John Benjamins Publishing Company

SHORT PAPERS

Making Google Books n-grams useful for a wide range of research on language change

Mark Davies
Brigham Young University

The “standard” Google Books n-grams were released by Google in 2010, and they include more than 155 billion words of data for the American English data alone. Unfortunately, the standard interface is far too simplistic to allow many types of useful research on this massive dataset. In this paper, I discuss an alternative “advanced” architecture and interface for these datasets, which is freely available at googlebooks.byu.edu. This resource allows for a wide range of research on lexical, phraseological, syntactic, and semantic changes in English, in ways that would not be possible with the standard interface. With this new resource, researchers now have access to hundreds of billions of words of data, and can map out changes in English in ways that were not previously possible.

Keywords: Google Books, historical, lexical, syntactic, semantic

1. Introduction

In the 1990s a number of historical corpora of English were created, which have been the “backbone” of research on language change in Late Modern English and Present-Day English since that time. These corpora included resources such as the Brown family of corpora, ARCHER, and CONCE. These corpora have been used for many insightful studies during the past two decades, even though their small size (typically just one to four million words in size) often limited their use to just high frequency syntactic phenomena.

In 2010, the 400 million word Corpus of Historical American English (COHA) was released, which is based on 400 million words of text from the 1810s-2000s (making it at least one hundred times as big as other historical corpora of English). As has been explained elsewhere (see Davies 2012a, 2012b, forthcoming), COHA allows for research on a wide range of phenomena that are difficult or impossible to study with the small first-generation historical corpora of English.

International Journal of Corpus Linguistics 19:3 (2014), 401–416. DOI 10.1075/ijcl.19.3.04dav
ISSN 1384–6655 / E-ISSN 1569–9811 © John Benjamins Publishing Company

Coincidentally, late 2010 also saw the release of the Google Books n-grams. These n-grams are based on hundreds of *billions* of words from scanned books, which obviously makes them much larger than even COHA — about four hundred times as large as COCA and about 40,000–50,000 times as large as the small first generation historical corpora. In other words, where a small 1–2 million word corpus might have just 4–5 tokens (often too small for meaningful analysis), the Google n-grams might have 200,000–300,000 tokens. As a result, when the n-grams were released, they were released with great fanfare as a resource that would revolutionize work on historical English (and other languages), especially with regards to language change as it relates to changes in culture (see Michel et al. 2011).[1]

And yet, to this point there have been virtually no large-scale studies of changes in English based on the Google Books n-grams. As I will discuss in this paper, this is likely because the standard Google Books architecture and interface (http://books.google.com/ngrams; hereafter GB-S(tandard)) are far too simplistic to be used for research on many types of language change in English. Researcher cannot search by wildcards, they cannot meaningfully use part of speech, and they cannot use collocates in their searches. Virtually all that one can do is find the frequency of exact words and phrases over time. Because of this, in-depth studies on lexical, phraseological, syntactic, and semantic change in English with GB-S are either very difficult or impossible.

With GB-S, then, we have hundreds of billions of words worth of data — which is potentially very useful for a wide range of research — essentially "trapped" within an architecture and interface that does not allow for advanced research on language change.

2. Creating Google Books — Advanced

Fortunately, the Google Books team has made the raw data that is used for GB-S freely available for other researchers, to use with their own architectures and interfaces. In early 2012, we downloaded all of the datasets for the American English portion of Google Books — representing about 155 billion words of data.[2] The number of words per decade is as follows in Table 1 (in billions of words):

[1] Nunberg (2009, 2010) and others have been highly critical of the Google Books project from the outset, because they feel that with a dataset this size, there are bound to be too many inaccuracies in the scanned text and in the metadata. We agree that there are certainly inaccuracies, but — as we believe the data for the phenomena studied in this paper suggest — the data is of sufficient quality that it can still be used for meaningful linguistic research.

[2] In this paper, we provide data from the 155 billion words from the 1810s-2000s. There is also a very small amount of pre-1810 data, but we will not use that data in this paper. In addition,

Table 1. Size of Google Books, by decade (1810s-2000s); billions of words

1810	0.4	1910	10.1
1820	0.7	1920	7.1
1830	1.4	1930	5.8
1840	1.9	1940	6.2
1850	3.0	1950	8.1
1860	2.4	1960	13.2
1870	2.8	1970	14.0
1980	4.4	1980	15.5
1890	5.6	1990	19.8
1900	7.5	2000	26.9

The data was then imported into a relational database architecture that is similar to that of COHA and the other corpora from http://corpus.byu.edu. After the billions of rows of data were processed, the data looked like that in Table 2, which is a very small portion of the 3-grams with the initial word *started*. For each unique three word string, we see the frequency in each decade of the corpus (only every other decade is shown here, for reasons of size in this print version), as well as the "total" in the entire 155 billion word dataset. Similar tables were created for the 1-grams, 2-grams, 4-grams, and 5-grams.

Overall, there are about 730 million rows of data in the databases (as in Table 2), and these serve as the basis for all of the types of searches that we will describe in this paper. As we will see, this resource — which is now freely available at googlebooks.byu.edu — allows for a wide range of research on lexical, phraseological, syntactic, and semantic changes in English, which are available exclusively via our Google Books — Advanced site (hereafter GB-Adv), but which are not possible via GB-S. In the sections that follow, I will provide a number of concrete examples of how this data can be used to carry out research on lexical, phraseological, syntactic, and semantic change in English.

3. Lexical changes

The one thing that GB-S does well is to show the frequency of a given word or exact phrase over time, which provides useful insight into lexical shifts in the language. For example, Figure 1 shows the frequency of the word *steamship*

there are other databases, such as British English. This paper, however, is based on just the American English dataset.

Table 2. Example of n-grams databases

Word1	Word2	Word3	1810s	1830s	1850s	1870s	1890s	1910s	1930s	1950s	1970s	1990s	2000s	Total
started	to	accelerate	0	0	0	0	0	0	4	5	36	126	192	462
started	to	accommodate	0	0	0	0	0	3	3	6	5	12	18	70
started	to	accompany	0	4	15	7	16	38	7	8	17	42	51	339
started	to	accrue	0	0	0	0	0	2	0	11	3	21	25	71
started	to	adjust	0	0	0	0	0	3	9	13	22	48	107	288
started	to	advertise	0	0	0	0	0	27	13	9	21	30	67	274
started	to	affect	0	0	0	0	0	1	2	16	49	174	288	650
started	to	allow	0	0	0	0	0	5	1	1	33	83	152	337
started	to	anticipate	0	0	0	0	0	0	1	0	13	12	38	72

(notice how changes in lexical frequency is often related to cultural and societal changes).

