Mark Davies and Don Chapman

The effect of representativeness and size in historical corpora: An empirical study of changes in lexical frequency

1 Introduction

There are at least two basic considerations in the creation and use of a corpus, especially historical corpora. First, the corpus should be large enough to accurately reflect what was really happening in the language at a particular point in time. It would probably not make sense to create a corpus from just ten or twenty texts – maybe 20,000 to 40,000 words of text in total – and expect such a small corpus to accurately reflect the entirety of the language.

Secondly, the texts will ideally be "representative" of the entire universe of texts for a particular period – and hopefully even the actual language of that period, including informal genres like spoken language (see Biber 1990, 1993; Leech 2006). If we create a corpus that is composed strictly of newspapers, for example, then we may find out a great deal about the language of newspapers over time, but this may have very little to do with other genres, or the language as a whole.

In this paper, we will consider both the issue of size and representativeness by looking at (primarily) lexical data from three historical corpora. We will first examine the issue of size. Can a relatively small but well-designed corpus yield data on lexical frequency that is similar to that of a much larger corpus? Second, we will compare the lexical data from the two large corpora. Do they yield similar lexical frequency data, in spite of the fact that one corpus is designed to be representative while the other is not? The answers to these questions may be surprising to some researchers.

2 The corpora

In this study, we will compare the following three corpora:

– The Brown family of corpora: A Standard Corpus of Present-Day Edited American English (Brown), The Lancaster-Olso/Bergen Corpus (LOB), The Freiburg-Brown Corpus (Frown), and The Freiburg-LOB Corpus of British English (FLOB). These corpora were designed to be very representative of

(Unicode 9 24/5/16 11:21) WDG-New (155mm×230mm) DGMetaSerifScience (OpenType) pp. 131–150 1732 Chapman_06_Davies (p. 131)

the language as a whole, but they are rather small in size. The four corpora are Brown (American texts from 1961), LOB (British, 1961), Frown (American, 1991), and FLOB (British, 1991). Each of the four corpora contains one million words of text from 500 different texts (2,000 words each), with essentially the same genres and domains – one half of which is "imaginative" (= fiction) and the other half of which is "informational" (= nonfiction).

– The *Corpus of Historical American English* (COHA; corpus.byu.edu/coha). COHA was released in 2010, and it was designed to be both large and representative (see Davies 2012a, 2012b). It contains four hundred million words of texts in more than 100,000 different texts from the 1810s to the 2000s, including at least 10 million words in each decade from the 1830s on, and at least twenty million words in each decade from the 1880s on. Overall, it is about one hundred times as large as the four combined corpora in the Brown family of corpora. It is designed to be representative as well.

From the 1870s on, each decade has essentially the same proportion of fiction, popular magazines, newspapers, and nonfiction books (for a total of two hundred million words from fiction, a hundred million from popular magazines, forty million from newspapers, and sixty million from non-fiction books). In addition, the corpus was carefully designed to be balanced and representative at the level of sub-genres and domains as well. For example, the sixty million words of text from nonfiction books have texts from twenty distinct Library of Congress categories (e.g., religion, history, "domestic arts", agriculture, and engineering), and the balance between these categories stays essentially the same from decade to decade.

– Google Books (N-grams; books.google.com/ngrams). In terms of size, Google Books n-grams (hereafter "Google Books") dwarfs any other carefully constructed corpus. For American English alone, it is based on more than 155 billion words of text. Whereas COHA is one hundred times as large as the Brown family of corpora, Google Books is about four hundred times as large as COHA. On the other hand, Google Books was not designed to be representative; the Google Books team simply scanned everything they could find in several large university libraries. This resulted in more than fifteen million books scanned, five million of which were processed to form Google Books.

In summary, then, we have three historical corpora, which will form the basis for the comparisons in this study. The Brown family of corpora was designed to be small but representative. COHA was designed to be both large and representative. And Google Books is very large, but it makes no pretension of being representative.

3 The importance of size

English historical linguistics has a strong tradition of small, well-designed corpora, in the range of one to five million words each. 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 million words, UK 1800s), and the *Helsinki Corpus of English Texts* (1.6 million words, Old English through the early 1700s). These are all "general" historical corpora – they cover 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.

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 (Hofland and Johansson 1982; Leech and Fallon 1992; Oakes and Farrow 2007; Baron, Rayson and Archer 2009; and Baker 2010, 2011). 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 section, we will continue Baker's line of research and show empirically what types of lexical data we can extract from a small 2- to 4-million-word historical corpus 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 we will examine how well the one million–word Brown and Frown corpora (US, 1960s and 1990s) provide comparable 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 in Figure 1.

