This was an invited chapter for a book dealing with the history of English, which was published by a major university press in 2015. After two positive reviews, it was removed at the last minute because of objections from a third reviewer. This person apparently had a vested interest in preventing others from seeing **Section 2**, which s/he felt was too explicit in showing the limitations of small corpora for lexical studies.

I believe that this paper explains some very important concepts about corpora and lexical change, which are often misunderstood (or willfully ignored and even suppressed) by other researchers. As a result, I've decided to make the paper freely available via the Web – where it will likely have wider circulation than it would have had, sitting on the shelf in a library.

Using corpora to research lexical changes in Late (and Early) Modern English

Mark Davies
Brigham Young University

The focus of this chapter is the different ways in which corpora English can be used to research lexical changes in Late (and Early) Modern English. Although we could examine any type of language change (syntactic, morphological, pragmatic, and so on), in this case we have focused on lexical change, because it is an often-overlooked area of change in English – especially by those who are used to working with small corpora, where there are typically not enough tokens to look at medium and low-frequency words. In addition, we will focus primarily on Late Modern English – especially the period from the early 1800s through the mid-1900s – because until recently there were very few large, structured corpora for this period. As we will see in this chapter, however, there are a number of corpora and corpus-related resources that have become available in just the past four or five years, which have greatly expanded our ability to carry out extremely detailed investigations on lexical changes in English.

We will begin this chapter by looking at the type of research that one can do with the 400 million word Corpus of Historical American English (COHA). We will then consider what type of lexical research we can (or cannot) do with smaller corpora, a handful of large text archives, the quotations in the online Oxford English Dictionary (OED), and Google Books.

As we consider all of these resources, we will focus on how the corpora and related resources can be used to generate the following types of data:

1. Concordances for a particular word or phrase, to see the patterns in which it occurs
2. The frequency of specific words and phrases over time
3. Finding all words that are more frequent in one period than in another
4. Collocates of a particular word or phrase (over time) in order to see semantic change
5. Easy access to the relative frequency of related words, such as synonyms or words in a user-defined list

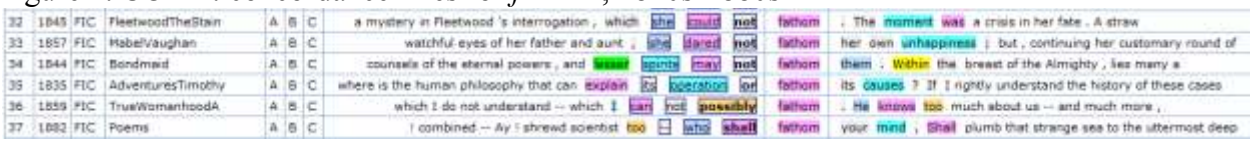
1. The Corpus of Historical American English (COHA)

The Corpus of Historical American English (COHA) was released in 2010 as a "companion" corpus to the 450 million word Corpus of Contemporary American English (COCA), which contains texts from just 1990 to the current time. COHA, on the other hand, contains 400 million words from 1810-2009, which makes it at least 100 times as large as any other publicly-available, structured corpus of English (as opposed to just text archives). It is composed of texts from fiction, popular magazines, newspapers, and non-fiction books, and it is genre-balanced, in the sense that it has roughly the same balance of these genres from decade to decade. (For an overview of COHA, see Davies 2012a and 2012b).

Most importantly – for the purpose of this paper – it allows for an extremely wide range of research on lexical changes in Late Modern English, which in many cases are not available from other corpora. Due to limited space, we will provide just one sample of each type of search, but extensive tutorials, with many additional sample searches, can be found at the corpus website.

At the most basic level, we can use COHA to see concordances for a particular word or phrase, to see the patterns in which it occurs. Users simply enter a word, phrase, or construction, and then (optionally) select a range of years, to see up to 1000 entries at a time. For example, at the current time, *fathom* as a verb is used almost exclusively in negative and interrogative contexts, and COHA shows that the same holds true when we examine the concordance lines from the 1820s-1880s, e.g.:

Figure 1. COHA: concordance lines for *fathom*, 1820s-1880s



Concordance lines can also provide evidence for semantic change over time. For example, the following two figures are for the word *gay* in the 1860s-1910s and the 1970s-2000s:

Figure 2. COHA: concordance lines for *gay*, 1860s-1910s

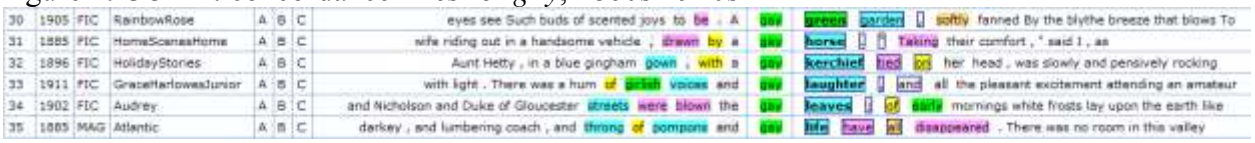
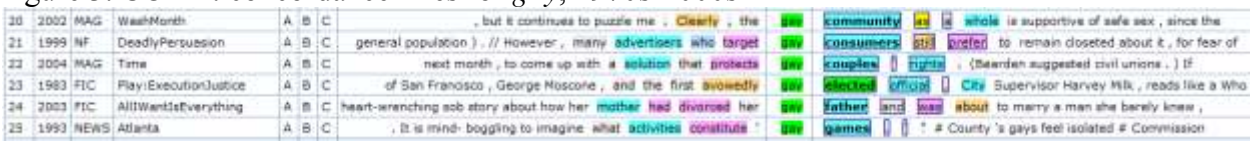


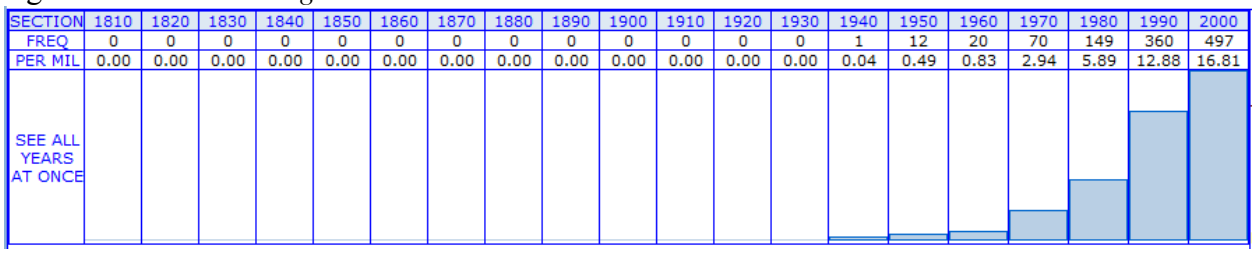
Figure 3. COHA: concordance lines for *gay*, 1970s-2000s



Second, COHA allows users to quickly and easily see the frequency of specific words and phrases over time. Figures 1-3 below are just a handful of examples, showing the frequency of

teenager (increasing over time), *steamship* (highest frequency in the early 1990s, and related to changes in society and culture), and *many a time* (decreasing over time).

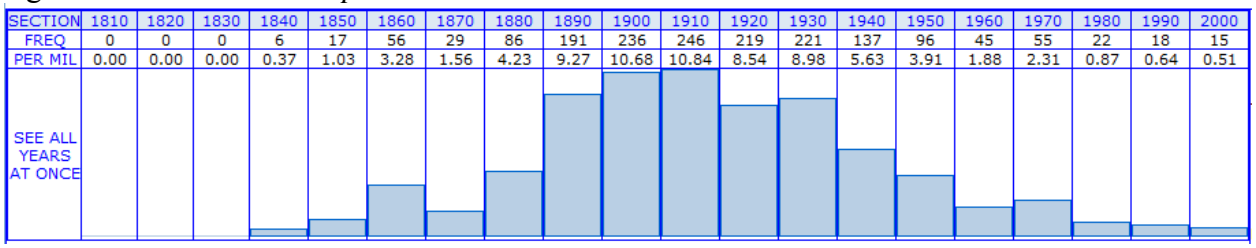
Figure 4. COHA: *teenager*



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	0	1	12	20	70	149	360	497
PER MIL	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.49	0.83	2.94	5.89	12.88	16.81

SEE ALL YEARS AT ONCE

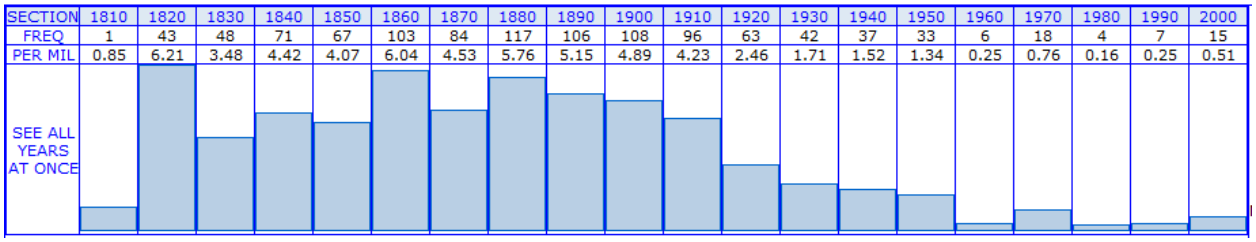
Figure 5. COHA: *steamship*



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	6	17	56	29	86	191	236	246	219	221	137	96	45	55	22	18	15
PER MIL	0.00	0.00	0.00	0.37	1.03	3.28	1.56	4.23	9.27	10.68	10.84	8.54	8.98	5.63	3.91	1.88	2.31	0.87	0.64	0.51

SEE ALL YEARS AT ONCE

Figure 6. COHA: *many a time*

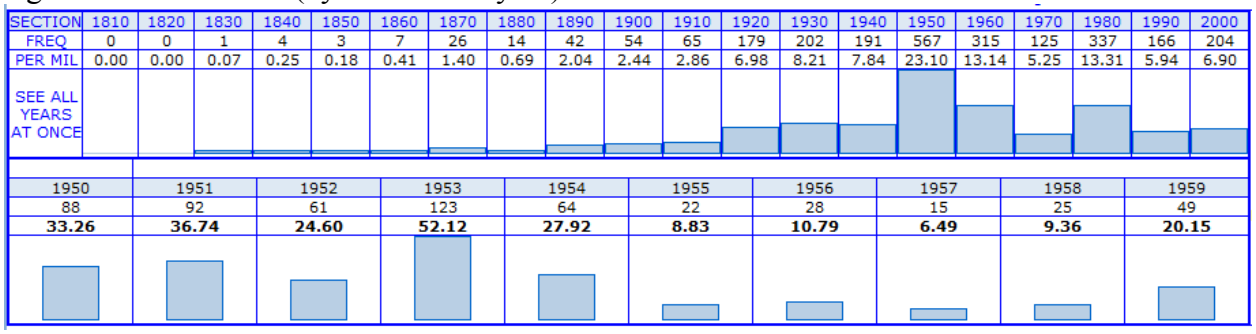


SECTION	1810	1820	1830	1840	1850	1860	1870	1880	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000
FREQ	1	43	48	71	67	103	84	117	106	108	96	63	42	37	33	6	18	4	7	15
PER MIL	0.85	6.21	3.48	4.42	4.07	6.04	4.53	5.76	5.15	4.89	4.23	2.46	1.71	1.52	1.34	0.25	0.76	0.16	0.25	0.51

SEE ALL YEARS AT ONCE

It is also possible to see the frequency by individual year, as in the following chart for *reds*, which peaks in 1953 (the height of the McCarthy anti-Communist hearings in the US Senate):

Figure 7. COHA: *reds* (by decade and year)



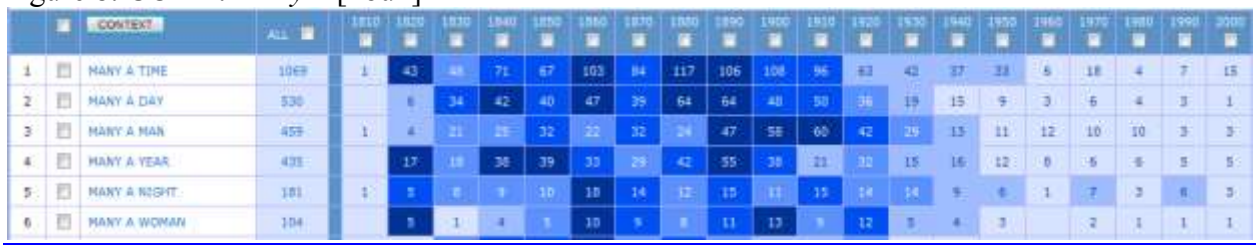
SECTION	1810	1820	1830	1840	1850	1860	1870	1880	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000
FREQ	0	0	1	4	3	7	26	14	42	54	65	179	202	191	567	315	125	337	166	204
PER MIL	0.00	0.00	0.07	0.25	0.18	0.41	1.40	0.69	2.04	2.44	2.86	6.98	8.21	7.84	23.10	13.14	5.25	13.31	5.94	6.90