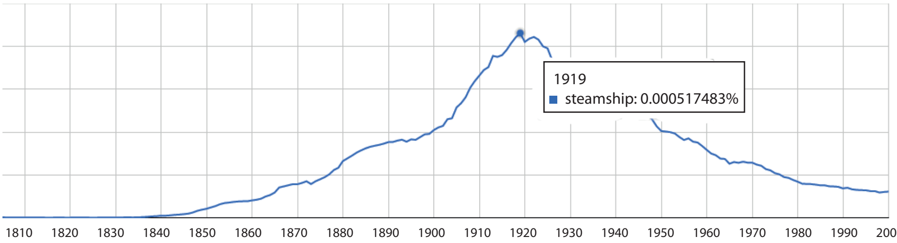
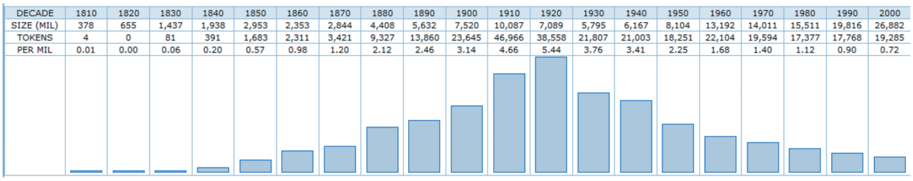


Figure 1. GB-S: frequency of *steamship*

Because the GB-Adv data is based on the same n-grams as GB-S, it will always give the same frequency as GB-S for these individual words and phrases (Figure 2):



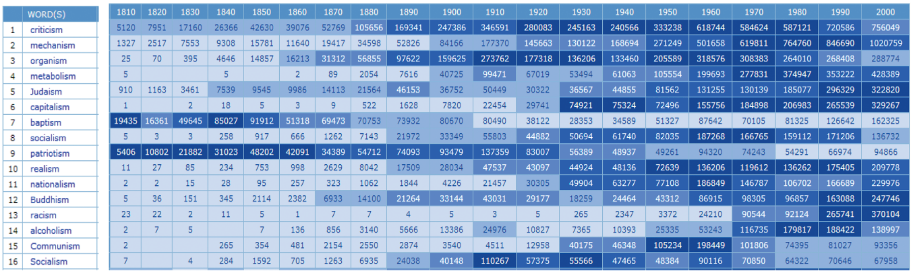
DECADE	1810	1820	1830	1840	1850	1860	1870	1880	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000
SIZE (MIL)	378	655	1,437	1,938	2,953	2,353	2,844	4,408	5,632	7,520	10,087	7,089	5,795	6,167	8,104	13,192	14,011	15,511	19,816	26,882
TOKENS	4	0	81	391	1,683	2,311	3,421	9,327	13,860	23,645	46,966	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

Figure 2. GB-Adv: frequency of *steamship*

As far as researchers being able to actually use this data, however, there is a huge difference between GB-S and GB-Adv. In the case of GB-S (Figure 1) all of the frequency data is "hidden" deep inside the "code" for the web page, and it takes some processing of this data to get it into a usable format.[3] In GB-Adv, on the other hand, the frequency data (both raw frequency [tokens] and normalized frequency per million words [per mil]) is displayed clearly in the chart (Figure 2), where it can easily be copied to a spreadsheet or database.

Although the results are similar in GB-S and GB-Adv for individual words, GB-Adv can do much more in terms of looking at lexical frequency, beyond the simplistic searches of GB-S. First, GB-Adv allows users to use wildcards to see the frequency of *all* matching words in each decade (researchers cannot search by wildcard in GB-S). For example, Figure 3 below shows the frequency of **ism* words by decade (note the increase in *criticism*, *organism*, *capitalism*, *Buddhism*, and *racism*, and the decrease in *baptism* and *patriotism*).

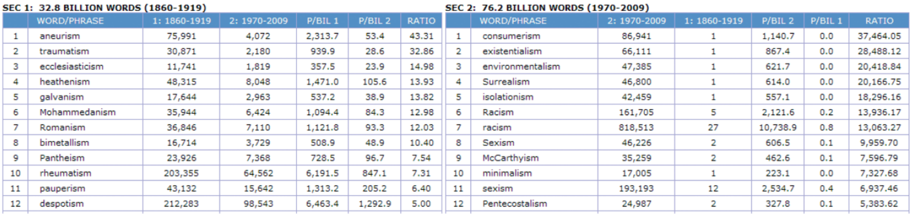
[3] And even then, this "underlying" data is available only in the format of frequency per million words, as with the 0.000517483 figure for 1919 in Figure 1. One would therefore have to convert all of these normalized figures into the actual number of tokens by creating a formula that incorporates the actual size of Google Books in words per decade, assuming that data is available at the Google Books n-grams site.



	WORD(S)	1810	1820	1830	1840	1850	1860	1870	1880	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000
1	criticism	5120	7951	17160	26366	42630	39076	52769	105656	169341	247386	346591	280083	245163	240566	333238	618744	584624	587121	720586	756049
2	mechanism	1327	2517	7553	9308	15781	11640	19417	34598	52826	84166	177370	145663	130122	168694	271249	501658	619811	764760	846690	1020759
3	organism	25	70	395	4646	14857	16213	31312	56855	97622	159625	273762	177318	136206	133460	205589	318576	308383	264010	268408	288774
4	metabolism	5			5		2	89	2054	7616	40725	99471	67019	53494	61063	105554	199693	277831	374947	353222	428389
5	Judaism	910	1163	3461	7539	9545	9986	14113	21564	46153	36752	50449	30322	36567	44855	81562	131255	130139	185077	296329	322820
6	capitalism	1		2	18	5	3	9	522	1628	7820	22454	29741	74921	75324	72496	155756	184898	206983	265539	329267
7	baptism	19435	16361	49645	85027	91912	51318	69473	70753	73932	80670	80490	38122	28353	34589	51327	87642	70105	81325	126642	162325
8	socialism	5	3	3	258	917	666	1262	7143	21972	33349	55803	44882	50694	61740	82035	187268	166765	159112	171206	136732
9	patriotism	5406	10802	21882	31023	48202	42091	34389	54712	74093	93479	137359	83007	56389	48937	49261	94320	74243	54291	66974	94866
10	realism	11	27	85	234	753	998	2629	8042	17509	28034	47537	43097	44924	48136	72639	136206	119612	136262	175405	209778
11	nationalism	2	2	15	28	95	257	323	1062	1844	4226	21457	30305	49904	63277	77108	186849	146787	106702	166689	229976
12	Buddhism	5	36	151	345	2114	2382	6933	14100	21264	33144	43031	29177	18259	24464	43312	86915	98305	96857	163088	247746
13	racism	23	22	2	11	5	1	7	7	4	5	3	5	265	2347	3372	24210	90544	92124	265741	370104
14	alcoholism	2	7	5		7	136	856	3140	5666	13386	24976	10827	7365	10393	25335	53243	116735	179817	188422	138997
15	Communism	2			265	354	481	2154	2550	2874	3540	4511	12958	40175	46348	105234	198449	101806	74395	81027	93356
16	Socialism	7		4	284	1592	705	1263	6935	24038	40148	110267	57375	55566	47465	48384	90116	70850	64322	70646	67958

Figure 3. **ism* words by decade

Second, GB-Adv can find all words that are more common in one period than another, for example **ism* words (Figure 4) that are more common in the 1860s-1910s (left) or the 1970s-2000s (right).