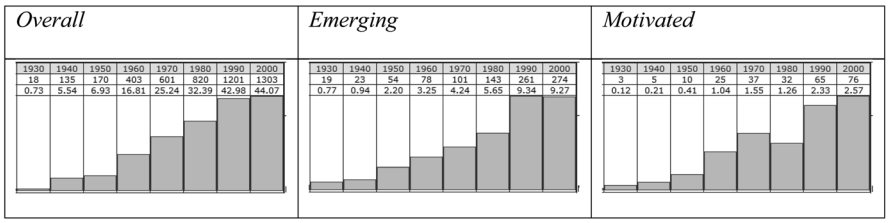


Figure 1: COHA: Adjectives doubling in frequency, 1960s–1990s

Table 1 shows that in COHA there are fifteen adjectives that have a combined frequency of between 800 and 1,600 tokens in COHA in the 1960s and 1990s (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 and 400 tokens in COHA in these two decades (e.g., *emerging* [shown above], *compelling*, *indoor*, *preferred*, and *unclear*) and 394 types with a frequency of between 50 and 100 tokens (e.g., *motivated* [shown in Figure 1], *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 fifteen COHA adjectives that have at least doubled in frequency and which have a combined token frequency of 800 to 1,600 in COHA in the 1960s and 1990s, all of these occur at least once in Brown/Frown, which is encouraging. One word occurs between one and nine times in Brown/Frown, and the other fourteen occur at least ten times (e.g., three tokens in Brown and seven tokens in Frown), which is perhaps enough to show an increase from the 1960s to the 1990s. Of these fourteen adjectives that occur at least ten times, six do show frequency that has doubled from the 1960s to the 1990s (e.g., Brown,

six; Frown, twelve, which is shown as "Support" (COHA) in Table 1 above). Another five adjectives show an increase but less than the doubling in COHA (e.g., Brown, six, and Frown, seven; shown as "? ? ?" above). And in three cases, the Brown/Frown data actually shows a decrease from the 1960s to the 1990s (e.g., Brown, seven, Frown, four; shown as "Contradict" above). Overall, then, six of the fifteen types (40%) of these high-frequency adjectives in Brown/Frown show the same doubling in frequency that is shown in the robust data (800–1,600 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, eight do not occur at all in Brown/Frown, and forty-six occur just one to nine times, which is probably too few to see an increase. Of those occurring ten times or more in Brown/Frown, fifty show a doubling, fifteen show a smaller increase, and eight 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 fifty and 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 ten times – probably too few to be useful. As a result, Brown/Frown 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 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, Figure 2 shows the collocates (nearby words) of *gay* in the 1800s and 1900s, and we can use these collocates to find evidence for semantic change during these two periods. For example, the older meaning of "happy, cheerful" is seen in lines 1 and 2, while the newer meaning related to sexual orientation is found in lines 4 and 6.

	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	129																	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	6	10	2	4	4	4			
8	MARRIAGE	93				1		1	1					1					1		7	81

Figure 2: COHA: Collocates of *gay*, by decade

With COHA, it is also possible to compare the collocates in different periods. For example, Figure 3 shows the collocates of *gay* in the 1830s–1910s (left) compared to the 1970s–2000s (right):

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	38.78
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	49	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

Figure 3: COHA: ADJ/NOUN collocates near the noun *gay*

Another example of comparing collocates is Figure 4, which shows the adjectival collocates preceding *women* in the 1830s–1890s (left) and the 1960s–2000s (right) and how women are represented and portrayed in the two periods:

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	264	3	2.02	0.02	82.75
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

Figure 4: COHA: Adjectival collocates of *women*

The important issue for our purposes here is the fact 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 one 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 thirty years, any changes that take place in between these years are 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 5):

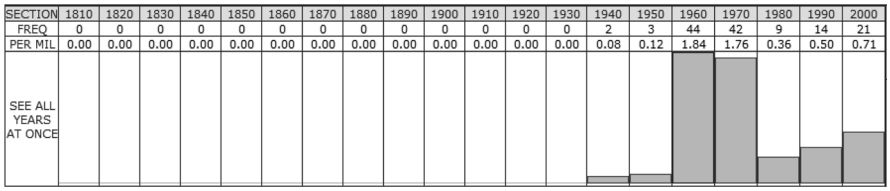


Figure 5: COHA: *Groovy*

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 6):

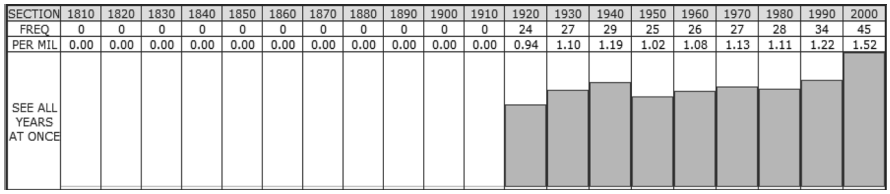


Figure 6: COHA: *Normalcy*

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 had only 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. 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 when looking at lexical change, which can occur quite suddenly.