SEE ALL YEARS AT ONCE

1950	1951	1952	1953	1954	1955	1956	1957	1958	1959
88	92	61	123	64	22	28	15	25	49
33.26	**36.74**	**24.60**	**52.12**	**27.92**	**8.83**	**10.79**	**6.49**	**9.36**	**20.15**

Users can also see the frequency (by decade) of each matching string for a particular construction. For example, the following table shows the first few entries for the construction *many a* [noun] (for a in-depth discussion of *many a* NOUN in COHA, see Hilpert 2012):

Figure 8: COHA: *many a* [noun]



	CONTEXT	ALL	1810	1820	1830	1840	1850	1860	1870	1880	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000
1	MANY A TIME	1069	1	43	48	71	67	103	84	117	106	108	96	63	42	37	33	6	18	4	7	15
2	MANY A DAY	530		6	34	42	40	47	39	64	64	48	50	36	19	15	9	3	6	4	3	1
3	MANY A MAN	459	1	4	21	25	32	22	32	24	47	58	60	42	29	13	11	12	10	10	3	3
4	MANY A YEAR	435		17	18	38	39	33	29	42	55	38	21	32	15	16	12	6	6	6	5	5
5	MANY A NIGHT	181	1	5	6	9	10	18	14	12	15	11	15	14	14	9	6	1	7	3	6	3
6	MANY A WOMAN	104		9	1	4	5	10	9	8	11	13	9	12	5	4	3		2	1	1	1

Users can then click on the bar for any decade or year in the frequency chart display (Figures 4-7) or words and phrases in the table display (Figure 8) to see the KWIC entries (Keyword in Context); for example the entries for *many a time* in the 1850s:

Figure 9. COHA: Keyword in Context lines (KWIC)

26	1853	FIC	UncleSamsEmancipation	A	B	C	Druids -- graybeards, dusky garments and all, on the shores of Maine, **many a time**: and if anybody wants to feel the beauty and grandeur of the
27	1854	FIC	TempestSunshine	A	B	C	, and which Julia, in the bitterness of her heart, cursed many and **many a time**. In the early part of the morning Dr. Lacey wandered down to
28	1854	FIC	TempestSunshine	A	B	C	had resolved that if there still was hope for him he would find it. **Many a time** during the succeeding days he prayed in secret, not that Fanny might
29	1854	FIC	TempestSunshine	A	B	C	the bookcase the old family Bible, on whose dark dusty covers she remembered having **many a time** written her name. All was now explained. Her father's
30	1854	FIC	RhymesWithReason	A	B	C	storm. " Behold upon the kitchen wall That old and rusty gun! Full **many a time** the same I've borne Till setting of the sun; On training-days
31	1854	FIC	ThisThatOther	A	B	C	touches to a beautiful Madonna. Wearily had the girl-artist toiled and studied, and **many a time** had her lamp grown dim, in the gray light of morning,
32	1854	FIC	MartinMerrivale	A	B	C	's disconsolate soul. His step-mother was remembered with a sort of tenderness, and **many a time** he wished himself sitting in her kitchen corner, as of old
33	1855	FIC	ScenesCharacters	A	B	C	Reginald's. It was not very often that quarrels went so far, but **many a time** in thought, word, and deed was the rule of love transgressed

Third, with one simple search, in COHA it is possible to see all words that are more frequent in one period than in another. For example, the following table compares *-ism* words in the 1860s-1910s (left) with the 1970s-2000s (right), and it is quite interesting to see how these relate to cultural changes in the United State. Note that users could also compare any set of words – for example, all verbs or all nouns that are more common in one period than another.[1]

Figure 10. COHA: *–ism* words, 1860s-1910s vs 1970s-2000s

SEC 1: 121,332,176 WORDS

	WORD/PHRASE	TOKENS 1	TOKENS 2	PM 1	PM 2	RATIO
1	PAUPERISM	217	1	1.79	0.01	190.72
2	FETICHISM	113	0	0.93	0.00	93.13
3	BIMETALLISM	62	0	0.51	0.00	51.10
4	ROMANISM	62	0	0.51	0.00	51.10
5	HEATHENISM	94	2	0.77	0.02	41.31
6	DEMAGOGISM	41	1	0.34	0.01	36.04
7	PROPAGANDISM	43	0	0.35	0.00	35.44
8	ECCLESIASTICISM	42	0	0.35	0.00	34.62
9	MOHAMMEDANISM	117	3	0.96	0.03	34.28
10	MONOPOLISM	40	0	0.33	0.00	32.97
11	SPIRITISM	38	0	0.31	0.00	31.32
12	INVALIDISM	70	2	0.58	0.02	30.76

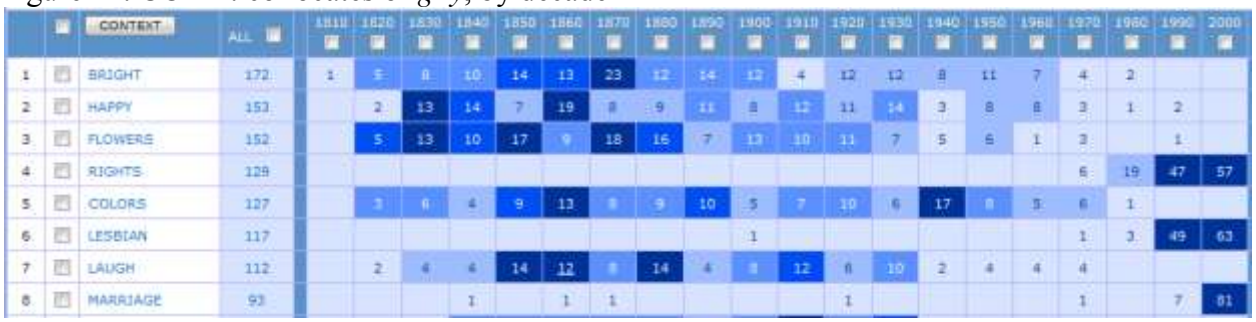
SEC 2: 106,640,094 WORDS

	WORD/PHRASE	TOKENS 2	TOKENS 1	PM 2	PM 1	RATIO
1	RACISM	994	0	9.32	0.00	932.11
2	TOURISM	759	0	7.12	0.00	711.74
3	ACTIVISM	405	1	3.80	0.01	460.80
4	MARXISM	342	2	3.21	0.02	194.56
5	FUNDAMENTALISM	176	0	1.65	0.00	165.04
6	MULTICULTURALISM	156	0	1.46	0.00	146.29
7	COUNTERTERRORISM	134	0	1.26	0.00	125.66
8	DYNAMISM	130	0	1.22	0.00	121.91
9	AUTHORITARIANISM	101	1	0.95	0.01	114.92
10	EXPRESSIONISM	100	1	0.94	0.01	113.78
11	CONSUMERISM	121	0	1.13	0.00	113.47
12	SEXISM	116	0	1.09	0.00	108.78

Fourth, in COHA we can see the collocates of a particular word or phrase (over time) in order to see semantic change. For example, to continue the example shown above, the following chart shows the collocates of *gay* by decade. The older meaning of "happy. cheerful" is seen in lines like 1 and 2, while the newer meaning related to sexual orientation is found in lines like 4 and 6.

[1] To see the actual query online, see http://corpus.byu.edu/coca/?c=coca&q=20252401. These are adjectives that occur at least two times in each of the two time periods. The [ratio] column shows how much more common the words is in the one time period than the other, taking into account the overall size (in words) of each time period.

Figure 11. COHA: collocates of *gay*, by decade



	CONTEXT	ALL	1810	1820	1830	1840	1850	1860	1870	1880	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000
1	BRIGHT	172	1	5	8	10	14	13	23	12	14	12	4	12	12	8	11	7	4	2		
2	HAPPY	153		2	13	14	7	19	8	9	11	8	12	11	14	3	8	8	3	1	2	
3	FLOWERS	152		5	13	10	17	9	18	16	7	13	10	11	7	5	6	1	3		1	
4	RIGHTS	128																	6	19	47	57
5	COLORS	127		3	6	4	9	13	8	9	10	5	7	10	6	17	8	5	6	1		
6	LESBIAN	117										1							1	3	49	63
7	LAUGH	112		2	4	4	14	12	8	14	4	8	12	8	10	2	4	4	4			
8	MARRIAGE	93				1		1	1					1					1		7	81

It is also possible to compare the collocates in different periods. For example, the following table shows the collocates of *gay* in the 1830s-1910s (left) compared to the 1970s-2000s (right):

Figure 12. COHA: ADJ/NOUN collocates near the noun *gay*

SEC 1: 167,626,806 WORDS

	WORD/PHRASE	TOKENS 1	TOKENS 2	PM 1	PM 2	RATIO
1	GRAVE	63	1	0.38	0.01	40.08
2	LADY	66	0	0.39	0.00	39.37
3	GALLANT	65	0	0.39	0.00	38.79
4	BRILLIANT	57	0	0.34	0.00	34.00
5	HEART	51	1	0.30	0.01	32.44
6	GRAVE	51	0	0.30	0.00	30.42
7	THRONG	50	0	0.30	0.00	29.83
8	VOICES	46	1	0.27	0.01	29.26
9	GLAD	48	0	0.29	0.00	29.23
10	SPIRITS	48	0	0.29	0.00	28.64
11	SONG	46	0	0.27	0.00	27.44
12	ATTIRE	42	1	0.25	0.01	26.72

SEC 2: 106,640,094 WORDS

	WORD/PHRASE	TOKENS 2	TOKENS 1	PM 2	PM 1	RATIO
1	LESBIAN	116	1	1.09	0.01	182.34
2	RIGHTS	129	0	1.21	0.00	120.97
3	COMMUNITY	81	2	0.76	0.01	63.66
4	BAR	38	1	0.36	0.01	59.73
5	STRAIGHT	35	1	0.33	0.01	55.02
6	LESBIAN	35	1	0.33	0.01	55.02
7	ACTIVISTS	31	1	0.29	0.01	48.73
8	MARRIAGE	89	3	0.83	0.02	46.63
9	COUPLES	27	1	0.25	0.01	42.44
10	NATIONAL	20	1	0.19	0.01	31.44
11	BISEXUAL	31	0	0.29	0.00	29.07
12	LESBIANS	27	0	0.25	0.00	25.32

Another example of comparing collocates is Figure 13, which shows the adjectival collocates preceding *women* in the 1830s-1890s (left) and the 1960s-2000s (right), and which shows how woman are represented and portrayed in the two periods:

Figure 13. COHA: Adjectival collocates of *women*

SEC 1: 122,828,575 WORDS

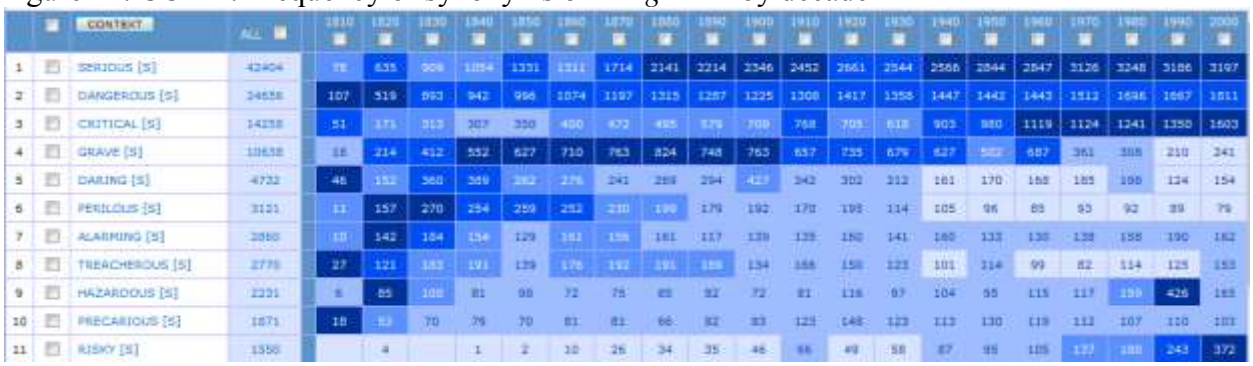
	WORD/PHRASE	TOKENS 1	TOKENS 2	PM 1	PM 2	RATIO
1	STRONG-MINDED	23	1	0.19	0.01	24.46
2	NOBLE	30	2	0.24	0.02	15.95
3	TRUE	13	1	0.11	0.01	13.82
4	AGED	24	2	0.20	0.02	12.76
5	CULTIVATED	12	0	0.10	0.00	9.77
6	DEFENCELESS	11	0	0.09	0.00	8.96
7	ELDER	11	0	0.09	0.00	8.96
8	FINEST	8	1	0.07	0.01	8.51
9	LOVELIEST	8	1	0.07	0.01	8.51
10	LITERARY	10	0	0.08	0.00	8.14
11	HEROIC	10	0	0.08	0.00	8.14
12	DEVOTED	10	0	0.08	0.00	8.14