SEC 1: 32.8 BILLION WORDS (1860-1919)

	WORD/PHRASE	1: 1860-1919	2: 1970-2009	P/BIL 1	P/BIL 2	RATIO
1	aneurism	75,991	4,072	2,313.7	53.4	43.31
2	traumatism	30,871	2,180	939.9	28.6	32.86
3	ecclesiasticism	11,741	1,819	357.5	23.9	14.98
4	heathenism	48,315	8,048	1,471.0	105.6	13.93
5	galvanism	17,644	2,963	537.2	38.9	13.82
6	Mohammedanism	35,944	6,424	1,094.4	84.3	12.98
7	Romanism	36,846	7,110	1,121.8	93.3	12.03
8	bimetallism	16,714	3,729	508.9	48.9	10.40
9	Pantheism	23,926	7,368	728.5	96.7	7.54
10	rheumatism	203,355	64,562	6,191.5	847.1	7.31
11	pauperism	43,132	15,642	1,313.2	205.2	6.40
12	despotism	212,283	98,543	6,463.4	1,292.9	5.00

SEC 2: 76.2 BILLION WORDS (1970-2009)

	WORD/PHRASE	2: 1970-2009	1: 1860-1919	P/BIL 2	P/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

Figure 4. Comparison of **ism* words, 1860s–1910s vs 1970s–2000s

In essence, then, GB-Adv allows us to find all words that have appeared or disappeared between different time periods (or which have increased or decreased greatly in frequency between these time periods) — even when we do not know ahead of time what these words are. In GB-S, on the other hand, we can only get frequency information on specific, already-determined words.

4. Changes in phraseology

As with words, phrase-based searches are very simplistic in GB-S. Again, one can only search for exact phrases, such as *as though to* (decreasing since about the 1960s) and *a lot of* (increasing since the mid 1850s, but especially since about the 1960s), as seen in Figure 5.

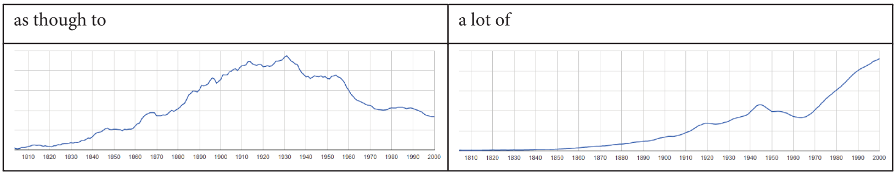


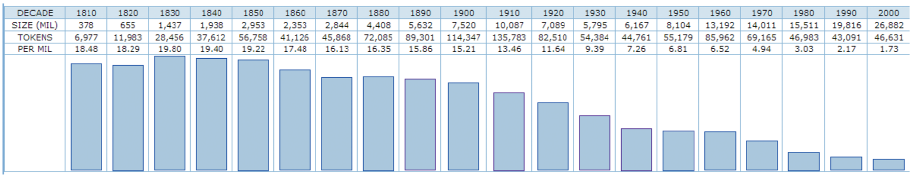
Figure 5. GB-S: Frequency of exact phrases: "*as though to*" and "*a lot of*"

But GB-S is unable to deal with phrases like those in Table 3, which have a variable "slot" for a given part of speech:[4]

Table 3. Examples of "variable slot" phrases

Phrase	**Examples**
a most ADJ NOUN	a most important part, a most difficult task
many a NOUN	many a time, many a night
have quite V-ed	had quite forgotten, has quite changed
NOUN [be] that of a	effect is that of a, role was that of a

In GB-Adv, on the other hand, it is possible to search for phrases that include part of speech, and then to see a list of the most frequent matching strings, and then see each of these matching strings in context.[5] For example, Figure 6 shows the overall frequency of "*a most* ADJ NOUN":



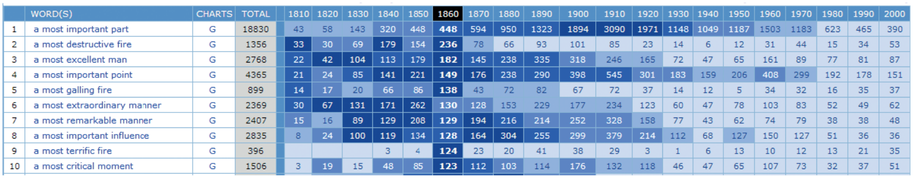
DECADE	1810	1820	1830	1840	1850	1860	1870	1880	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000
SIZE (MIL)	378	655	1,437	1,938	2,953	2,353	2,844	4,408	5,632	7,520	10,087	7,089	5,795	6,167	8,104	13,192	14,011	15,511	19,816	26,882
TOKENS	6,977	11,983	28,456	37,612	56,758	41,126	45,868	72,085	89,301	114,347	135,783	82,510	54,384	44,761	55,179	85,962	69,165	46,983	43,091	46,631
PER MIL	18.48	18.29	19.80	19.40	19.22	17.48	16.13	16.35	15.86	15.21	13.46	11.64	9.39	7.26	6.81	6.52	4.94	3.03	2.17	1.73

Figure 6. Overall frequency of the phrase "*a most* ADJ NOUN"

Users can click on any decade to see the most frequent strings for that particular decade (or they could also see a "Table format" display with the most frequent strings for all decades). For example, users could click on the [1860] decade in Figure 6 to see Figure 7, which lists the most frequent matching phrases in that decade (*a most important part*, *a most destructive fire*, etc.):

[4] Google Books recently released a new version of the n-grams, which they claimed allows users to search for phrases like "*many a* NOUN", by using "part of speech" placeholders (see https://books.google.com/ngrams/info). However, such searches are almost meaningless, since it is impossible to (i) see the "matching strings" for such a search (e.g. *many a time*, *many a day*), or (ii) to see the "Word in Context" display for such searches, since all links to the "Word in Context" display mysteriously disappear when a part of speech code has been used in the search.

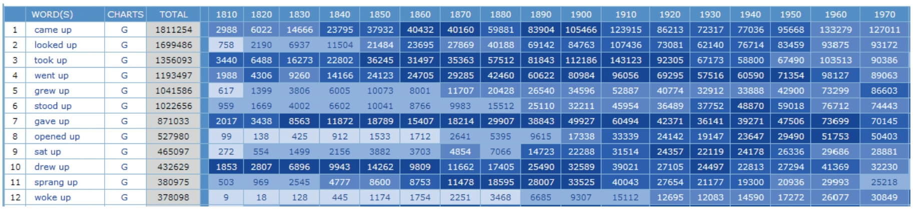
[5] To search by part of speech, GB-Adv uses frequency data from COHA (the 400 million word Corpus of Historical American English) to see what word forms are tagged with a given part of speech, and then it uses this information as part of the GB-Adv search. For example, to search for "a most ADJ NOUN", it sees which words in COHA are tagged as ADJ or NOUN at least 50% of the time, and then uses this list of words as part of the GB-Adv search. If too many incorrect word forms are found, users can simply include a code in the search string to set the figure to be more restrictive, such as 80% or 90%. Or to see more forms (but with potentially more incorrect forms) they could lower the accuracy to 30% or even 10%.