In summary, we would agree with Baker (2011) that small one to four million–word corpora – while useful for high-frequency grammatical constructions – are in most cases inadequate for lexical studies (especially historical lexis), except for perhaps a handful of extremely frequent words.

4 Representativeness

We have now seen the crucial importance of size for historical corpora. In this section, we will consider the issue of representativeness. As we mentioned previously, a cardinal belief in corpus linguistics is that a corpus needs to be representative – meaning that the texts are carefully selected to produce a meaningful sample of the entire population. Ideally that population would be all English for a given time period, but in historical corpora, where we are limited to the writing that has survived, the population might more properly be said to be all writing that has survived. And since size is important, the population may need to be qualified once more to be all writing that has survived in sufficient quantities to be useful.

A key consideration for achieving representativeness, whatever the population, is widespread sampling among the groups that make up the population. Since the differences between groups will usually be greater than the differences within a group, the more groups that are included, the greater the chance that sample will not be skewed by the idiosyncrasies of one group. For similar reasons, balance across groups will be important as well: if one group (say newspapers) predominates in a corpus, the sample may be skewed toward the idiosyncrasies of that group, even if other groups are sampled. Size can also be framed as an issue of representativeness, in that the sample needs to be large enough for relevant linguistic features to show up. Biber (1993) calls this "linguistic representativeness".

All three considerations – widespread sampling, balance, and size – were important in the design of COHA. The range of sampling for COHA is seen partially in its four broad register divisions – fiction, nonfiction, magazines, and newspapers – but more emphatically in the wide variety of subtypes of each of these registers (http://corpus.byu.edu/coha/ > 400 MILLION WORDS, 1810-2009). Under fiction, for example, are samples from drama, movie scripts, novels, poetry, and short stories. Under nonfiction are samples from all twenty-one divisions of the Library of Congress classifications system. Under magazines are samples from 127 different magazines, representing a range of styles and

subjects, from general middle-brow magazines like *Harper's* to more specialized magazines like *National Geographic*. Under newspapers are newspaper samples from a range of geographic areas and subgenres within the newspapers, such as editorials and letters to the editor. When the subtypes are considered, it is clear that COHA samples from a wide variety of subtypes. The balancing in COHA and the size have already been demonstrated.

In contrast, the collection making up the Google Books corpus was assembled without any consideration of representativeness. As Michel et al. (2011) explain, the Google Books creators basically just went into large university libraries and scanned everything they could find. From the point of view of corpus linguistics, this is heresy. We should never expect the data from such a haphazardly created corpus to produce the same quality of data as a well-designed corpus like COHA. But as we will see in this section, that is precisely what happens (at least in terms of lexis), and this should come as quite a surprise to historical and corpus linguists.

Let us first examine some single-word and ad-hoc evidence for the similarity of (lexical) data from COHA and Google Books, after which we will carry out a much more systematic comparison. First, consider cases in Figures 7–12, which show the frequency of the words *bosom*, *steamship*, and *teenager* in both COHA and Google Books. (Note that all of the frequency charts for Google Books in this paper actually come from the googlebooks.byu.edu version rather than the standard interface at books.google.com/ngrams, although both versions are based on the same underlying frequency data.)

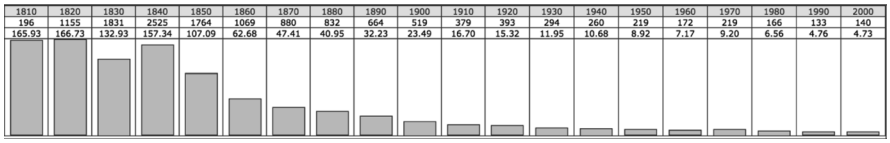


Figure 7: Frequency of *bosom* in COHA

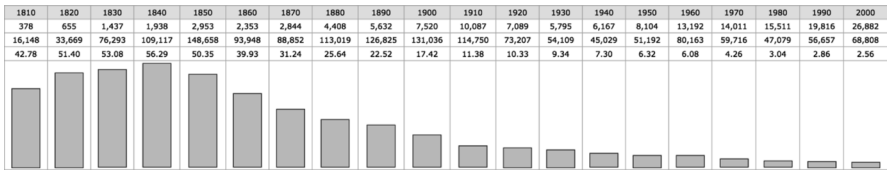
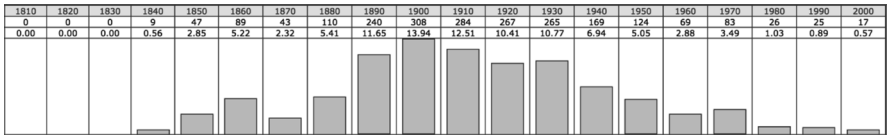
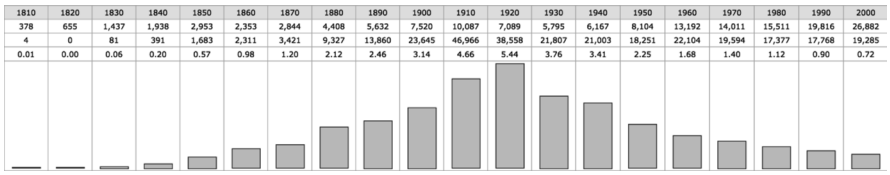