SEC 2: 130,617,326 WORDS

	WORD/PHRASE	TOKENS 2	TOKENS 1	PM 2	PM 1	RATIO
1	PREGNANT	364	3	2.02	0.02	82.73
2	BATTERED	70	0	0.54	0.00	53.59
3	NATIONAL	70	0	0.54	0.00	53.59
4	AFRICAN-AMERICAN	61	0	0.47	0.00	46.70
5	PROFESSIONAL	47	1	0.36	0.01	44.20
6	JAPANESE	39	1	0.30	0.01	36.67
7	BLACK	493	14	3.77	0.11	33.11
8	NAKED	69	2	0.53	0.02	32.44
9	SOVIET	32	0	0.24	0.00	24.50
10	CATHOLIC	25	1	0.19	0.01	23.51
11	DIVORCED	28	0	0.21	0.00	21.44
12	HOMELESS	22	1	0.17	0.01	20.69

Fifth and finally, in COHA it is possible to see the relative frequency of related words, such as synonyms or words in a user-defined list. For example, the following table shows the frequency of synonyms of *dangerous* by decade, where we see the increase of words like *critical* and *risky*, and the decrease of words like *perilous*, *alarming*, and *treacherous*:

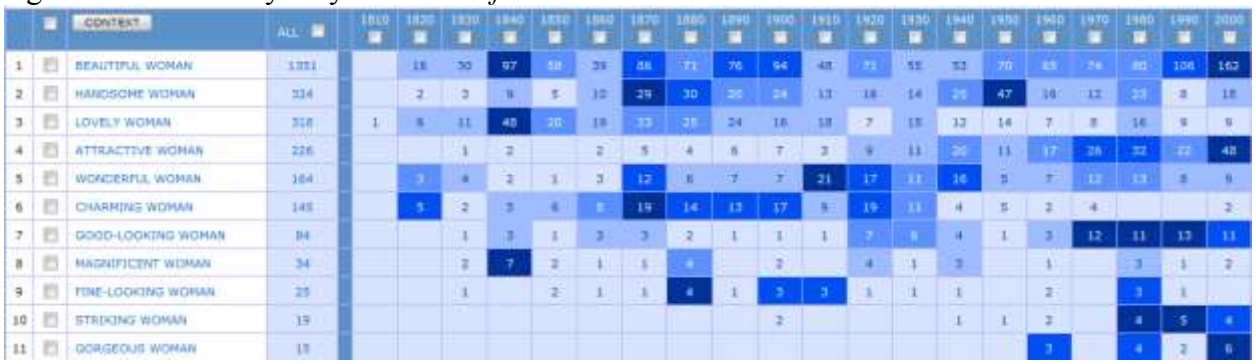
Figure 14. COHA: Frequency of synonyms of *dangerous* by decade



	CONTEXT	ALL	1810	1820	1830	1840	1850	1860	1870	1880	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000
1	SERIOUS [S]	42404	78	635	909	1084	1331	1311	1714	2141	2214	2346	2452	2661	2544	2566	2844	2847	3126	3248	3186	3197
2	DANGEROUS [S]	34658	107	519	893	942	996	1074	1197	1315	1287	1225	1308	1417	1358	1447	1442	1443	1513	1696	1667	1611
3	CRITICAL [S]	14258	51	171	313	307	350	400	472	485	579	709	768	705	618	903	980	1119	1124	1241	1350	1503
4	GRAVE [S]	13638	18	214	412	552	627	710	783	824	748	765	657	735	679	627	582	687	361	308	210	241
5	DARING [S]	4732	46	152	360	389	382	276	241	288	294	417	342	302	212	181	170	168	185	198	124	154
6	PERILOUS [S]	3121	11	157	270	254	259	252	230	190	179	192	170	193	114	105	96	85	93	92	89	79
7	ALARMING [S]	2860	10	142	184	154	129	161	156	161	117	139	139	160	141	160	133	130	138	158	190	162
8	TREACHEROUS [S]	2770	27	121	183	193	139	176	192	191	188	134	168	150	123	101	114	99	82	114	125	153
9	HAZARDOUS [S]	2231	8	85	108	81	98	72	75	85	82	72	81	116	87	104	95	115	117	150	426	165
10	PRECARIOUS [S]	1671	18	53	70	79	70	81	81	66	82	83	125	148	123	113	130	119	112	107	110	103
11	RISKY [S]	1350		4		1	2	10	26	34	35	46	66	49	58	87	95	105	122	188	243	372

Another example is the search *[=beautiful] woman*, which finds any synonym of beautiful followed by *woman* (note the decrease in *lovely* and *charming* and the increase in *attractive*, *striking*, and *gorgeous*):

Figure 15. COHA: Synonyms: "*beautiful*" + *woman*



	CONTEXT	ALL	1810	1820	1830	1840	1850	1860	1870	1880	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000
1	BEAUTIFUL WOMAN	1351		18	30	97	58	39	88	71	76	94	48	71	55	53	70	65	74	80	106	162
2	HANDSOME WOMAN	334		2	3	9	5	10	29	30	20	24	13	18	14	25	47	10	12	23	8	18
3	LOVELY WOMAN	318	1	6	11	48	20	18	33	25	24	16	18	7	15	12	14	7	8	16	9	9
4	ATTRACTIVE WOMAN	226			1	2		2	5	4	6	7	3	9	11	20	11	17	28	32	22	48
5	WONDERFUL WOMAN	184		3	4	2	1	3	12	8	7	7	21	17	11	16	5	7	12	13	8	9
6	CHARMING WOMAN	145		5	2	3	6	8	19	14	13	17	9	19	11	4	5	2	4			2
7	GOOD-LOOKING WOMAN	84			1	3	1	3	3	2	1	1	1	7	8	4	1	3	12	11	13	11
8	MAGNIFICENT WOMAN	34			2	7	2	1	1	4		2		4	1	3		1		3	1	2
9	FINE-LOOKING WOMAN	25			1		2	1	1	4	1	3	3	1	1	1		2		3	1	
10	STRIKING WOMAN	19										2				1	1	2		4	5	4
11	GORGEOUS WOMAN	17																3		4	2	6

We have shown how COHA – which extends from the 1810s-2000s, can be used to examine a wide range of lexical changes in English. As is discussed in Davies (2011), it is possible to use the related Corpus of Contemporary American English (currently 450 million words, 1990-2012) to carry out similar searches for just the last two decades or so. For example, we could see charts (like Figures 1-3 above) showing the frequency by year of words and phrases that have increased in frequency during this time (e.g. *old-school*, *morph (verb)*, *freak out*, *perfect storm*, *think outside the box*, or *throw someone under the bus*.

As we can do for COHA for the period of the 1810s-2000s, we can use COCA to look at just the 1990s-2000s to generate lists of all words that have increased or decreased in frequency during this time. For example, with one quick search we could quickly find all adjectives that have decreased in frequency from the 1990s to the 2000s (e.g. *Croatian*, *Kuwaiti*, *Bosnian*, *US-Soviet*, *Soviet*, *anti-apartheid*) or that have increased in frequency (e.g. *clean-energy*, *anti-doping*, *low-carb*, *anti-terror*, *high def*). We can compare concordance lines for a given word or phrase, where there have been changes over time (such as *web* or *green*). As with COHA, we can compare collocates in two recent time periods. For example, collocates of *web* that were more common in the early 1990s than the late 2000s include *strand*, *spider*, *relation*, *life*, and *image*, where those from the late 2000s include *site*, *page*, *e-mail*, *browser*, and *link*. Finally, we can –

with just one quick search – see the relative frequency of related words (such as synonyms of a given word) over the past two decades.

Having now provided samples of the different types of lexical data that a large, well-designed corpus can provide, let us now turn to several examples of other corpora and lexical resources for Late Modern English, to see what types of insight they give for lexical change.

2. Small corpora

English historical linguistics has a strong tradition of small, well-designed corpora, in the range of one to five million words each. As López-Couso (this volume) mentions, these include the Brown family of corpora – one million words each in Brown (US 1960s), LOB (UK 1960s), Frown (US 1990s), and FLOB (UK 1990s). They also include ARCHER (1.8 million words, 1650-1999) CONCE (Corpus of Nineteenth Century Texts) (1,000,000 words, UK, 1800s), and the Helsinki Corpus (1.6 million words, Old English through the early 1700s). These are all "general" historical corpora – covering a wide range of genres and (like COHA) balanced by genre from decade to decade. There are also many small corpora of particular genres, such as letters, newspapers, or court proceedings.[2] Again, for a very complete and balanced overview of such corpora, please see López-Couso (this volume), especially Section 2 of that chapter.

These small corpora have certainly proven their value in research on high-frequency syntactic constructions, such as modals and other auxiliaries, pronouns, and prepositions, where even in one million words there might be hundreds or even thousands of tokens. But much less has been done – or can be done – in terms of lexical change, where there are just a handful of tokens for most words. A few studies have attempted to use these smaller corpora in looking at changes in lexis, including Leech 1992, Baker 2010, Baker 2011, Baker 2012, Baron et al 2009, Hofland and Johansson 1982, Leech and Fallon 1992, Oakes and Farrow 2007, and Sigley and Holmes 2002. But as one of the most active researchers in this field notes (Baker 2011:70):

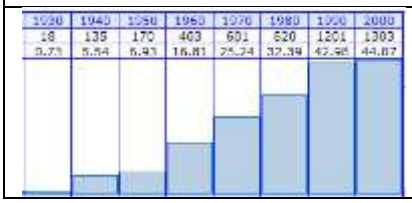
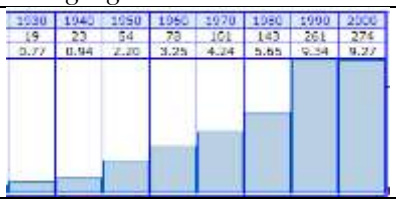
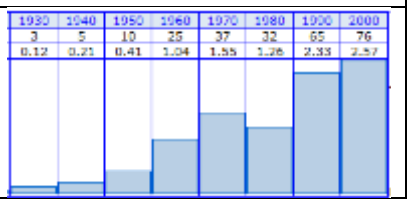
> Leech and Fallon (1992) point out that the corpora in the Brown family contain only about 50,000 word types in total, which is relatively small for lexical research, and that the majority of words will be too infrequent to give reliable guidance on British and American uses of language.
>
> For that reason, this study focuses only on frequent words in the corpora. It was stipulated that for a word to be of interest to this study, it would need to occur at least 1,000 times when its frequencies in all four corpora were added together. Three hundred eighty words met this criteria, but a number of high frequency words (e.g., *class, miss, black, true,* and *English*) were excluded because they missed the cutoff.

In this chapter, we will continue Baker's line of research, and show empirically what types of lexical data we can extract from small 1-2 million word corpora, compared to a much larger corpus like COHA. As a test case, we will briefly consider adjectives that have (at least) doubled in (normalized) frequency in COHA from the 1960s to the 1990s, and then examine how well the one million word Brown and Frown corpora (US, 1960s and 1990s) provide comparable

[2] See http://www.helsinki.fi/varieng/CoRD/index.html for a good overview of these genre-specific corpora.

evidence for this increase in frequency. In other words, in the data below we will be considering adjectives like *overall*, *emerging*, and *motivated*, whose charts in COHA are shown below.

Figure 16. COHA: Adjectives doubling in frequency, 1960s-1990s

overall	*emerging*	*motivated*
		

The following table shows that in COHA there are 15 adjectives that have a combined frequency of between 800-1600 tokens in the 1960s and 1900s (words such as *overall* (shown above), *amazing*, *long-term*, and *alternative*) and which have at least doubled in frequency during this time. There are another 127 types with a frequency of between 200-400 tokens in these two decades (e.g. *emerging* (shown above), *compelling*, *indoor*, *preferred*, and *unclear*), and 394 types with a frequency between of between 50 and 100 tokens (e.g. *motivated* (shown above), *first-time*, *blurry*, *impaired*, *viral*, *obnoxious*, and *luscious*).