	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	a most important part	G	18830	43	58	143	320	448	448	594	950	1323	1894	3090	1971	1148	1049	1187	1503	1183	623	465	390
2	a most destructive fire	G	1356	33	30	69	179	154	236	78	66	93	101	85	23	14	6	12	31	44	15	34	53
3	a most excellent man	G	2768	22	42	104	113	179	182	145	238	335	318	246	165	72	47	65	161	89	77	81	87
4	a most important point	G	4365	21	24	85	141	221	149	176	238	290	398	545	301	183	159	206	408	299	192	178	151
5	a most galling fire	G	899	14	17	20	66	86	138	43	72	82	67	72	37	14	12	5	34	32	16	35	37
6	a most extraordinary manner	G	2369	30	67	131	171	262	130	128	153	229	177	234	123	60	47	78	103	83	52	49	62
7	a most remarkable manner	G	2407	15	16	89	129	208	129	194	216	214	252	328	158	77	43	62	74	79	38	38	48
8	a most important influence	G	2835	8	24	100	119	134	128	164	304	255	299	379	214	112	68	127	150	127	51	36	36
9	a most terrific fire	G	396				3	4	124	23	20	41	38	29	3	1	6	13	10	12	13	21	35
10	a most critical moment	G	1506	3	19	15	48	85	123	112	103	114	176	132	118	46	47	65	107	73	32	37	51

Figure 7. Forms of the phrase "*a most* ADJ NOUN" by decade

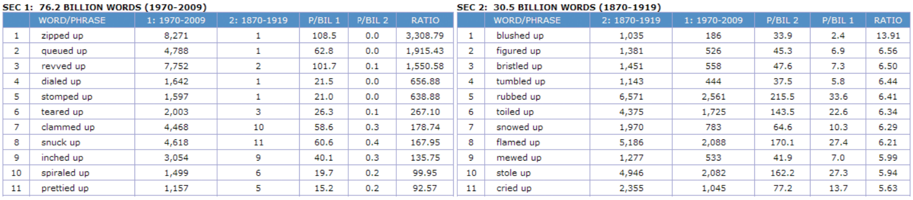
Another example of changes in phraseology might be phrasal verbs, such as phrasal verbs with *up* (*make up*, *show up*, *look up*, etc.). As with the previous examples, with GB-Adv we can see the frequency of each matching string in each decade (Figure 8):



	WORD(S)	CHARTS	TOTAL	1810	1820	1830	1840	1850	1860	1870	1880	1890	1900	1910	1920	1930	1940	1950	1960	1970
1	came up	G	1811254	2988	6022	14666	23795	37932	40432	40160	59881	83904	105466	123915	86213	72317	77036	95668	133279	127011
2	looked up	G	1699486	758	2190	6937	11504	21484	23695	27869	40188	69142	84763	107436	73081	62140	76714	83459	93875	93172
3	took up	G	1356093	3440	6488	16273	22802	36245	31497	35363	57512	81843	112186	143123	92305	67173	58800	67490	103513	90386
4	went up	G	1193497	1988	4306	9260	14166	24123	24705	29285	42460	60622	80984	96056	69295	57516	60590	71354	98127	89063
5	grew up	G	1041586	617	1399	3806	6005	10073	8001	11707	20428	26540	34596	52887	40774	32912	33888	42900	73299	86603
6	stood up	G	1022656	959	1669	4002	6602	10041	8766	9983	15512	25110	32211	45954	36489	37752	48870	59018	76712	74443
7	gave up	G	871033	2017	3438	8563	11872	18789	15407	18214	29907	38843	49927	60494	42371	36141	39271	47506	73699	70145
8	opened up	G	527980	99	138	425	912	1533	1712	2641	5395	9615	17338	33339	24142	19147	23647	29490	51753	50403
9	sat up	G	465097	272	554	1499	2156	3882	3703	4854	7066	14723	22288	31514	24357	22119	24178	26336	29686	28881
10	drew up	G	432629	1853	2807	6896	9943	14262	9809	11662	17405	25490	32589	39021	27105	24497	22813	27294	41369	32230
11	sprang up	G	380975	503	969	2545	4777	8600	8753	11478	18595	28007	33525	40043	27654	21177	19300	20936	29993	25218
12	woke up	G	378098	9	18	128	445	1174	1754	2251	3468	6685	9307	15112	12695	12083	14590	17272	26077	30849

Figure 8. Forms of the phrase "VERB *up*" by decade

But remember that we can also compare the results from one historical period against another. This allows us to see which phrasal verbs with *up* are more frequent in one period than another (in just the past tense, in this example). For example, Figure 9 lists phrasal verbs with *up* that are more common in the 1970s-2000s (on the left: *zipped up*, *revved up*, *teared up*, etc.) compared to the 1870s-1910s (on the right; most of these sound quite old-fashioned now, e.g. *blushed up*, *figured up*, *bristled up*).



SEC 1: 76.2 BILLION WORDS (1970-2009)

	WORD/PHRASE	1: 1970-2009	2: 1870-1919	P/BIL 1	P/BIL 2	RATIO
1	zipped up	8,271	1	108.5	0.0	3,308.79
2	queued up	4,788	1	62.8	0.0	1,915.43
3	revved up	7,752	2	101.7	0.1	1,550.58
4	dialed up	1,642	1	21.5	0.0	656.88
5	stomped up	1,597	1	21.0	0.0	638.88
6	teared up	2,003	3	26.3	0.1	267.10
7	clammed up	4,468	10	58.6	0.3	178.74
8	snuck up	4,618	11	60.6	0.4	167.95
9	inched up	3,054	9	40.1	0.3	135.75
10	spiraled up	1,499	6	19.7	0.2	99.95
11	prettied up	1,157	5	15.2	0.2	92.57

SEC 2: 30.5 BILLION WORDS (1870-1919)

	WORD/PHRASE	2: 1870-1919	1: 1970-2009	P/BIL 2	P/BIL 1	RATIO
1	blushed up	1,035	186	33.9	2.4	13.91
2	figured up	1,381	526	45.3	6.9	6.56
3	bristled up	1,451	558	47.6	7.3	6.50
4	tumbled up	1,143	444	37.5	5.8	6.44
5	rubbed up	6,571	2,561	215.5	33.6	6.41
6	toiled up	4,375	1,725	143.5	22.6	6.34
7	snowed up	1,970	783	64.6	10.3	6.29
8	flamed up	5,186	2,088	170.1	27.4	6.21
9	mewed up	1,277	533	41.9	7.0	5.99
10	stole up	4,946	2,082	162.2	27.3	5.94
11	cried up	2,355	1,045	77.2	13.7	5.63

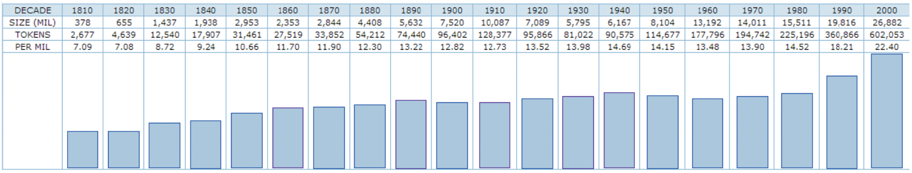
Figure 9. Comparison of "VERB *up*" in the 1970s–2000 and the 1870s–1910s

Although this search may seem simple, it would be completely impossible in GB-S, where it is impossible to (i) search for "variable" phrases such as this with part of speech, (ii) display the most frequent matching forms, or (iii) compare between different historical periods. But in GB-Adv, it takes just 2–3 seconds to compare all cases of "VERB *up*" in the two historical periods, and thus compare phrases over time.