Figure 8: Frequency of *bosom* in Google Books



1810	1820	1830	1840	1850	1860	1870	1880	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000
0	0	0	9	47	89	43	110	240	308	284	267	265	169	124	69	83	26	25	17
0.00	0.00	0.00	0.56	2.85	5.22	2.32	5.41	11.65	13.94	12.51	10.41	10.77	6.94	5.05	2.88	3.49	1.03	0.89	0.57

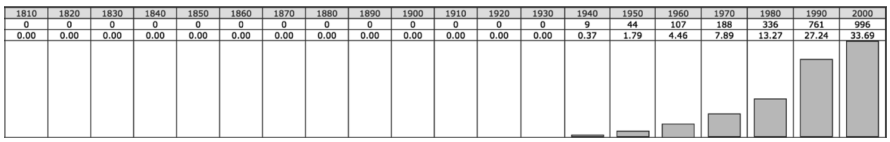
Figure 9: Frequency of *steamship* in COHA



1810	1820	1830	1840	1850	1860	1870	1880	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000
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
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
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 10: Frequency of *steamship* in Google Books

In both corpora, the frequency starts decreasing in the 1850s and is almost uniformly consistent since that time.



1810	1820	1830	1840	1850	1860	1870	1880	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000
0	0	0	0	0	0	0	0	0	0	0	0	0	9	44	107	188	336	761	996
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	0.37	1.79	4.46	7.89	13.27	27.24	33.69

Figure 11: Frequency of *teenager* in COHA

1810	1820	1830	1840	1850	1860	1870	1880	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000
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
1	5	0	0	0	0	0	2	1	3	1	13	20	39	1,321	8,422	21,496	50,645	97,783	151,757
0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.16	0.64	1.53	3.27	4.93	5.65

Figure 12: Frequency of *teenager* in Google Books

In both corpora, there are very few tokens before the 1940s, but there has been a consistent increase in frequency – decade by decade – since that time.

We should note that the similar frequencies over time are limited not just to single words like *bosom*, *steamship*, and *teenager*. They also extend to phrases, such as "so ADJ as to VERB" (e.g., *so good as to tell me*) or "have quite VERB-ed" (*he had quite forgotten her name*), as seen in Figures 13–16.