Table 1. Evidence for increase in adjective frequency, COHA and Brown family

COHA: token range			800-1600	200-400	50-100
COHA: # of types			15	127	394
# Brown/Frown tokens	0		0	8	114
	1-9		1	46	264
	>= 10	Support	6	50	12
	>= 10	? ? ?	5	15	0
	>= 10	Contradict	3	8	4
Brown/Frown "correct"			0.40	0.39	0.03

Table 1 shows that for the 15 COHA adjectives that have at least doubled in frequency and which have a combined token frequency of 800-1600 in COHA in the 1960s and 1990s, all of these occur at least once in Brown/Frown, which is encouraging. One word occurs 1-9 times in Brown/Frown, and the other 14 occur at least 10 times (e.g. 3 tokens in Brown and 7 tokens in Frown), which is perhaps enough to show an increase from the 1960s to the 1990s. Of these 14 adjectives that occur at least 10 times, 6 do show frequency that has doubled from the 1960s-1990s (e.g. Brown 6, Frown 12, which is shown as "Support" (COHA) in Table 1 above). Another 5 adjectives show an increase, but less than the doubling in COHA (e.g. 6 Brown and 7 Frown, shown as "? ? ?" above). And in 3 cases, the Brown/Frown data actually shows a decrease from the 1960s to the 1990s (e.g. 7 Brown, 4 Frown; shown as "Contradict" above). Overall, then, 6 of the 15 types (40%) of these high-frequency adjectives in Brown / Frown show the same doubling in frequency that is shown in the robust data (800-1600 tokens) in COHA.

The situation is a bit less encouraging for the 127 medium-frequency adjectives (token count of 200-400 for the 1960s/1990s in COHA). Of these, 8 do not occur at all in Brown/Frown and 46

occur just 1-9 times, which is probably too few to see an increase. Of those occurring 10 times or more in Brown/Frown, 50 show a doubling, 15 show a smaller increase, and 8 show a decrease.

The situation with lower-frequency words is very poor. Remember, these are adjectives like *first-time*, *blurry*, *impaired*, *viral*, *obnoxious*, *luscious*, and *motivated* – less common to be sure, but certainly still the type of adjectives that most speakers of English would be familiar with. Of the 394 types in COHA with a frequency of between 50-100 and which have at least doubled in frequency, 114 of these do not occur at all in Brown/Frown, and another 264 occur less than 10 times – probably too few to be useful. As a result, Brown/Brown provides evidence for doubling in frequency for only about 3% of all of these lower-frequency adjectives from COHA.

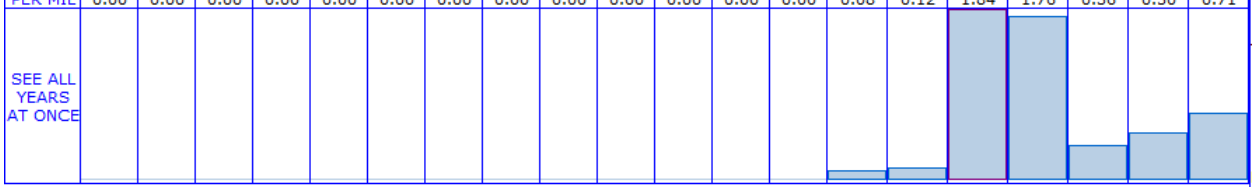
We should also realize that for some types of searches, the situation is even much worse than the searches just described, where we are simply looking at the frequency of a given word or phrase over time. For example, Figures 11-13 above shows that collocates are extremely sensitive to corpus size. In a one million word corpus (1/400th the size of COHA), virtually none of the collocates of *gay* or *woman* would occur more than one or two times. In summary, we argue that a corpus of 1 million words – while perhaps useful for high frequency grammatical changes – is simply too small to examine lexical changes with the vast majority of the words in the language.

In addition to size, one other problem with some of these small corpora is the issue of granularity. For example, the Brown family of corpora have texts from 1961 and 1991 (and work is proceeding on a similar corpus from 1931 and then 1901). But because there are texts from only every 30 years, any changes that take place in between these years is essentially "invisible", and in terms of lexical change, this is often too long of a gap.

Let us briefly consider two examples related to granularity, which are representative of tens of thousands of words. First, let us consider the frequency for *groovy* in COHA:

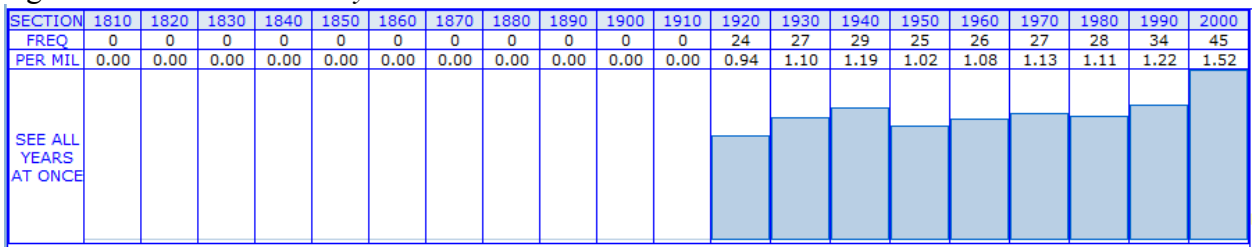
Figure 17. COHA: *groovy*

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	0	2	3	44	42	9	14	21
PER MIL	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.08	0.12	1.84	1.76	0.36	0.50	0.71



Imagine that our two corpora contained texts thirty years apart – from 1955 and 1985. In this case, it would appear (based on the COHA data from the 1950s and the 1980s) that *groovy* is on the increase. While it has increased slightly in these 30 years, we would miss entirely the steep decrease from the 1960s/1970s to the 1980s. Second, consider the case of *normalcy*:

Figure 18. COHA: *normalcy*



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	24	27	29	25	26	27	28	34	45
PER MIL	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.94	1.10	1.19	1.02	1.08	1.13	1.11	1.22	1.52

SEE ALL YEARS AT ONCE

This word was famously "rescued" from obscurity by President Warren G. Harding in 1920, who (according to purists) mistakenly used it instead of the more "correct" *normality*. The word caught on with a public tired of World War I and other foreign involvements, and Harding went on to win the election. But imagine that we only had two corpora from 1901 and 1931 (as with the planned extensions in the Brown family of corpora). There would obviously be a large increase in frequency between 1901 and 1931, but there would be no way to know if that predated Harding, whether his campaign caused the increase in usage, or whether it was after his time. In summary, corpora that have texts that are spaced decades apart may be adequate for looking at much more gradual grammatical change, but they are much more problematic in looking at lexical change, which can occur quite suddenly.

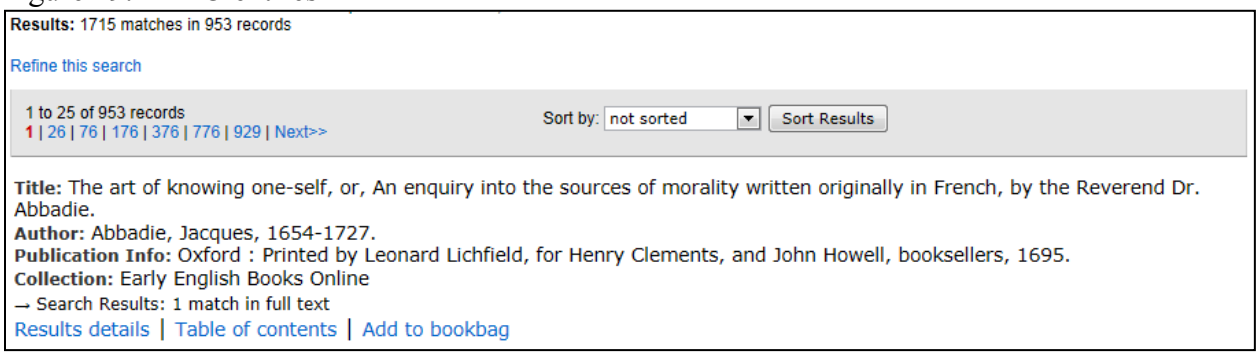
3. Text archives

If small corpora are of limited use in looking at lexical changes, then perhaps we might turn to text archives, some of which contain hundreds of millions of words of text (see López-Couso, this volume (especially Section 3) for a very good overview of available text archives). As we will see in this section, however, while the *textual* component of these text archives is truly impressive, they are often paired with an architecture and interface that make it very difficult to extract the lexical data in any meaningful way.

In this section, we will consider how we can find lexical data using Early English Books Online (EEBO), Literature Online (LION), and the New York Times newspaper archive. Note that as we discuss these archives, the precise type of data that we can get from each text archive is not the important thing – undoubtedly this will change over the next 5-10 years, and most of this will be outdated by then. In addition, there will undoubtedly be bigger and better text archives. Rather, this discussion and these resources are meant to serve simply as *examples* of some of the challenges inherent in using text archives.

EEBO represents one end of the spectrum: text archives that – while they have excellent textual data – have an interface that makes them quite problematic for lexical research. As of late 2012, EEBO contains about 700 million words in about 28,000 texts from 1470-1699, which is quite impressive. But consider what type of information is (and is not) available from the search interface. Suppose that we search for the word *polite*:

Figure 19. EEBO entries

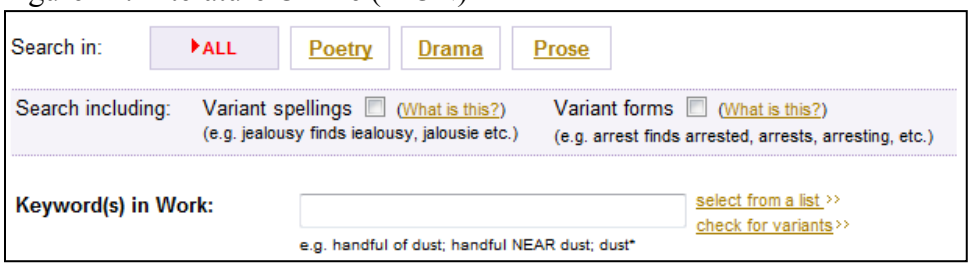


As with a typical Google web search, the results tell us how many texts the word or phrase appears in (953 texts, in this case). But there is no indication of frequency over time, as we can so easily find in COHA. Theoretically, it is possible to write a script to automatically retrieve all pages and then parse the page to extract dates and number of tokens and then import and process this in a database, but this is probably beyond the ability of most users. As far as collocates or concordances, we would need to click individually – one after another – on all 953 texts to see the word or phrase in context, as in the following.

Figure 20. EEBO concordance lines

An example of a somewhat more useful text archive is Literature Online, which contains about 600-700 million words of text from Early Modern English and Late Modern English, and which continues to grow in size. It provides some very useful features that EEBO does not. First, it is possible to search for spelling variants and for variant forms, as well as with simple proximity and Boolean operators:

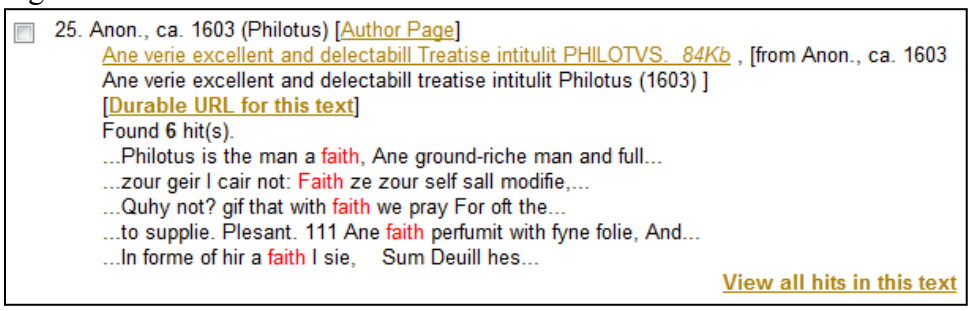
Figure 21. Literature Online (LION)



While it is still not possible to see the frequency of a word over time, with LION one could potentially search for a word in (for example) 1500-1509, and then 1510-1519, and so on, and after each search copy the number of hits to a program like Excel, and then generate a frequency chart from all of these searches. Finally, in Literature Online (unlike in EEBO) it is not necessary

to click on each text to see the word in context. Each page contains the word or phrase in context for up to ten texts at a time:

Figure 22. Literature Online concordance lines



It would still be necessary to write a script to extract all of this information from (perhaps) hundreds of pages, and then process it in another program to create concordance lines and to extract collocates for all tokens, but it is probably still more doable than in EEBO.

As a final example of text archives, let us consider online historical newspapers. In the past 5-10 years, many historical archives for many newspapers have become available online, and many more will undoubtedly become available in the future as well. The question is whether and how this data can be used for lexically-oriented research. As an example, we will briefly consider the New York Times archive, which contains 300-400 million words from the 1850s-2000s. A search for the phrased *messed up* yields results like the following:

Figure 23. New York Times: concordance lines and frequency chart

As with other text archives, some results (as in entry #1 above) yield short Keyword in Context entries, which with some effort could be aggregated to created KWIC and/or collocates for all entries. Unfortunately, many other entries (such as in #2) only provide links to a PDF file, which means that users have to scan through that entire graphics image to find the phrase *messed up* (the PDF file is not searchable), and researchers would have to do this for each of the hundreds or thousands of entries. (This type of linking to graphical/non-text PDF images is common in text archives.) In this sense, the New York Times archive is even less linguistics-friendly than Literature Online.