5. Syntactic change

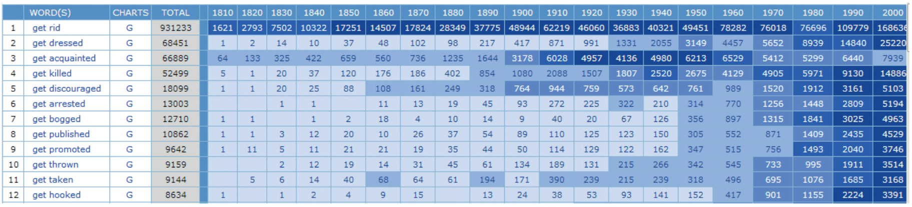
As we have seen, GB-Adv allows us to search for phrases in some fairly advanced ways. It should come as no surprise, then, that GB-Adv also allows researchers to gather data on historical changes in syntax in ways that could never be done with GB-S.

Let us briefly consider two quick examples of how GB-Adv can search through the billions of words of data to provide information on other syntactic changes in American English. Figure 10 and Figure 11 provide data on the "*get* passive" construction (e.g. *got returned*, *get fired*; search = [get] [vvn*]).



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	2,677	4,639	12,540	17,907	31,461	27,519	33,852	54,212	74,440	96,402	128,377	95,866	81,022	90,575	114,677	177,796	194,742	225,196	360,866	602,053
PER MIL	7.09	7.08	8.72	9.24	10.66	11.70	11.90	12.30	13.22	12.82	12.73	13.52	13.98	14.69	14.15	13.48	13.90	14.52	18.21	22.40

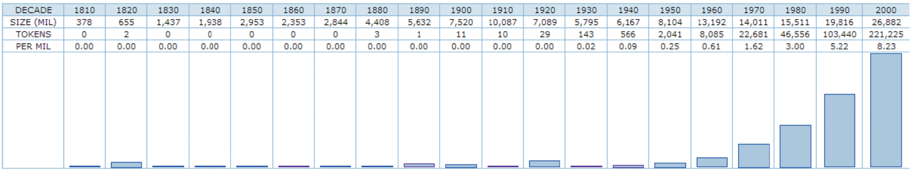
Figure 10. Overall frequency of the construction "*get* V-ed"



	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	get rid	G	931233	1621	2793	7502	10322	17251	14507	17824	28349	37775	48944	62219	46060	36883	40321	49451	78282	76018	76696	109779	168636
2	get dressed	G	68451	1	2	14	10	37	48	102	98	217	417	871	991	1331	2055	3149	4457	5652	8939	14840	25220
3	get acquainted	G	66889	64	133	325	422	659	560	736	1235	1644	3178	6028	4957	4136	4980	6213	6529	5412	5299	6440	7939
4	get killed	G	52499	5	1	20	37	120	176	186	402	854	1080	2088	1507	1807	2520	2675	4129	4905	5971	9130	14886
5	get discouraged	G	18099	1	1	20	25	88	108	161	249	318	764	944	759	573	642	761	989	1520	1912	3161	5103
6	get arrested	G	13003			1	1		11	13	19	45	93	272	225	322	210	314	770	1256	1448	2809	5194
7	get bogged	G	12710	1	1		1	2	18	4	10	14	9	40	20	67	126	356	897	1315	1841	3025	4963
8	get published	G	10862	1	1	3	12	20	10	26	37	54	89	110	125	123	150	305	552	871	1409	2435	4529
9	get promoted	G	9642	1	11	5	11	21	21	19	35	44	50	114	129	122	162	347	515	756	1493	2040	3746
10	get thrown	G	9159			2	12	19	14	31	45	61	134	189	131	215	266	342	545	733	995	1911	3514
11	get taken	G	9144		5	6	14	40	68	64	61	194	171	390	239	215	239	318	496	695	1076	1685	3168
12	get hooked	G	8634	1		1	2	4	9	15		13	24	38	53	93	141	152	417	901	1155	2224	3391

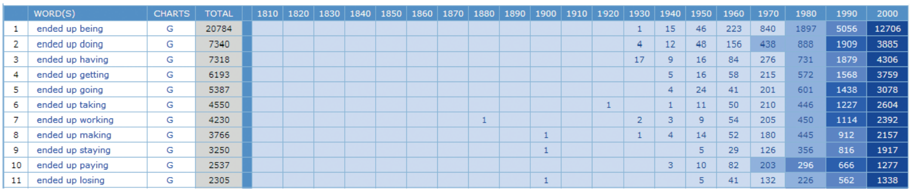
Figure 11. Forms of the construction "*get* V-ed" by decade

Figure 12 and Figure 13, on the other hand, provide data on the construction "*end up* V-ing" (here limited to just the form *ended*, e.g. *ended up paying*; search = [end] up [v?g*]). As can be seen, both this construction and the "*get* passive" are increasing over time. In addition, "*end up* V-ing" is a relatively new construction, and has only been used since about the 1920s.



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	0	2	0	0	0	0	0	3	1	11	10	29	143	566	2,041	8,085	22,681	46,556	103,440	221,225
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	0.02	0.09	0.25	0.61	1.62	3.00	5.22	8.23

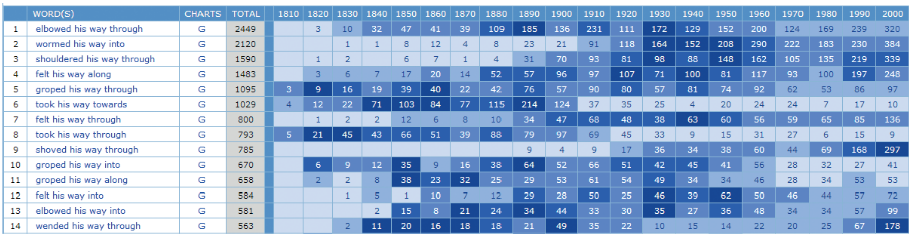
Figure 12. Overall frequency of the construction "*end up* V-ing"



	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	ended up being	G	20784													1	15	46	223	840	1897	5056	12706
2	ended up doing	G	7340													4	12	48	156	438	888	1909	3885
3	ended up having	G	7318													17	9	16	84	276	731	1879	4306
4	ended up getting	G	6193														5	16	58	215	572	1568	3759
5	ended up going	G	5387														4	24	41	201	601	1438	3078
6	ended up taking	G	4550												1		1	11	50	210	446	1227	2604
7	ended up working	G	4230								1					2	3	9	54	205	450	1114	2392
8	ended up making	G	3766										1			1	4	14	52	180	445	912	2157
9	ended up staying	G	3250										1					5	29	126	356	816	1917
10	ended up paying	G	2537														3	10	82	203	296	666	1277
11	ended up losing	G	2305										1					5	41	132	226	562	1338

Figure 13. Forms of the construction "*end up* V-ing" by decade

To take a somewhat more complex construction, consider the "*way* construction", which has been the focus of a great deal of research in construction grammar (Figure 14). In GB-Adv we can simply search for "[vv*] [ap*] way [i*]" to find more than 1,083,000 tokens for 3,000 unique strings like *find their way into*, *make his way through*, *groping their way into*, and so on. If desired, we could also compare the verbs (*feel*, *shove*, *grope*, *elbow*, etc.) that are used in different periods, to see the influence of semantic factors over time.