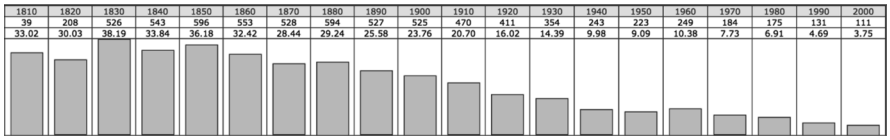


Figure 13: Frequency of *so ADJ as to VERB* in COHA

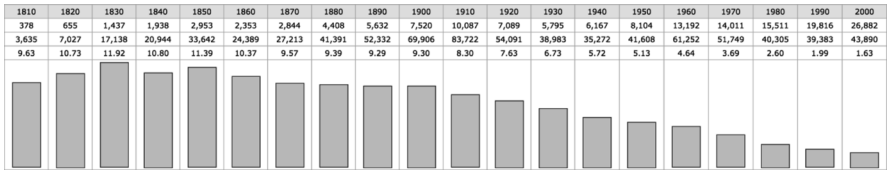


Figure 14: Frequency of *so ADJ as to VERB* in Google Books

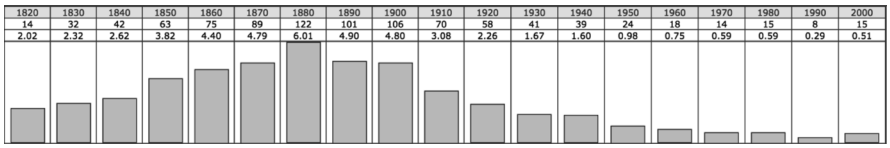


Figure 15: Frequency of *have quite VERB-ed* in COHA

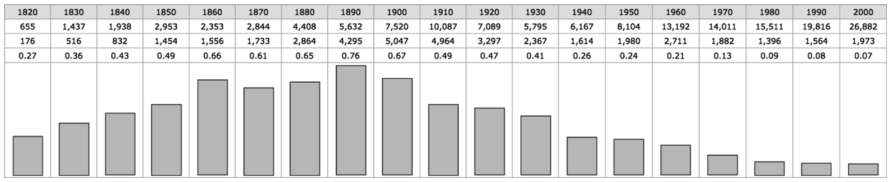


Figure 16: Frequency of *have quite VERB-ed* in Google Books

And although the focus of this paper is a comparison of lexical frequency, we should also mention that there is typically very good similarity in terms of syntactic constructions as well. To take just one quick example, consider the frequency of "NEED NEG VERB" over time in Figures 17 and 18 (e.g., *you needn't worry*; cf. the alternative *you don't need to worry*):

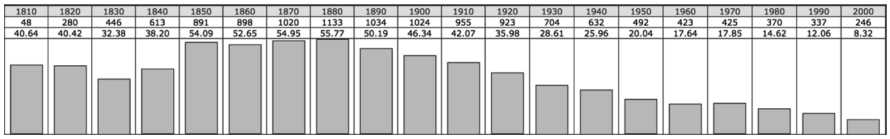
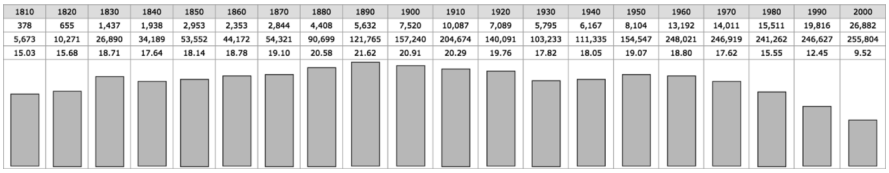


Figure 17: Frequency of *NEED NEG VERB* in COHA



1810	1820	1830	1840	1850	1860	1870	1880	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000
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
5,673	10,271	26,890	34,189	53,552	44,172	54,321	90,699	121,765	157,240	204,674	140,091	103,233	111,335	154,547	248,021	246,919	241,262	246,627	255,804
15.03	15.68	18.71	17.64	18.14	18.78	19.10	20.58	21.62	20.91	20.29	19.76	17.82	18.05	19.07	18.80	17.62	15.55	12.45	9.52

Figure 18: Frequency of *NEED NEG VERB* in Google Books

We see an increase throughout the 1800s, peaking in the late 1800s or the first decade of the 1900s, and then a fairly sustained decrease since that time.

All of the preceding, however, is merely anecdotal, and perhaps we are just "cherry-picking" the best examples to show a similarity between COHA and Google Books. What we need is a much more systematic comparison of the two corpora. The following is a description of the process we used to carry out a comparison of lexical frequency for thousands of different lexical items in the two corpora.

We first created a list of all nouns, verbs, adjectives, and adverbs that occurred at least fifteen times in COHA, which amounted to 116,630 different word forms – *statues*, *remarked*, *massive*, *particularly*, and others. We then took a sample of every tenth word (11,663 words total) and found the frequency by decade in both COHA and Google Books. Table 2 provides an example of these frequencies and shows the partial data for the word *steamship*. For reasons of space on this printed page, we show only the data for the 1870s–1950s, but in the study we looked at the frequency in each of the twenty decades from the 1810s–2000s. This table shows the raw frequency (e.g., COHA #) in both COHA and Google Books, as well as the normalized frequency (per million words) in each corpus (e.g., COHA PM).

Table 2: COHA/Google Books correlation

	1870s	1880s	1890s	1900s	1910s	1920s	1930s	1940s	1950s
COHA #	29	86	191	236	246	219	221	137	96
COHA PM	1.56	4.23	9.27	10.68	10.84	8.54	8.98	5.63	3.91
GB #	3,421	9,327	13,860	23,645	46,966	38,558	21,807	21,003	18,251
GB PM	1.20	2.12	2.46	3.14	4.66	5.44	3.76	3.41	2.25

For each of the 11,663 words, we then computed the Pearson correlation between the normalized frequency in the twenty decades of COHA and the equivalent twenty decades in Google Books. For example, in the case of *steamship*, the correlation is 0.89, which is extremely high (and a glance at the frequency charts from COHA and Google Books in Figures 9 and 10 show this as well).

The general rule of thumb with the Pearson correlation coefficient is that anything over about 0.50 is considered to be a moderate to high correlation and is assumed to be statistically significant. Let us take a look at the word *shuddering*, which has a correlation coefficient of 0.55, in Figures 19 and 20.