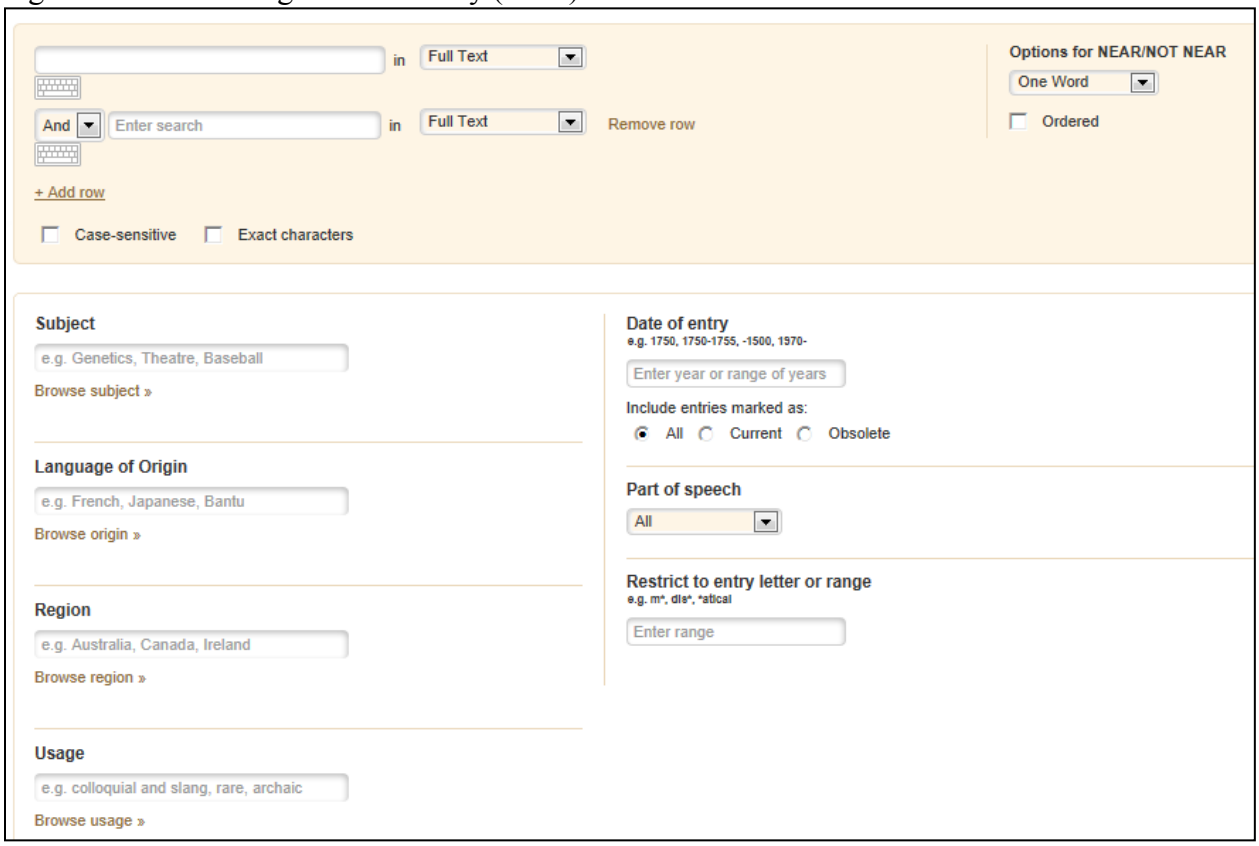
However, note the frequency chart (#3 above), which shows the frequency of *messed up* in each decade from the 1890s to the 2000s. Users can click on any decade to see the frequency by year, and they can also limit the KWIC entries to just that decade or year. While it is interesting and encouraging to see such linguistics-friendly features in the interface, there are some important limitations. Most importantly, the frequency data is not normalized for corpus size in each

decade. This means that researchers would have to first calculate the "base frequency" for a very common word like *the* or *and* (which presumably change very little from decade to decade), and then compare the frequency of any given word or phrase (like *messed up*) to this base frequency. Nevertheless, the inclusion of frequency information in a non-linguistic-oriented resource– in any form – is still a start.

4. The Oxford English Dictionary

As anyone who has used it can testify, in many respects the OED is unparalleled in terms of the information that it can provide on lexical changes in English. As the following screen shot of the online OED interface shows, we can perform searches using an incredible range of information in the OED dictionary entries– subject, language of origin, region, usage, and part of speech.

Figure 24. Oxford English Dictionary (OED) interface



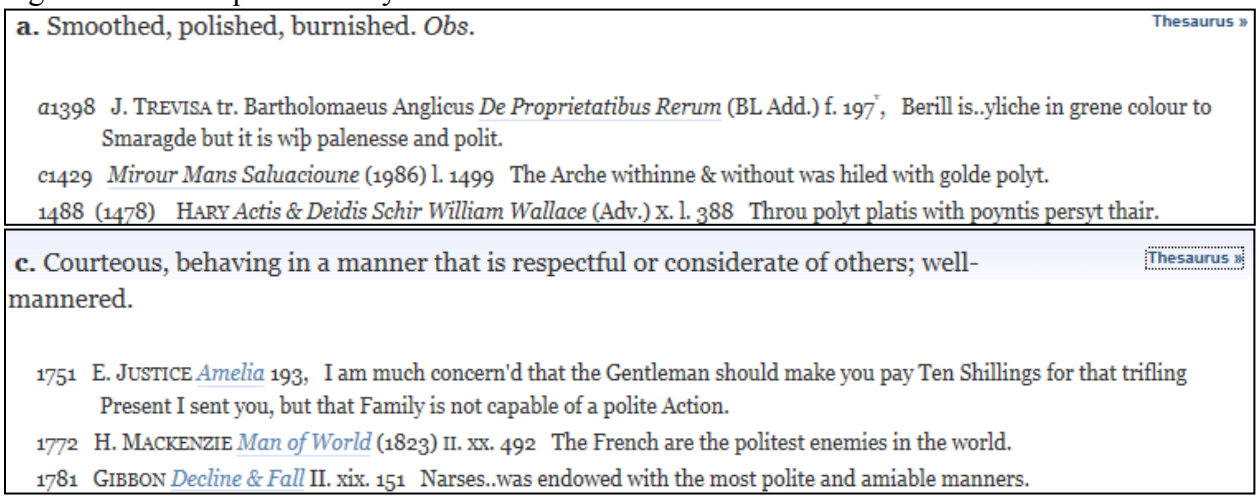
For example, we could search for Subject = Arts, Language of Origin = Italian, Date of entry = 1700-1720, and Part of speech = noun, and we would retrieve the words and phrases *al fresco, allegretto, capitolo, claro obscuro, mandolin, mezzanine, pianissimo, ricercata, ridotto, rondo, smalto*, and *stuccature*.

For the purpose of this chapter, however, which is more oriented towards corpus-based approaches to lexical changes in English, we might consider how the OED can be used as a corpus – a searchable collection of texts – rather than just a collection of entries in a dictionary. Although some may not be accustomed to thinking of the OED as a corpus, we should remember

that it is based on approximately 38 million words of text from about 2.8 million entries, which means that we could – at least in principle – search it as a "corpus".

At the most basic level, we can search for a word or phrase and see that word or phrase in context. Consider first a few entries for the word *polite*:

Figure 25. OED: quotations by headword



In the case of the COHA data shown above (Figures 2-3, 11), we used re-sortable concordance lines and collocates to separate entries with different meanings. In the case of the OED, this has already been done for us, as seen above for the two meanings of *polite*.

But the OED interface also allows us to search for *any* words in the 2.8 million quotations, whether or not they are the headword (e.g. *polite*, above). For example, we could search for *trifling* and see the quotation for the headword *polite* from 1751 above or search for *endowed* and see the quotation for the headword *polite* from 1781. As a concrete example, the following are a handful of entries for *trifling*:

Figure 26. OED: searching for quotations by keyword

Even with such entries, however, we would still need to write a script to parse the web pages and extract the text and the dates from all of the quotation. In addition, as displayed in this view

(Advanced / Quotations search), the quotation only show 5-6 words to the left and to the right of the search term (rather than the entire quotation), which is a bit limiting. Finally, from a corpus linguistics perspective (which is the focus of this chapter), we might also wish that we had somewhat more control over the quotation searches in the OED, by being able to search in more advanced ways for strings of words (including variant forms and substrings) and then being able to quickly and easily display the results in frequency charts, concordance lines, or with collocates, none of which is possible.

The fact that the OED is still limited as a linguistic *corpus* should not, however, take away from its unique role as a resource for lexical changes in English, through the massive amount of data in its dictionary entries. And hopefully, at some point, the interface for the OED will be improved (as it has been in many other respects over the past few years), to allow an even greater range of lexically-oriented searches of the 38 million words of data in the 2.8 million quotations.

5. Google Books

The Google Books n-grams are based on hundreds of *billions* of words in more than a million books. It is composed of a number of different "datasets", such as American English (155 billion words), British English (34 billion words), and the "One Million Books" dataset (89 billion words). When the Google Books n-grams were released in late 2010, it was heralded as a resource that would revolutionize not only historical linguistics, but also a wide range of other fields within the scope of what was called "Culturomics", such as history, cultural studies, and sociology, by showing trends in word and phrase usage over the past few centuries (see Michel, Lieberman, et al, 2010).

The n-grams – single words and phrases up to five words in length – are searchable via the web interface, and they show the frequency of the words and phrases in each decade and year. For example, the following three charts show the frequency from the 1810s-2000s for the words *teenager*, *steamship*, and *many a time*. Compare these to the equivalent searches in COHA (Figures 1-3 above), and note how similar the frequency data is in the two resources.

Figure 27. Google Books: *teenager*

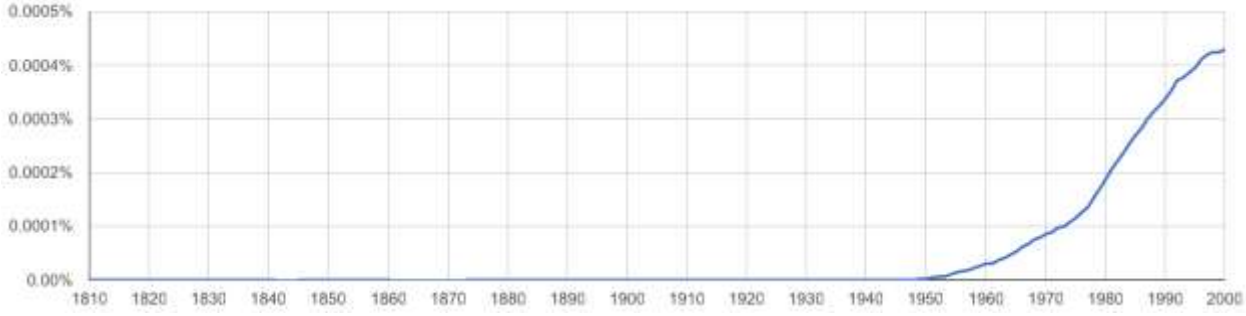


Figure 28. Google Books: *steamship*

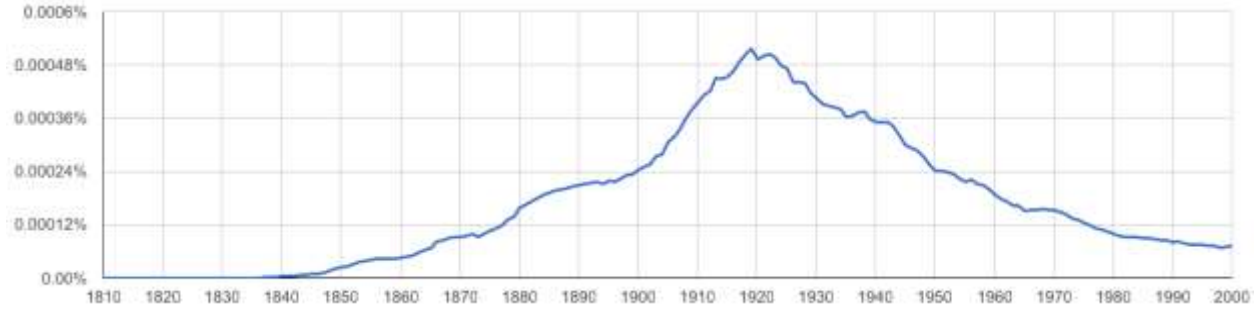


Figure 29. Google Books: *many a time*

While the total number of texts and words for the Google Books n-grams is huge, there are some critical weaknesses in terms of the interface to this data. First, the only frequency data are the pictures themselves – there is no actual frequency data that can be copied and pasted from the Google Books website. For example, [0.0002] in the chart for *many a time* is not an actual number – it is just a picture of that number. Second, the frequency data does not show the number of tokens – just these percentages. There is no way of knowing how many times a word or phrase actually occurs in a given decade or year. Third, because there are no actual token counts, it is impossible to calculate the total number of tokens for a given construction or even a single lemma. For example, one can search for *steamship* and then for *steamships* (cf. Figure 28), but these two "pictures" cannot be "added" together, as if there were actual frequency data. Fourth, it is not possible to search by part of speech. For example, one can search for *many a time* (Figure 29), *many a man*, or *many a mile*, but there is no way to search for *many a* [noun].