	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	elbowed his way through	G	2449		3	10	32	47	41	39	109	185	136	231	111	172	129	152	200	124	169	239	320
2	wormed his way into	G	2120			1	1	8	12	4	8	23	21	91	118	164	152	208	290	222	183	230	384
3	shouldered his way through	G	1590		1	2		6	7	1	4	31	70	93	81	98	88	148	162	105	135	219	339
4	felt his way along	G	1483		3	6	7	17	20	14	52	57	96	97	107	71	100	81	117	93	100	197	248
5	groped his way through	G	1095	3	9	16	19	39	40	22	42	76	57	90	80	57	81	74	92	62	53	86	97
6	took his way towards	G	1029	4	12	22	71	103	84	77	115	214	124	37	35	25	4	20	24	24	7	17	10
7	felt his way through	G	800		1	2	2	12	6	8	10	34	47	68	48	38	63	60	56	59	65	85	136
8	took his way through	G	793	5	21	45	43	66	51	39	88	79	97	69	45	33	9	15	31	27	6	15	9
9	shoved his way through	G	785									9	4	9	17	36	34	38	60	44	69	168	297
10	groped his way into	G	670		6	9	12	35	9	16	38	64	52	66	51	42	45	41	56	28	32	27	41
11	groped his way along	G	658		2	2	8	38	23	32	25	29	53	61	54	49	34	34	46	28	34	53	53
12	felt his way into	G	584			1	5	1	10	7	12	29	28	50	25	46	39	62	50	46	44	57	72
13	elbowed his way into	G	581				2	15	8	21	24	34	44	33	30	35	27	36	48	34	34	57	99
14	wended his way through	G	563			2	11	20	16	18	18	21	49	35	22	10	15	14	22	20	25	67	178

Figure 14. Forms of the construction "V-ed *his way* PREP" by decade

In the three examples above, we searched for just one particular string (such as "[end] *up* [vvg*]" or "[vv*] [ap*] *way* [i*]") and then retrieved the overall frequency (e.g. Figure 12) or the frequency of each matching string (e.g. Figure 13). But it is also possible to carry out more advanced research as well. For example, we could compare the frequency of two competing constructions to see how one construction is increasing at the expense of the other.

For example, consider the two competing options deals with the complements of verbs such as *start* and *begin*, which can take either [to V] or [V-ing]: *he started* [*to walk*/*walking*] *down the street*. As many researchers have shown, there has been a "Great Complement Shift" underway (analogous in some ways to the Great Vowel Shift) since at least the 1800s, in which [V-ing] has been increasing at the expense of [to V] (for the historical development of this construction, based on much smaller corpora, see for example de Smet 2008).

The GB-Adv data (Table 4) shows this change in complement structures quite nicely, via four simple searches: [to V] complements with both *start* and *begin*, as well as [V-ing] complements with both verbs. The 26,125,000 tokens show that [V-ing] is increasing with both verbs over time (note parenthetically that the

Table 4. *start/begin* [to V] vs. [V-ing]

	1820	1840	1860	1880	1900	1920	1940	1960	1980	2000
start										
to V	128	1,128	2,900	8,241	32,348	53,315	85,121	163,112	273,197	865,672
V-ing	41	56	64	263	2,523	16,154	58,647	136,785	275,322	968,535
% V-ing	0.24	0.05	0.02	0.03	0.07	0.23	0.41	0.46	0.50	0.53
begin										
to V	58,937	193,017	242,340	473,600	902,423	880,834	795,967	1,616,536	1,853,586	3,118,572
V-ing	903	3,469	6,758	18,684	46,851	68,029	92,539	192,828	347,207	864,433
% V-ing	0.02	0.02	0.03	0.04	0.05	0.07	0.10	0.11	0.16	0.22

number of tokens for this one construction — more than 26 million tokens — is 10–20 times as large as the entire size of many historical corpora of English!).

Whereas the two verbs had more or less the same degree of [V-ing] in the late 1800s, *start* began moving towards [V-ing] in the early 1900s much more than *begin*. As Figure 15 indicates, while the rate of change has slowed somewhat in the last 50–60 years, there is still a large difference between the two verbs — *start* takes [V-ing] complements at more than twice the rate of *begin*.

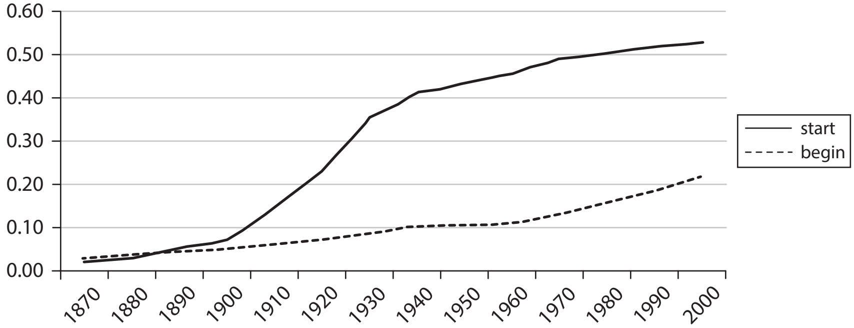


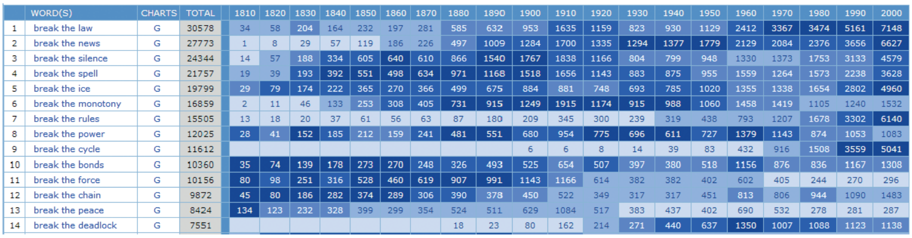
Figure 15. Percentage of clauses with [V-ing] (vs. [to V])

The only real option for researching this construction in GB-S would be to search for each possible combination of a form of *start* or *begin*, followed by *to* + thousands of individual verbs (e.g. *started to notice*, *beginning to consider*). We would also need to search for a form of one of these two verbs followed by the [V-ing] form of thousands of individual verbs (e.g. *starts talking*, *began eating*). Obviously, such a solution would take hundreds of hours. With GB-Adv, on the other hand, we can get the data for millions of tokens in just 1–2 minutes.

6. Semantic changes and changes in discourse

We can tell a great deal about the meaning of a word by the other words with which it co-occurs. As Firth (1957: 11) noted, "you shall know a word by the company it keeps". Unfortunately, with GB-S, there is no way to look at collocates. We cannot enter a word into the search interface and then find the most frequently co-occurring words. All we can do is see the frequency of the word in isolation (as in Figure 1), which is of little or no value in terms of looking at meaning.