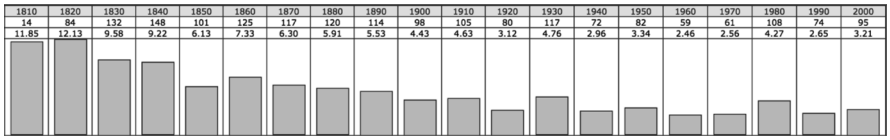


Figure 19: Frequency of *shuddering* in COHA

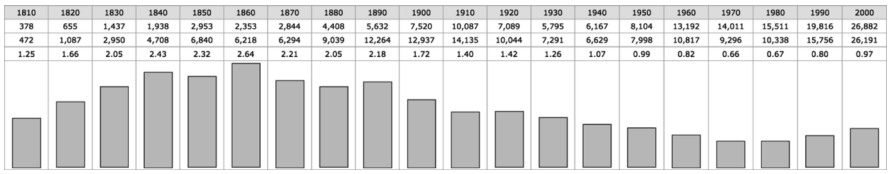


Figure 20: Frequency of *shuddering* in Google Books

We notice here that both corpora show a general decrease over time, but in COHA it is a fairly sustained decrease since the early 1800s, whereas in Google Books there is an increase through the mid-1800s. It is not nearly as similar in the two corpora as it is with *steamship* (Figures 9 and 10), which had a correlation coefficient of 0.89.

Often the Pearson correlation coefficient is lower than we would expect, based on a quick examination of the frequency charts in the two corpora. For example, Figures 21 and 22 show the frequency of the word *sense* over time.

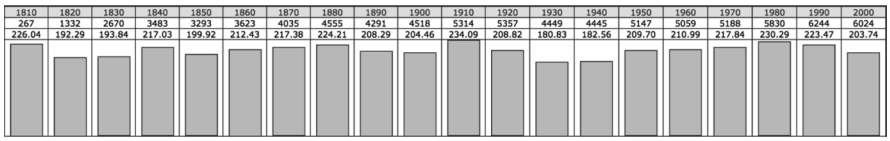


Figure 21: Frequency of *sense* in COHA

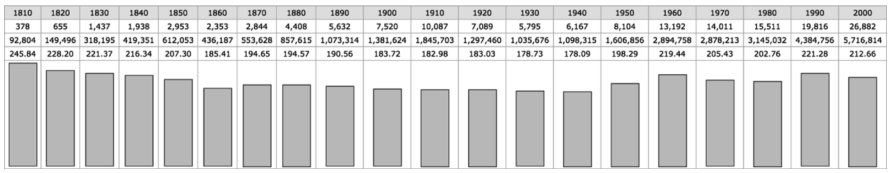


Figure 22: Frequency of *sense* in Google Books

Both corpora show that the frequency has been relatively flat for the last 150 years or so. But where there might be a slight increase between two decades in COHA, there is perhaps a slight decrease in Google Books (or vice versa). As a result, the Pearson correlation coefficient for this word is only 0.20.

Overall, the correlation between COHA and Google Books was 0.51, which shows a fairly large correlation between the two corpora. But we also wanted to see if there was some relationship between the frequency of a word and the COHA/Google Books correlation. Table 3 shows the correlation for ten different word-frequency bands – the top 10% most frequent of the 11,663 words (words 1–1,166 words; frequency-band 1), the next highest 10% (1,167–2,332; band 2), and so on.

Table 3: COHA/Google Books correlation by frequency band

Frequency band	Words	Correlation
1	1–1,166	0.620
2	1,167–2,332	0.591
3	2,333–3,498	0.548
4	3,499–4,664	0.530
5	4,665–5,830	0.489
6	5,831–6,996	0.468
7	6,997–8,162	0.436
8	8,163–9,328	0.401
9	9,329–10,494	0.383
10	10,495–11,660	0.379

This same data can also be represented in Figure 23, which shows the average Pearson correlation coefficient (Y axis) for the ten frequency bands (X axis).

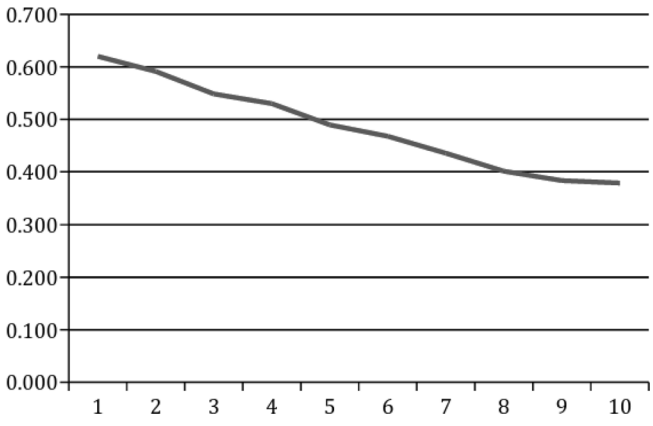


Figure 23: COHA/Google Books correlation by frequency band

What this shows, perhaps not surprisingly, is that the most frequent words in COHA correlate the best with the same words in Google Books.