Finally, the only information available is the frequency of a given word or phrase over time. It would be difficult or impossible to create concordance lines or collocates for a given word or phrase. Researchers can click on a range of years in the chart to see Google-like "snippet" views with the word in context, as in Figure 30. But depending on the copyright status of the book, the context view may (#1 below) or may not (#2) be available. Even if it is, one would have to write a script to process the snippet entries and extract the relevant text (and years), and this is made difficult by the fact that Google limits snippet views to the first 1000 entries.

Figure 30. Google Books: "snippet" view

1	Improving teenage nutrition - Page 8 books.google.com/books?id=p2gvAAAAYAAJ United States. Dept. of Agriculture, United States. Federal Extension Service - 1963 - Read Friends. The **teenager** wants his peer group to like him. ... The **teenager** wants to be with others of his age group. ... The **teenager** needs to know how nutrition contributes to shiny hair, clear skin, good posture, and, of course, correct weight.

2	Adolescence and religion: the Jewish **teenager** in American society  books.google.com/books?id=5flCAAAAIAAJ Bernard Carl Rosen - 1965 - Snippet view

6. Summary of resources

At this point, it might be useful to review what types of lexical data can be obtained from different types of resources[3]. The following table might serve as a useful summary. Again, we recognize that the features of a given resource will change over time, and that new resources will become available. The point here is to see how well even the best resources are at providing information on lexical change.

Table 2. Lexically-oriented queries with different resources

	COHA	Small	EEBO	LION	NYT	OED	GB
1. Concordances for a particular word or phrase, to see the patterns in which it occurs	+	((~))	((+))	((+))	((+))	((+))	((+))
2. The frequency over time of specific words and phrases	+	((~))			(+)		+
3. Lists showing all words that are more frequent in one period than in another	+					(+)	
4. Collocates of a particular word or phrase (over time) in order to see semantic change	+						
5. Easy access to the relative frequency of related words, such as synonyms or words in a user-defined list	+	((~))				((+))	

[1] Other than COHA, none of the resources does a particularly good of providing concordance lines for all tokens of a word or phrase, although with some scripting and programming (or third party software, like WordSmith or AntConc) this can be extracted. **[2]** COHA and Google Books provide frequency data over time, but Google Books is limited by just showing pictures of frequency, with no real data, and the New York times frequency data is not normalized. **[3]** The OED comparisons of all words are based on first occurrences of words, but (unlike COHA) not on actual frequency data. **[4]** Other than COHA, none of the resources provide information on collocates. **[5]** The OED shows related words, via the Historical Thesaurus of English, but unlike COHA, it does not provide data to see the relative frequency of these words over time. Finally, a few of the features are possible with small corpora (Section 3), but only for high frequency words – and only if they are somehow accessed with the right corpus interface.

[3] COHA = Corpus of Historical American English, EEBO = Early English Books Online, LION = Literature Online, NYT = New York Times, OED = Oxford English Dictionary, GB = Google Books

7. Extending useful lexical searches to other corpora

As we have seen, COHA (and COCA) are able to carry out a wide range of searches on lexical changes in Late Modern English, which in many cases would be difficult or impossible with other corpora. However, it is important to understand that once this architecture and interface are available for corpora like COHA and COCA, they can in principle he applied to any other set of texts, including those from text archives. For example, the TIME Magazine corpus (Davies 2012c) contains 100 million words in more than 275,000 texts from the Time Magazine text archive from the 1920s-2000s (http://www.time.com/time/archive), and all of the lexically-oriented searches that are possible with COHA and COHA can be carried out in this "corpus" as well.

In addition, a modified very of the COHA/COCA interface has been applied to the Google Books n-grams data (see http://googlebooks.byu.edu), to allow a wide range of searches on this 200+ billion words of data, which are not possible with the standard Google Books interface. In the interest of space, we will provide only four examples of the types of searches that can be done with the BYU/Advanced Google Books interface (GB-BYU), which are not possible with the standard Google Books interface (GB-Standard), but many more sample searches are available at the corpus website.

First, the GB-BYU interface shows the raw and normalized frequency for words and phrases – they are not just pictures of frequency as in GB-Standard:

Figure 31. Frequency of *steamship* in BYU/Advanced Google Books (GB-BYU)

DECADE	1810	1820	1830	1840	1850	1860	1870	1880	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000
SIZE (MIL)	378	655	1,437	1,938	2,953	2,353	2,844	4,408	5,632	7,520	10,087	7,089	5,795	6,167	8,104	13,192	14,011	15,511	19,816	26,882
TOKENS	4	0	81	391	1,683	2,311	3,421	9,327	13,860	23,645	46,966	38,558	21,807	21,003	18,251	22,104	19,594	17,377	17,768	19,285
PER MIL	0.01	0.00	0.06	0.20	0.57	0.98	1.20	2.12	2.46	3.14	4.66	5.44	3.76	3.41	2.25	1.68	1.40	1.12	0.90	0.72

Because there are actual numbers, the charts in GB-BYU can represent the totals for lemmas or entire constructions as well. It is also possible to see the frequency of each matching string, as with the construction *many a* [noun] (cf. COHA in Figure 8 above). Note that GB-BYU allows searching by part of speech, unlike GB-Standard.

Figure 32. GB-BYU: Frequency of *many a* [noun]

	WORD(S)	CHARTS	TOTAL	1810	1820	1830	1840	1850	1860	1870	1880	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000
1	many a time	G B	100393	238	534	1533	2318	4440	3915	4832	7827	11123	11791	11680	7342	5057	3721	4081	5483	4141	2997	3341	3861
2	many a day	G B	64963	131	425	1084	1922	3555	2998	3461	5173	8067	8012	7875	4449	2874	2098	2265	3198	2358	1452	1704	1861
3	many a man	G B	60125	157	309	790	1024	2012	1877	2381	3678	5204	6775	8161	5224	3369	2805	3036	4407	2748	1813	2134	2201
4	many a year	G B	53485	132	334	924	1532	2789	2178	2726	4182	6197	5768	6440	3788	2806	2083	2148	2839	2010	1385	1587	1860
5	many a night	G B	15933	38	85	208	335	686	576	623	932	1371	1486	1746	1117	687	640	734	841	813	735	924	1339
6	many a mile	G B	12920	60	76	291	473	809	586	734	1146	1392	1441	1317	896	599	430	444	655	437	239	303	369
7	many a heart	G B	12174	69	146	452	955	1577	1007	895	1174	1157	1193	1118	679	333	219	238	367	391	194	215	235
8	many a woman	G B	8821	23	26	80	123	248	229	266	488	732	944	1130	702	429	374	474	516	609	387	497	535
9	many a battle	G B	7074	13	27	110	240	361	459	360	595	695	745	753	494	320	283	220	450	271	188	223	278
10	many a tale	G B	6557	23	81	239	328	529	356	372	488	513	642	771	379	311	210	197	303	234	181	165	273

As with COHA, but unlike GB-Standard, the GB-BYU interface and architecture allows users to use wildcards to see the frequency of all matching forms in each decade, for example the

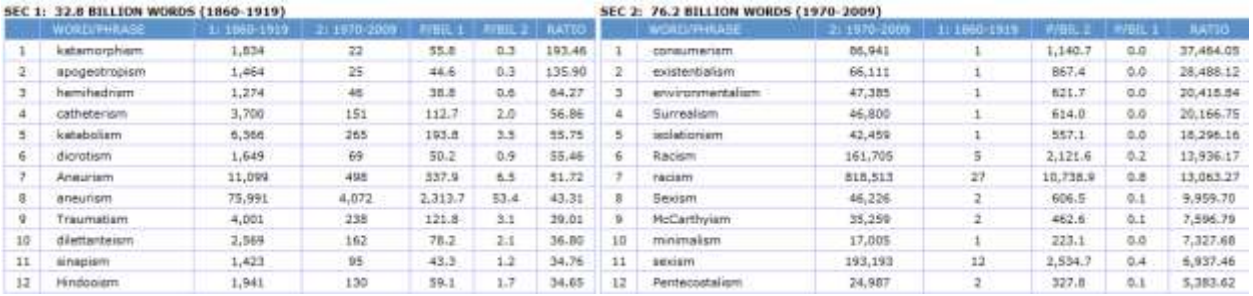
frequency of **ism* words (note the increase in *criticism*, *organism*, and *capitalism*, and the decrease in *baptism* and *patriotism*):

Figure 33. GB-BYU: **ism* words by decade

	WORD(S)	CHARTS	TOTAL	1810	1820	1830	1840	1850	1860	1870	1880	1890	1900	1910	1920	1930	1940	1950	1960
1	criticism	G B	5426220	5120	7951	17160	26366	42630	39076	52769	105656	169341	247386	346591	280083	245163	240586	333238	618744
2	mechanism	G B	4885909	1327	2517	7553	9308	15781	11640	15417	34598	52826	84166	177370	145663	130122	188694	271249	501658
3	organism	G B	2756106	25	70	395	4646	14857	16213	31312	56855	97622	159625	273762	177318	136206	133460	205589	318576
4	metabolism	G B	2071179	5			8		2	89	2054	7616	40725	99471	67019	53494	61063	103894	199693
5	Judaism	G B	1460561	910	1163	3461	7539	9543	9886	14113	21564	46153	36752	50449	30322	36967	46855	81562	131255
6	capitalism	G B	1427387	1		2	18	5	3	9	522	1628	7820	22454	29741	74921	75324	72496	155756
7	baptism	G B	1369446	19435	16361	49645	85027	91912	51318	59473	70753	73932	80670	80490	58122	26353	34589	51327	87642
8	socialism	G B	1181815	5	3	3	258	917	666	1262	7143	21972	33349	55803	44882	50694	61740	82035	187268
9	patriotism	G B	1175726	5406	10802	21882	31023	48202	42091	34389	54712	74093	93479	137359	83007	56389	48837	49261	94320

Like COHA, but unlike GB-Standard, in GB-BYU one can find all words that are more common in one period than another, for example **ism* words more common in the 1860s-1910s or the 1970s-2000s (compare Figure 10 from COHA, above):

Figure 34. GB-BYU: Comparison of **ism* words, 1860s-1910s vs 1970s-2000s



SEC 1: 32.8 BILLION WORDS (1860-1919)

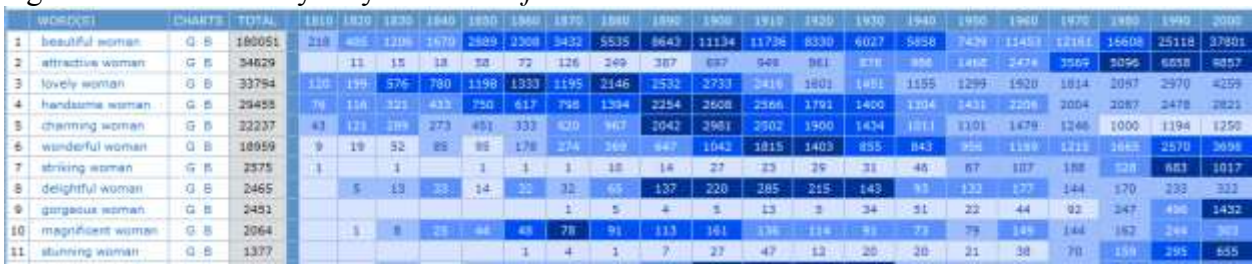
	WORD/PHRASE	1: 1860-1919	2: 1970-2009	#/BIL 1	#/BIL 2	RATIO
1	katamorphism	1,834	22	55.8	0.3	193.46
2	apogeotropism	1,464	25	44.6	0.3	135.90
3	hemihedrism	1,274	46	38.8	0.6	84.27
4	catheterism	3,700	151	112.7	2.0	56.86
5	katabolism	6,566	265	199.8	3.5	55.75
6	dicrotism	1,649	69	50.2	0.9	55.46
7	Aneurism	11,099	498	337.9	6.5	51.72
8	aneurism	75,991	4,072	2,313.7	53.4	43.31
9	Traumatism	4,001	238	121.8	3.1	39.01
10	dilettanteism	2,569	162	78.2	2.1	36.80
11	sinapism	1,423	95	43.3	1.2	34.76
12	Hindooism	1,941	130	59.1	1.7	34.65

