having good granularity is also important in terms of looking at related shifts. For example, as we have discussed, the largest increase in [V-ing] with *start* occurred in the 1920s, whereas with the emotion verbs *love, hate*, and *like* it occurred somewhat later (1950s–2000s; see Table 23). 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, grammaticalization, specific functional and stylistic motivations, and so on.

7.3 Corpus architecture

'Unstructured corpora' and text archives like Google Books, Google News Archive, and other collections of historical newspapers and magazines are much larger than the 400 million word Corpus of Historical American English. So why not use these larger resources instead of COHA? The answer lies with corpus architecture. With the unstructured corpora and text archives, it would be difficult or even impossible to study the wide range of language changes that can be studied quickly and easily with COHA.

In terms of lexical change, with the larger unstructured corpora and text archives, one can probably find the 'first occurrence' of a word or phrase with more precision than with COHA. However, it may not be possible to (accurately) measure frequency over time. Most interfaces do not allow users to see frequency by decade or year. Rather, one would have to carry out the search for the word or phrase for each individual decade, and then somehow 'normalize' the data (per million words, in each decade). As far as comparing all words and phrases in the corpus in two different time periods (as in Tables 3, 5, 9, 10, 18, 20, and 22), this would not be possible with unstructured corpora and text archives – it is only possible with COHA.

With unstructured corpora and text archives, it is also difficult or impossible to carry out studies on morphological change, since these resources do not allow users to search by wildcard, as in our *-ist* searches above. It is also difficult or impossible to carry out syntactic research, because the unstructured corpora and text archives are not lemmatized or tagged for part of speech. For example, if we are interested in the 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 overgenerate results. With COHA, we can carry out this search (*[vv*] [p*] into [v?g*]*) ; to find all 1669 tokens with an embedded clause subject that is a pronoun) in less than two seconds.

Finally, COHA allows us to look at semantic change much more easily than we could with unstructured corpora and text archives. This is a result of the fact that COHA allows us to extract collocates and to compare them in different historical periods. With unstructured corpora and text archives, we would have to write a program to input the node word into the search interface, retrieve the hits, find and copy the 4–5 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 2–3 seconds.

In summary, the 400 million word Corpus of Historical American English allows us to carry out a wide range of studies on changes in American English (1810s–2000s) in ways that are probably not possible with any other corpus.

References

Biber, Douglas, Edward Finegan & Dwight Atkinson. 1994. "ARCHER and its Challenge: Compiling and Exploring a Representative Corpus of Historical English Registers". *Creat ing and Using English Language Corpora: Papers from the Fourteenth International Confer ence on English Language Research on Computerized Corpora, Zürich 1993* ed. by Udo Fries Gunnel Tottie & Peter Schneider, 1–13. Amsterdam: Rodopi.

Davies, Mark. 2010. "The Corpus of Contemporary American English as the First Reliable Mon itor Corpus of English". *Literary and Linguistic Computing* 25: 4.447–464.

Davies, Mark. 2009. "The 385+ Million Word Corpus of Contemporary American Englisl (1990–2008+): Design, Architecture, and Linguistic Insights". *International Journal o Corpus Linguistics*. 14.159–190.

Hunston, Susan & Gill Francis. 2000. *Pattern Grammar: A Corpus-driven Approach to the Lexica Grammar of English*. (= *Studies in Corpus Linguistics*, 4.) Amsterdam: John Benjamins.

Rohdenburg, Günter. 2006. "The Role of Functional Constraints in the Evolution of the Englisl Complementation System". *Syntax, Style and Grammatical Norms: English from 1500–2000* (= *Linguistic Insights – Studies in Language and Communication*, 39.) ed. by Christian Dalton-Puffer, Dieter Kastovsky, Nikolaus Ritt & Herbert Schendl, 143–166. Bern Peter Lang.