On the other hand, words at the bottom of the frequency band (which occurred just 15–20 times overall in COHA), have only a 0.379 correlation with Google Books. This makes sense, because with just 15–20 tokens, the data is too sparse and the frequency data is too "spikey" in COHA. For example, consider Figure 24, which shows the frequency for *humpbacks*, which occurs only seventeen times in COHA, in contrast to Figure 25. The data (especially in COHA) is just too sparse to get a good correlation between the two corpora.

1810	1820	1830	1840	1850	1860	1870	1880	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000
0	0	1	0	0	0	0	0	0	0	3	0	0	1	0	0	2	5	1	7
0.00	0.00	0.07	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.13	0.00	0.00	0.04	0.00	0.00	0.08	0.20	0.04	0.24

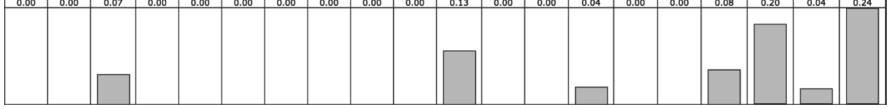


Figure 24: Frequency of *humpbacks* in COHA

1810	1820	1830	1840	1850	1860	1870	1880	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000
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
0	0	1	4	14	20	20	53	306	245	193	188	102	177	188	164	518	1,076	1,495	1,660
0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.01	0.05	0.03	0.02	0.03	0.02	0.03	0.02	0.01	0.04	0.07	0.08	0.06

Figure 25: Frequency of *humpbacks* in Google Books

Perhaps the best evidence for a strong overall correlation between COHA and Google Books comes from looking at words that have a strong overall trend toward increasing or decreasing use in the language. For example, consider *grieved*, which is clearly decreasing in frequency over time, in Figure 26:

1810	1820	1830	1840	1850	1860	1870	1880	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000
103	512	968	910	1009	679	590	561	543	499	388	304	182	131	138	134	159	125	146	156
87.20	73.91	70.27	56.70	61.26	39.81	31.78	27.61	26.36	22.58	17.09	11.85	7.40	5.38	5.62	5.59	6.68	4.94	5.23	5.28

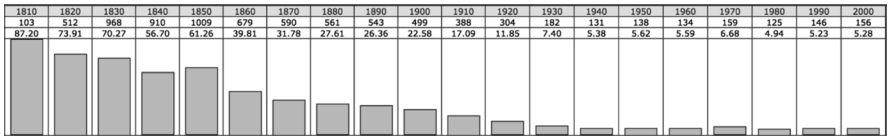


Figure 26: Frequency of *grieved* in COHA

The question is how well COHA and Google Books agree on these words that are strongly trending one way or the other. In order to measure this, we again looked at all 11,663 words in the study. We looked for words that changed in the same "direction" in four successive decades, spaced four decades apart. For example, we looked for the frequency of each word in the 1860s, 1900s, 1940s, and 1980s or in the 1880s, 1920s, 1960s, and 2000s. If the frequency was higher in the 1880s than in the 1920s, and more in the 1920s than in the 1960s (and so

on), then we would say that the word was in overall "decrease". And of course the overall frequency could be increasing as well.

The question, then, is whether words that had an overall increase in COHA had the same overall increase in Google Books, and the answer is that they clearly did. Table 4 shows that 354 words have a strongly increasing frequency in both COHA and Google Books, whereas 399 words have a strongly decreasing frequency in both corpora. There were only five words that were increasing overall in COHA but decreasing in Google Books, and only one that was decreasing in COHA but increasing in Google Books. Clearly, the two corpora are in very good agreement with each other, even though COHA was designed to be representative and Google Books was not.

Table 4: COHA/Google Books correlation: Overall increasing and decreasing

		COHA	
		increase	**decrease**
Google Books	increase	354	1
	decrease	5	399

5 Discussion and conclusion

In this study, we have looked empirically at the effect that corpus size and representativeness have on lexical robustness and accuracy (as measured by the frequency of words over time). The first important conclusion of this study is that size really does matter. Previous studies involving COHA (e.g., Davies 2012a, 2012b, 2012c) have looked at syntactic phenomena, and they show that a 400-million-word corpus like COHA allows for research on a wide range of syntactic phenomena that could never be studied effectively with a small 1- to 5- million-word corpus like ARCHER or the Brown family of corpora.

No previous study, however, has looked carefully at the effect of corpus size on lexical phenomena, such as the accuracy of word frequency over time. Small corpora have not been used much for lexical studies, and researchers perhaps have intuitively known that small corpora would not be robust enough for lexical studies (especially the frequency of words over time). But ours is the first study to empirically compare lexis in small and large historical corpora.