SEC 2: 76.2 BILLION WORDS (1970-2009)

	WORD/PHRASE	2: 1970-2009	1: 1860-1919	#/BIL 2	#/BIL 1	RATIO
1	consumerism	86,941	1	1,140.7	0.0	37,464.05
2	existentialism	66,111	1	867.4	0.0	28,488.12
3	environmentalism	47,385	1	621.7	0.0	20,418.84
4	Surrealism	46,800	1	614.0	0.0	20,166.75
5	isolationism	42,459	1	557.1	0.0	18,296.16
6	Racism	161,705	5	2,121.6	0.2	13,936.17
7	racism	818,513	27	10,738.9	0.8	13,063.27
8	Sexism	46,226	2	606.5	0.1	9,959.70
9	McCarthyism	35,259	2	462.6	0.1	7,596.79
10	minimalism	17,005	1	223.1	0.0	7,327.68
11	sexism	193,193	12	2,534.7	0.4	6,937.46
12	Pentecostalism	24,987	2	327.8	0.1	5,383.62

GB-BYU allows users to use semantically-oriented queries with synonyms, such as *[=beautiful] woman* (compare the COHA data in Figure 15 above):

Figure 35. GB-BYU: Synonyms: "*beautiful*" + *woman*



	WORD(S)	CHARTS	TOTAL	1810	1820	1830	1840	1850	1860	1870	1880	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000
1	beautiful woman	G B	180051	210	486	1206	1670	2899	2308	3432	5535	8643	11134	11736	8330	6027	5858	7429	11453	12161	16608	25118	37801
2	attractive woman	G B	34629		11	15	18	58	72	126	249	387	687	948	961	878	886	1468	2474	3569	5096	6658	9857
3	lovely woman	G B	33794	120	199	576	780	1198	1333	1195	2146	2532	2733	2419	1801	1481	1155	1299	1920	1814	2097	2970	4259
4	handsome woman	G B	29455	79	116	325	433	750	617	798	1394	2254	2608	2566	1791	1400	1204	1431	2206	2004	2087	2478	2821
5	charming woman	G B	22237	43	121	289	273	451	333	620	967	2042	2981	2502	1900	1424	1011	1101	1479	1246	1000	1194	1250
6	wonderful woman	G B	18959	9	19	52	89	86	178	274	369	847	1042	1815	1403	855	843	956	1199	1215	1665	2570	3698
7	striking woman	G B	2575	1		1		1	1	1	10	14	27	23	29	31	46	67	107	188	328	663	1017
8	delightful woman	G B	2465		5	13	33	14	32	32	65	137	220	285	215	143	93	132	177	144	170	238	322
9	gorgeous woman	G B	2451							1	5	4	5	13	3	34	51	22	44	92	247	496	1432
10	magnificent woman	G B	2064		1	8	23	46	48	78	91	113	161	136	114	91	73	79	149	144	162	244	303
11	stunning woman	G B	1377						1	4	1	7	27	47	12	20	20	21	38	70	159	295	655

Like COHA, but unlike GB-Standard, in GB-BYU one can compare collocates in different periods, for example adjectives with *woman* in the 1850s-1910s (left) and the 1970s-2000s (right) (compare the COHA data in Figure 13 above):

Figure 36. GB-BYU: Comparison of collocates with *women*, 1850s-1910s vs 1970s-2000s

SEC 1: 35.8 BILLION WORDS (1850-1919)

	WORD/PHRASE	1: 1850-1919	2: 1970-2009	P/BIL 1	P/BIL 2	RATIO
1	feeble women	816	149	22.8	2.0	11.66
2	fair women	8,212	1,529	229.4	20.1	11.44
3	Fair women	524	104	14.6	1.4	10.73
4	chief women	699	139	19.5	1.8	10.71
5	delicate women	2,651	605	74.1	7.9	9.33
6	defenceless women	1,474	341	41.2	4.5	9.20
7	tender women	816	190	22.8	2.5	9.14
8	noblest women	975	229	27.2	3.0	9.07
9	handsomest women	1,037	249	29.0	3.3	8.87
10	nervous women	1,895	476	52.9	6.2	8.48

SEC 2: 76.2 BILLION WORDS (1970-2009)

	WORD/PHRASE	2: 1970-2009	1: 1850-1919	P/BIL 2	P/BIL 1	RATIO
1	bisexual women	10,784	1	141.5	0.0	5,064.71
2	battered women	83,346	10	1,093.5	0.3	3,914.35
3	heterosexual women	21,388	4	280.6	0.1	2,511.22
4	academic women	4,253	1	55.8	0.0	1,997.42
5	negative women	3,696	2	48.5	0.1	867.91
6	urban woman	10,521	7	138.0	0.2	705.88
7	Black women	102,960	69	1,350.8	1.9	700.80
8	overweight women	4,287	3	56.2	0.1	671.13
9	Inuit women	1,163	1	15.3	0.0	548.20
10	Jamaican women	1,153	1	15.1	0.0	541.51

As we have seen, GB-Standard and GB-BYU use exactly the same n-grams datasets, which Google has graciously made available to others. But due to its unique architecture and interface, GB-BYU allows a wide range of lexically-oriented searches that are quite impossible with the standard, simplistic Google Books interface.

To conclude this section, one might ask why one would want to use a "small" corpus like COHA (at a "mere" 400 million words) when they could use the Google Books n-grams, which is based on a dataset that is nearly 400 times as large, at 155 billion words.

First, Google Books are just n-grams – nothing longer than a five word string can be searched (where in COHA, the search string can be up to 21 words in length). Even more importantly, these Google Book n-grams are limited to just those that occur 40 times, which dramatically reduces the number of types available (where in COHA, *all* strings – even those occurring just once or twice – are included). In addition, there is no contextual part of speech disambiguation in Google Books (as there is in COHA), making it impossible to search for "nouns" or "verbs" (or any other part of speech) in isolation. And finally, Google Books (whether GB-Standard or GB-BYU) only displays the actual text – the original sentences and perhaps paragraphs – in individual "snippets" (see Figure 30 above), whereas they are displayed in COHA in a much more readable and usable format (see Figure 1-3, 9).

So even with the important improvements that GB-BYU makes on the basic GB-Standard interface, there are still many types of lexically-oriented searches that are only available with COHA, since it alone of the two is a true "linguistic corpus".

8. Early Modern English

Nearly all of the discussion to this point has focused on Late Modern English, via COHA, Google Books, text archives, and smaller corpora like the Brown family of corpora. But there is no reason that the same interface that is used for COHA, COCA, TIME, and GB-BYU could not be applied to texts from Early Modern English as well, to provide for a wide range of searches on lexical change and variation. In fact, we have applied this same interface to 400 million words of data in nearly 30,000 texts from Early English Books Online. Due to space limitations here, we will provide only a handful of examples of lexically-oriented searches from EEBO-BYU.

As with the other corpora, we can see concordance lines for a particular word, phrase, or construction. For example, the following are a handful of lines from EEBO-BYU for the word *wise* from the 1470s-1550s:

Figure 37. EEBO-BYU: Concordance lines for *wise*, 1470s-1550s

36	1506	The mirroure of gold	A	B	C	of Iugemente shall see the blessed creatures / not in that maner	wyse	to knowe their loué what it is : but only they shall
37	1507	Scala perfectionis.	A	B	C	and punfyed twenty wynter / And therfore vpon this maner	wise	take in my sayenge as I haue sayd . and na ?-*ly
38	1534	A mustre of scismaty	A	B	C	sende euery good man in a iust cause [] somwhat more	wyser	cou?sayle than he were lyke to haue of suche men , whiche
39	1549	A contrarye (to a ce	A	B	C	; preacher of Christ . Wherfore Christe willeth not in my	wise	these kodes or damel of hereti-*ques , and not of Adulterers
40	1533	The dabellacyon of S	A	B	C	Page xvsiAnd therfore me thynkath that this deuice is not mych	wyser	, than the de-*uice that a good felow deuised ones for his
41	1553	The vocacyon of Ioha	A	B	C	my tryall / I desired to go a la?de/ but it no	wyse	wolde it be graunted .* After that we passed more than the
42	1537	The boke named the G	A	B	C	sustained at the batayle of Cannas , they could in no	wise	delyuer them . Wherfore they dyscharged them of theyr promyse ,

Second, we can search for frequency of words and phrases in Early Modern English. In the following, we search for the verb *secure* as part of the string *to secure*, to distinguish it from *secure* as an adjective, since the corpus is not contextually disambiguated.

Figure 38. EEBO-BYU: Frequency of *to secure*

SECTION	1470s	1480s	1490s	1500s	1510s	1520s	1530s	1540s	1550s	1560s	1570s	1580s	1590s	1600s	1610s	1620s	1630s	1640s	1650s	1660s	1670s	1680s	1690s
FREQ	0	0	0	0	0	0	0	0	0	1	3	1	16	78	109	217	232	583	549	617	737	899	1040
PER MIL	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.09	0.17	0.05	0.67	2.53	3.58	6.99	7.57	18.73	17.63	20.49	24.06	29.17	33.19

SEE ALL SUB-SECTIONS AT ONCE

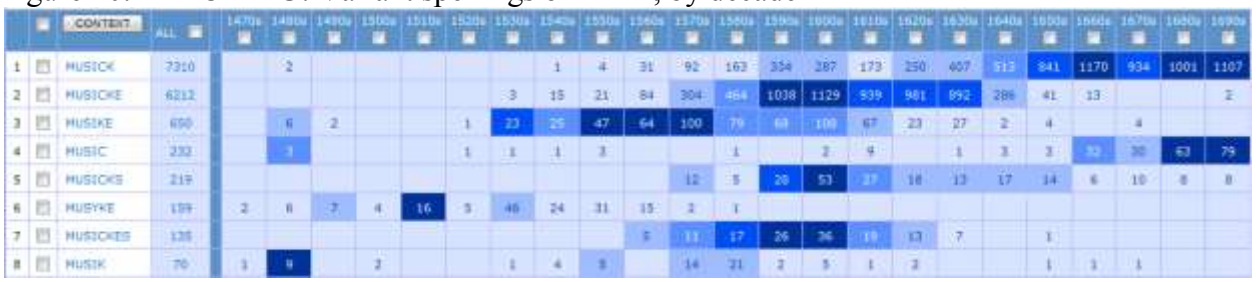
As with the other corpora, we can see the frequency of all matching strings by decade, as with words ending in **ism*. (Note that earlier spellings with **isme* or **ysm*(*e*) are not included here.)

Figure 39. EEBO-BYU: **ism* words by decade

	CONTEXT	ALL	1470s	1480s	1490s	1500s	1510s	1520s	1530s	1540s	1550s	1560s	1570s	1580s	1590s	1600s	1610s	1620s	1630s	1640s	1650s	1660s	1670s	1680s	1690s
1	BAPTISM	14051								8	6	12	9	75	28	74	251	185	146	578	2885	1757	2309	2785	2943
2	SCHISM	4962												6	1	10	16	23	3	150	867	467	900	1487	1132
3	ATHEISM	1833												1	2	6	6	11	5	23	119	392	372	399	497
4	CATECHISM	1396										1		6	3		5	3	20	34	90	189	278	421	346
5	SYLLOGISM	634												3	2	1	1	8		17	45	63	132	105	257
6	PAGANISM	552													1	1	16	5		21	40	129	105	111	123
7	INFANT-BAPTISM	524																		9	173	36	82	70	154
8	ANTICHRISTIANISM	332														2	18			11	7	218	43	21	12
9	JUDAISM	314																		9	49	43	65	87	81
10	QUAKERISM	284																			2	12	96	93	121

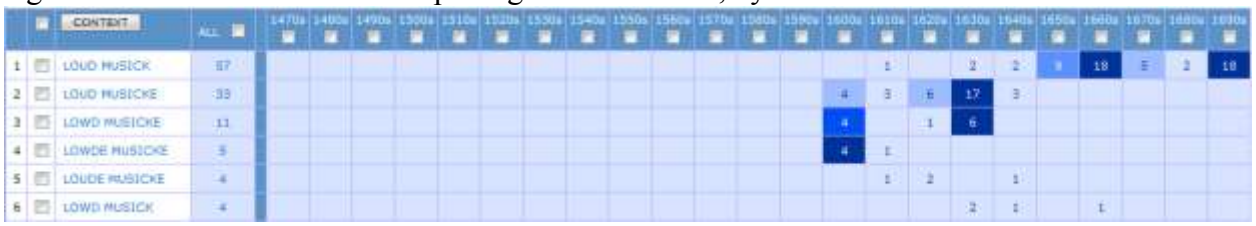
In a text of Early Modern English, there are of course several challenges. The first, of course, is spelling variation (see Smitterberg, this volume). However, we have received from Martin Mueller at Northwestern University a database of more than 1,350,000 variant spellings for more than 1,020,000 word forms (from approximately 73,000 lemmas). Because these forms are in a database that is linked to the standard search interface and architecture, we can include variant forms as part of queries. For example, Figure 40 shows a few of the 25 different forms for *music* and their frequency in each decade, and Figure 41 shows the frequency for a few of the variants of *loud music*.