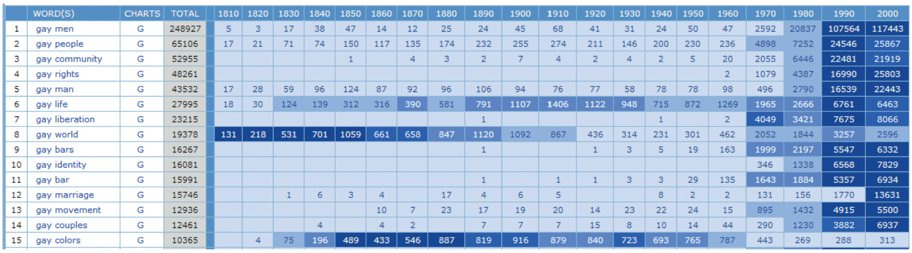
GB-Adv, on the other hand, can easily find the collocates of a given word. For example, Figure 16 shows the most common nouns occurring after *break the*, by decade.[6] Note in Figure 16 the increase in the collocates *law*, *cycle*, and *deadlock*, and the decrease in the collocates *spell*, *bonds*, *force*, and *peace*.



	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	break the law	G	30578	34	58	204	164	232	197	281	585	632	953	1635	1159	823	930	1129	2412	3367	3474	5161	7148
2	break the news	G	27773	1	8	29	57	119	186	226	497	1009	1284	1700	1335	1294	1377	1779	2129	2084	2376	3656	6627
3	break the silence	G	24344	14	57	188	334	605	640	610	866	1540	1767	1838	1166	804	799	948	1330	1373	1753	3133	4579
4	break the spell	G	21757	19	39	193	392	551	498	634	971	1168	1518	1656	1143	883	875	955	1559	1264	1573	2238	3628
5	break the ice	G	19799	29	79	174	222	365	270	366	499	675	884	881	748	693	785	1020	1355	1338	1654	2802	4960
6	break the monotony	G	16859	2	11	46	133	253	308	405	731	915	1249	1915	1174	915	988	1060	1458	1419	1105	1240	1532
7	break the rules	G	15505	13	18	20	37	61	56	63	87	180	209	345	300	239	319	438	793	1207	1678	3302	6140
8	break the power	G	12025	28	41	152	185	212	159	241	481	551	680	954	775	696	611	727	1379	1143	874	1053	1083
9	break the cycle	G	11612										6	6	8	14	39	83	432	916	1508	3559	5041
10	break the bonds	G	10360	35	74	139	178	273	270	248	326	493	525	654	507	397	380	518	1156	876	836	1167	1308
11	break the force	G	10156	80	98	251	316	528	460	619	907	991	1143	1166	614	382	382	402	602	405	244	270	296
12	break the chain	G	9872	45	80	186	282	374	289	306	390	378	450	522	349	317	317	451	813	806	944	1090	1483
13	break the peace	G	8424	134	123	232	328	399	299	354	524	511	629	1084	517	383	437	402	690	532	278	281	287
14	break the deadlock	G	7551								18	23	80	162	214	271	440	637	1350	1007	1088	1123	1138

Figure 16. "*break* + *the* + NOUN", by decade

Assuming we have a large enough corpus (and 155 billion words certainly fits this definition), we can look at collocates over time, and see how changing collocates may serve as indicators of changes in meaning. For example, Figure 17 shows the collocates of *gay* in each decade since the early 1800s. Notice in Figure 17 the decrease in words like *world* and *colors*, and the increase in words like *rights*, *liberation*, and *identity* — all of which provide good data for the change in meaning of this word.

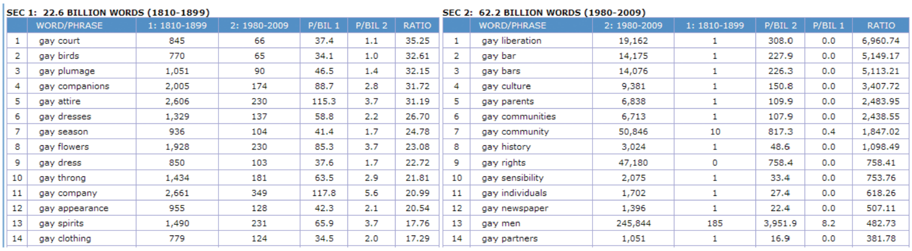


	WORD(S)	CHARTS	TOTAL	1810	1820	1830	1840	1850	1860	1870	1880	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000
1	gay men	G	248927	5	3	17	38	47	14	12	25	24	45	68	41	31	24	50	47	2592	20837	107564	117443
2	gay people	G	65106	17	21	71	74	150	117	135	174	232	255	274	211	146	200	230	236	4898	7252	24546	25867
3	gay community	G	52955					1		4	3	2	7	4	2	4	2	5	20	2055	6446	22481	21919
4	gay rights	G	48261																2	1079	4387	16990	25803
5	gay man	G	43532	17	28	59	96	124	87	92	96	106	94	76	77	58	78	78	98	496	2790	16539	22443
6	gay life	G	27995	18	30	124	139	312	316	390	581	791	1107	1406	1122	948	715	872	1269	1965	2666	6761	6463
7	gay liberation	G	23215									1					1		2	4049	3421	7675	8066
8	gay world	G	19378	131	218	531	701	1059	661	658	847	1120	1092	867	436	314	231	301	462	2052	1844	3257	2596
9	gay bars	G	16267									1			1	3	5	19	163	1999	2197	5547	6332
10	gay identity	G	16081																	346	1338	6568	7829
11	gay bar	G	15991									1		1	1	3	3	29	135	1643	1884	5357	6934
12	gay marriage	G	15746			1	6	3	4		17	4	6	5			8	2	2	131	156	1770	13631
13	gay movement	G	12936						10	7	23	17	19	20	14	23	22	24	15	895	1432	4915	5500
14	gay couples	G	12461				4		4	2		7	7	7	15	8	10	14	44	290	1230	3882	6937
15	gay colors	G	10365		4	75	196	489	433	546	887	819	916	879	840	723	693	765	787	443	269	288	313

Figure 17. "*gay* + NOUN", by decade

Finally, we can compare all of the collocates in different time periods. In Figure 18, we see that in the 1800s (left), there are collocates like *birds*, *dresses*, *flowers*, *spirits*, and *clothing*, whereas in the 1980s-2000s (right) there are collocates like *liberation*, *bar*, *history*, *community*, and *rights*.

[6] A significant limitation of the Google Books n-gram data (both at GB-S and GB-Adv) is that only those n-grams that occur 40 times or more are included in the n-grams datasets. For long phrases (e.g. 4-grams or 5-grams) with many possible words in multiple "slots", this is a serious limitation.



SEC 1: 22.6 BILLION WORDS (1810-1899)

	WORD/PHRASE	1: 1810-1899	2: 1980-2009	P/BIL 1	P/BIL 2	RATIO
1	gay court	845	66	37.4	1.1	35.25
2	gay birds	770	65	34.1	1.0	32.61
3	gay plumage	1,051	90	46.5	1.4	32.15
4	gay companions	2,005	174	88.7	2.8	31.72
5	gay attire	2,606	230	115.3	3.7	31.19
6	gay dresses	1,329	137	58.8	2.2	26.70
7	gay season	936	104	41.4	1.7	24.78
8	gay flowers	1,928	230	85.3	3.7	23.08
9	gay dress	850	103	37.6	1.7	22.72
10	gay throng	1,434	181	63.5	2.9	21.81
11	gay company	2,661	349	117.8	5.6	20.99
12	gay appearance	955	128	42.3	2.1	20.54
13	gay spirits	1,490	231	65.9	3.7	17.76
14	gay clothing	779	124	34.5	2.0	17.29

SEC 2: 62.2 BILLION WORDS (1980-2009)

	WORD/PHRASE	2: 1980-2009	1: 1810-1899	P/BIL 2	P/BIL 1	RATIO
1	gay liberation	19,162	1	308.0	0.0	6,960.74
2	gay bar	14,175	1	227.9	0.0	5,149.17
3	gay bars	14,076	1	226.3	0.0	5,113.21
4	gay culture	9,381	1	150.8	0.0	3,407.72
5	gay parents	6,838	1	109.9	0.0	2,483.95
6	gay communities	6,713	1	107.9	0.0	2,438.55
7	gay community	50,846	10	817.3	0.4	1,847.02
8	gay history	3,024	1	48.6	0.0	1,098.49
9	gay rights	47,180	0	758.4	0.0	758.41
10	gay sensibility	2,075	1	33.4	0.0	753.76
11	gay individuals	1,702	1	27.4	0.0	618.26
12	gay newspaper	1,396	1	22.4	0.0	507.11
13	gay men	245,844	185	3,951.9	8.2	482.73
14	gay partners	1,051	1	16.9	0.0	381.78

Figure 18. "*gay* + NOUN", 1800s vs 1980s–2000s

In addition to semantic change, however, we can also examine changes in collocates to look for evidence of changes in discourse — *what* we are saying about a particular topic over time — and each of these (as seen in Table 5) provides interesting insight into cultural and societal changes in the United States during the past 200 years.

Table 5. Culture: changing collocates over time

	Older period	More recent period
women	1930s-1950s: ridiculous, plump, loveliest, restless, agreeable	1960s-1980s: battered, militant, college-educated, liberated
art	1830s-1910s: noble, classic, Grecian	1960s-2000: abstract, Asian, African, commercial
fast	1850s-1910s: mail, train, horses, steamers	1960s-2000s: food, track, lane, buck
music	1850s-1910s: delightful, exquisite, sweeter, tender	1970s-2000s: Western, Black, electronic, recorded
food	1850s-1910s: spiritual, insufficient, unwholesome, mental	1970s-2000s: fast, Chinese, Mexican, organic

The insight into changes in culture that we gain from looking at collocates is unique to GB-Adv. With GB-S, all we can do is look at the frequency of the words *women, art*, *fast*, *music*, and *food* themselves (as in Figure 19), which is not overly insightful.

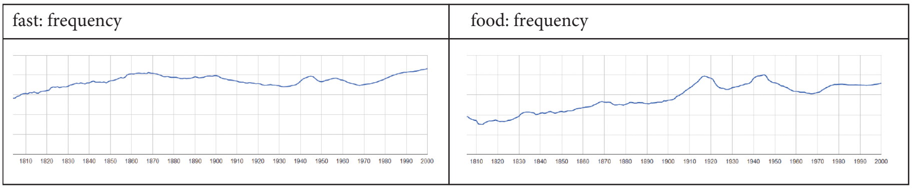


Figure 19. GB-S: Frequency of *fast* and *food*

In spite of all of the fanfare in the past few years about the potential in using Google Books data to gain insight into "culturomics" (cf. Michel et al. 2011), with GB-S we are often left just with simplistic charts like those above, showing the frequency of a single word itself. In order to gain the best insight into cultural shifts, we have to examine *what is being said* about a particular topic, and this is only possible via collocates in GB-Adv.

7. Conclusion

As we have seen, Google Books — Advanced (GB-Adv) allows us to gain insight into many linguistic changes in 155 billion words of American English in ways that are quite impossible with Google Books — Standard (GB-S).

Nevertheless, we should still keep in mind several limitations of the data in GB-Adv. First, as we have mentioned, a serious limitation is that only those words and strings that occur at least 40 times in the 155 billion words of data are included in the n-grams. Second, while the "on-the-fly" part of speech "tagging" occurs quite well, it is not the same as having a corpus that has been contextually tagged, word by word. Third, Google has limited the n-grams to 5-grams and less; therefore it is impossible to search for a string longer than five words. Fourth, while collocates work quite well (see Section 3 and Section 6), they are limited to a word and perhaps two words on each side (in the case of a 5-gram), which is often more narrow than we would like.

Parenthetically, for researchers who find these limitations to be overly restrictive for research on a particular phenomenon, it may make sense to use the freely-available, 400 million word Corpus of Historical American English (COHA), which has none of these limitations (but is of course much smaller than Google Books).

In summary, there are definitely significant limitations of the Google Books (Standard) interface, which only allows the simplest of searches. But the fact that Google has graciously made the n-grams data freely available to others to use in alternate architecture and interfaces (as we have done with Google Books — Advanced) means that researchers now have access to immense amounts of data (155 billion words) via a powerful architecture and interface, which will allow them to research a wide range of linguistic changes in English.

References

Davies, M. (2012a). "Expanding horizons in historical linguistics with the 400 million word Corpus of Historical American English". *Corpora*, 7 (2), 121–57. DOI: 10.3366/cor.2012.0024.

Davies, M. (2012b). "Examining recent changes in English: Some methodological issues". In T. Nevalainen & E. C. Traugott (Eds.), *The Oxford Handbook of the History of English*. Oxford: Oxford University Press, 263–87.

Davies, M. (forthcoming). "A corpus-based study of lexical developments in Early and Late Modern English". In M. Kytö & P. Pahta (Eds.), *Handbook of English Historical Linguistics*. Cambridge: Cambridge University Press.

de Smet, H. 2008. *Diffusional Change in the English System of Complementation: Gerunds, Participles and for...to-infinitives*. Unpublished Ph.D. dissertation. University of Leuven, Belgium.

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

Michel, J. B., Kui Shen, Y., Presser Aiden, A., Veres, A., Gray, M., The Google Books Team, Pickett, J., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M. & Lieberman Aiden, E. 2011. "Quantitative analysis of culture using millions of digitized books". *Science*, 331, 176–182. DOI: 10.1126/science.1199644

Nunberg, G. 2009. "Google's Book Search: A disaster for scholars". *The Chronicle of Higher Education*, August 31, 2009. Available at: http://chronicle.com/article/Googles-Book-Search-A/48245/ (accessed March 2014).

Nunberg, G. 2010. "Counting on Google Books". *The Chronicle of Higher Education*. December 16, 2010. Available at: https://chronicle.com/article/Counting-on-Google-Books/125735/ (accessed March 2014).

Author's address

Mark Davies
Department of Linguistics and English Language
Brigham Young University
4071 JFSB
Provo, UT 84602
USA

mark_davies@byu.edu