Figure 40. EEBO-BYU: Variant spellings of *music*, by decade



	CONTEXT	ALL	1470s	1480s	1490s	1500s	1510s	1520s	1530s	1540s	1550s	1560s	1570s	1580s	1590s	1600s	1610s	1620s	1630s	1640s	1650s	1660s	1670s	1680s	1690s
1	MUSICK	7310		2						1	4	31	92	163	354	287	173	250	407	513	841	1170	934	1001	1107
2	MUSICKE	6212							3	15	21	84	304	464	1038	1129	939	981	892	288	41	13			2
3	MUSIKE	690		6	2			1	23	25	47	64	100	79	68	108	67	23	27	2	4		4		
4	MUSIC	232		3				1	1	1	3			1		2	9		1	3	3	32	30	63	79
5	MUSICKS	219											12	5	20	53	27	18	13	17	14	6	10	8	8
6	MUSYKE	199	2	8	7	4	16	5	46	24	31	15	2	1											
7	MUSICKES	135										5	11	17	26	36	19	13	7		1				
8	MUSIK	70	1	9		2			1	4	5		14	21	2	5	1	3			1	1	1		

Figure 41. EEBO-BYU: Variant spellings of *loud music*, by decade

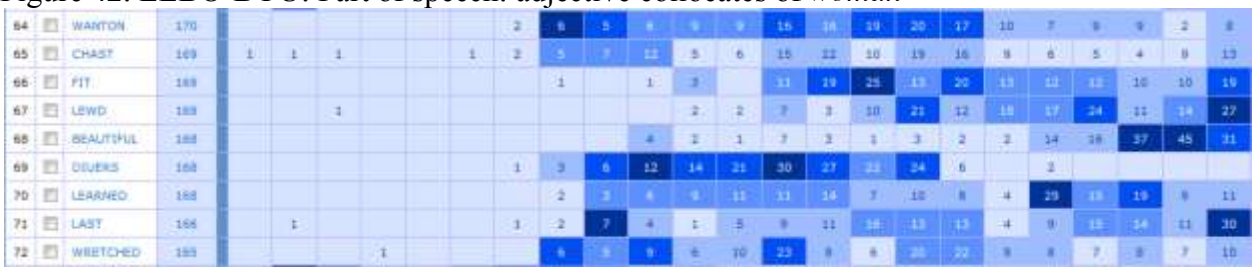


	CONTEXT	ALL	1470s	1480s	1490s	1500s	1510s	1520s	1530s	1540s	1550s	1560s	1570s	1580s	1590s	1600s	1610s	1620s	1630s	1640s	1650s	1660s	1670s	1680s	1690s
1	LOUD MUSICK	87															1		2	2	9	18	5	2	10
2	LOUD MUSICKE	33														4	3	6	17	3					
3	LOWD MUSICKE	11														4		1	6						
4	LOWDE MUSICKE	5														4	1								
5	LOUDE MUSICKE	4															1	2		1					
6	LOWD MUSICK	4																	2	1		1			

The second challenge is part of speech (see Smitterberg, this volume). We have extracted part of speech data from the Penn Helsinki Parsed Corpus of Early Modern English (PCEME). Words are assigned a given part of speech in the EEBO-BYU lexicon if they occur three times or more with that part of speech in the PCEME. For example, *poore*, *certaine*, *generall*, *faire*, *diuers*, *naturall*, *divers*, *lawfull*, *lyke*, *greate*, *neere*, *equall*, *litle* are all (non-PDE spelling) adjectives in our lexicon, because each of these is tagged as an adjective at least three times in the PCEME.

Although this solution is not perfect, it does allow it to use provisional part of speech tags in searches, as in the following collocates of *woman* in EEBO-BYU.[4] [5]

Figure 42. EEBO-BYU: Part of speech: adjective collocates of *woman*



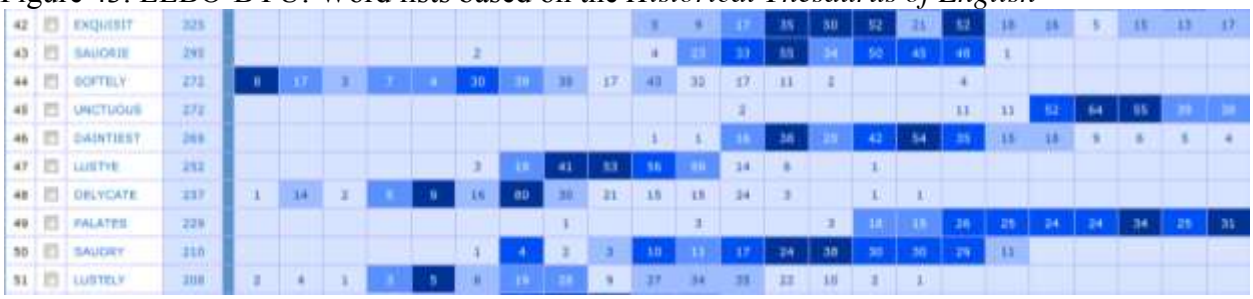
64	WANTON	170								2	6	5	8	9	9	15	16	19	20	17	10	7	9	9	2	8
65	CHAST	169	1	1	1			1	2	5	7	12	5	6	15	22	10	19	16	9	6	5	4	8	13
66	FIT	168								1		1	3		11	19	25	13	20	13	12	12	10	10	19
67	LEWD	168			1								2	2	7	3	10	21	12	18	17	24	11	14	27
68	BEAUTIFUL	168									4	2	1	7	3	1	3	2	2	14	18	37	45	31	
69	DIUERS	168							1	3	6	12	14	21	30	27	22	24	6		2				
70	LEARNED	168								2	3	8	9	11	11	14	7	10	8	4	29	15	19	9	11
71	LAST	166		1					1	2	7	4	1	5	9	11	18	13	13	4	9	15	14	11	30
72	WRETCHED	165				1				6	3	9	6	10	23	8	6	20	22	8	8	7	8	7	10

The third challenge are the synonyms. Obviously, a modern thesaurus (as we use with COHA or COHA or BYU Google Books) will be inadequate for Early Modern English. To compensate for this, we use an electronic version of the Historical Thesaurus of English (HTE), and allow users to create "customized word lists" based on the HTE. For example, the following table shows words from the HTE relating to "sweet" (note that variant spelling forms are included as well):

[4] Because entries are taken from further down in the list, the decade headers are not shown for the columns in this figure or in the following two figures.

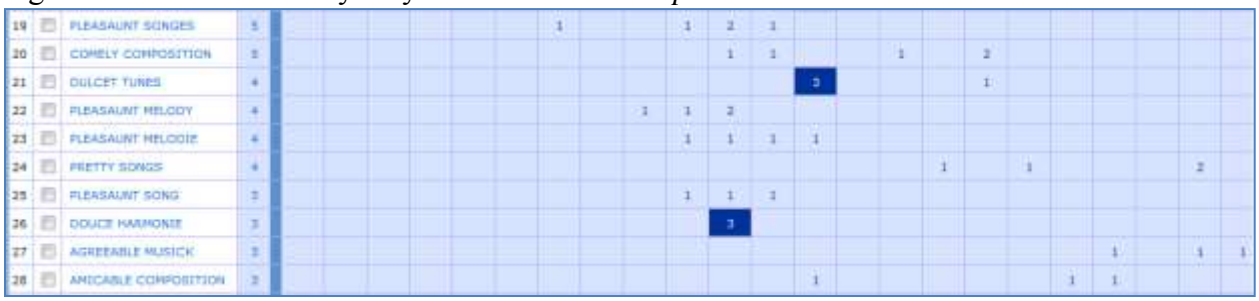
[5] *fit* (#66) is probably not an adjective in these cases (since there is no contextual disambiguation), the other words probably are adjectives

Figure 43. EEBO-BYU: Word lists based on the *Historical Thesaurus of English*



These HTE-based customized word lists can then be used as part of queries, such as the following, which finds words in the HTE-based list for "pleasant" followed by words in the HTE-based list for "music" (note the alternate spellings as well). While certainly not perfect, this approach does allow users to perform semantically-oriented searches, rather than just simply searching for exact strings.

Figure 44. EEBO-BYU: Synonym-based searches: "*pleasant*" + "*music*"



These are just a handful of examples of how we can use a robust textual corpus (the 400 million word EEBO) and a robust search interface to carry out lexically-oriented research on Early Modern English. Several problems remain, and the corpus is not yet publicly-available (due to licensing issues with the Text Creation Partnership, the creators of EEBO). Nevertheless, it will hopefully give an idea of what can be done even with Early Modern English texts, with the right tools.

9. Discussion and conclusions

As we have discussed in this paper, there are two crucial features of corpora that need to be present, in order for us to be able to look at this wide range of lexical (and semantic) shifts. The first is size. As we have discussed at some length in Section 2, small 1-2 million word corpora are perfectly fine for looking at high frequency syntactic constructions, but they are usually far too small to adequately investigate medium and low-frequency words and phrases. But size is not everything. As we have also discussed in Section 3-5, the data in extremely large text archives is essentially "trapped" inside, unless we have the right corpus architecture and interface to efficiently access this data.

As we have seen in Section 1 (as well as Section 7), there are a handful or corpora (such as COHA, TIME, and Google Books – Advanced) that have recently become available, which are both very large and which also have a robust architecture and interface. With these corpora, we

can (at the most basic level) use re-sortable concordances to search for individual tokens of a word or phrase, and to see the patterns in which they occur. We can also find the frequency of a word or phrase over time, and (as a more advanced search) we can produce lists of all words that are more frequent in one period than in another. Moving more towards semantic change, with these corpora we can also use collocates to look at changes in meaning over time. And since words do not occur in isolation paradigmatically, we can search for the frequency and use of related words over time.

In summary, due to recent advances in English historical corpora, we can now examine lexical change in English in a number of ways that would have been quite impossible even 4-5 years ago. And because these new architectures and interfaces can be applied to new datasets (such as Early English Books Online, as discussed in Section 8), the next few years will likely see even more advances in terms of examining historical English lexis.

References

Baker, Paul. 2009. The BE06 Corpus of British English and recent language change. *International Journal of Corpus Linguistics* 14(3). 312-337.

-------. 2010. 'Will Ms ever be as frequent as Mr? A corpus-based comparison of gendered terms across four diachronic corpora of British English.' *Gender and Language* 4.1: 125-129.

-------. 2011. 'Times may change but we'll always have money: a corpus driven examination of vocabulary change in four diachronic corpora.' *Journal of English Linguistics* 39: 65-88.

Baron, Alistair, Paul Rayson, and Dawn Archer. 2009. Word frequency and key word statistics in corpus linguistics. *Anglistik* 20: 41-67

Davies, Mark. 2011. "The Corpus of Contemporary American English as the First Reliable Monitor Corpus of English". *Literary and Linguistic Computing* 25: 447-65.

-------. 2012a. “Expanding Horizons in Historical Linguistics with the 400 million word Corpus of Historical American English”. *Corpora* 7: 121-57.

-------. 2012b. "Examining Recent Changes in English: Some Methodological Issues". In *Handbook on the History of English: Rethinking Approaches to the History of English*, eds. Terttu Nevalainen and Elizabeth Closs Traugott. Oxford: Oxford Univ. Press. 263-87.

-------. 2012c "Recent shifts with three nonfinite verbal complements in English: Data from the 100 million word TIME Corpus (1920s-2000s)". In *Current Change in the English Verb Phrase*, ed. Bas Aarts, et al. Cambridge: Cambridge Univ. Press. 46-67.

Hilpert, Martin. 2012. “Diachronic collostructional analysis meets the noun phrase: Studying many a noun in COHA.” In Terttu Nevalainen and Elizabeth Closs Traugott (eds.), *The Oxford handbook of the history of English*. Oxford: Oxford University Press. 233-44.

Hofland, K. & S. Johansson. 1982. *Word frequencies in British and American English*. Bergen: Norwegian Computing Centre for the Humanities/London: Longman.

Leech, Geoffrey & Roger Fallon. 1992. Computer corpora—What do they tell us about culture? *ICAME Journal* 16. 29-50.

Michel, Jean-Baptiste, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden. (2011) Quantitative Analysis of Culture Using Millions of Digitized Books. *Science* 331: 176-182.

Oakes, Michael & Malcolm Farrow. 2007. Use of the chi-square test to examine vocabulary differences in English-language corpora representing seven different countries. *Literary and Linguistic Computing* 22(1). 85-100.

Sigley, Robert & Janet Holmes. 2002. Looking at *girls* in corpora of English. *Journal of English Linguistics* 30(2). 138-157.