As we have seen in our comparison of COHA and the Brown family of corpora, the Brown family of corpora (four million words) provides poor data on lexical frequency for "medium"-frequency words like *emerging*, *compelling*,

indoor, *preferred*, and *unclear*. And it provides very poor data for lower-frequency words, such as *motivated*, *first-time*, *blurry*, *impaired*, *viral*, *obnoxious*, and *luscious*. For these words, the Brown family was able to show evidence for overall increases in the frequency of the word from the 1960s to the 1990s in only about 3% of all cases.

In addition, the Brown family of corpora are very poor in terms of "granularity" – the ability to track the frequency of words at, for example, the level of individual years. But as we have seen, lexical frequency can change dramatically in just a few years. Because COHA has at least two million words each year for every year since the 1870s, it can track such changes quite effectively.

These findings regarding the importance of corpus size for lexical studies are perhaps not surprising, and they probably confirm what others have suspected but never actually measured carefully. What is probably much more surprising are the findings regarding the importance of representativeness and the "correct" design of a corpus, at least as far as being able to accurately measure lexical change.

As we have discussed, a primary tenet of corpus linguistics is that representativeness is absolutely critical in corpus design. Corpus creators need to design a corpus so that it accurately reflects the target "population" of text or speech in the "real world". It is not enough to have a large corpus if it is not representative, and one of the key factors for making a corpus representative is to select from a wide range of texts and text-types. Proportion and balance from type to type and from decade to decade are also important considerations. Yet Google Books has disregarded these principles of representativeness. The creators of this collection simply scanned everything available in several large university libraries. Unlike COHA, there was no attempt to sample from multiple genres or to balance the selection across groups and decades. And yet Google Books provides data on lexical change (as measured by lexical frequency) that is very similar to that of COHA, which is a well-designed corpus. How can this be?

The answer may be simpler than we think. The concept of representativeness says that we should accurately "model" the entire target population of texts in the "real world". But what if you have, in effect, the entire target population at your disposal, or at least a sufficiently large percentage of it? In this case, modeling is not as important. The variety of text-types will be taken care of by a sample that is large enough to catch that variety. And this is precisely what Google Books has done.

Of course, a corpus sampled by genre will be important for investigations into genre differences, but when COHA is used as a single, undifferentiated corpus, it behaves remarkably similarly to a corpus that was never intentionally stratified

by genre to begin with. While it may still be important to make sure that a corpus representing "general language" is stratified by as many genres as possible, the data seem to indicate that it works just as well to come in and scan everything and not worry about genre. This is, of course, assuming that we have a corpus as massive as Google Books, where we have "the whole", rather than the case where we are merely sampling "the whole".

This conclusion might suggest that – at least in terms of historical data – corpus linguists ought to quit worrying about designing corpora and just include everything that is available. For periods covered by Google Books, we already have a corpus that performs as well as a carefully designed historical corpus like COHA. But this conclusion is too simplistic.

In the first place, the limitations of our interface with Google Books provides some serious obstacles to research. As we have discussed elsewhere (Davies 2014), Google Books is very good at tracking lexical change over time. And with the right architecture and interface (such as googlebooks.byu.edu), it can even be pressed into service to look at many types of syntactic change. But crucially, Google Books is not a real "corpus", in the sense that it contains sentences and paragraphs. It is composed of just n-grams – one-, two-, three-, four-, and five-word sequences. There is no context and no ability to search beyond those n-grams. This means that it is very difficult to extract collocates, even with an improved interface like googlebooks.byu.edu. In COHA, on the other hand, it is quite easy to find collocates and even to compare collocates in different historical periods (and this is an important method for examining semantic change; see Davies 2012a, 2012b, as well as Figures 2–4 above).

Even more seriously, there is an aspect of Google Books that few researchers seem to be aware of. Only those n-grams that occur forty times or more in the underlying corpus are searchable by end users, whether in the "standard interface" at books.google.com/ngrams or in alternative interfaces like googlebooks.byu.edu. Even in a massive corpus like Google Books, the vast majority of all two-, three-, four-, and five-word strings might occur just ten or twenty or thirty-nine times in the underlying corpus, but all of these would be "invisible" to the end user since they don't occur at least forty times. In COHA, on the other hand, all of the data is available, even if a string only appears one or two times. This makes COHA much more useful for looking at syntactic change (see Davies 2014).

In the second place, assembling a large-scale collection of texts for periods not covered by Google Books may be less practical than creating a smaller, widely sampled corpus. Perhaps university libraries will contain enough different kinds of books that the genre effects will not be large when the sample is large,

but for older texts, what similarly large collections do we have available? Will the *Early English Books On-line* (EEBO) collection, for example, be sufficiently stratified to ignore genres in selection? Or will it be idiosyncratic enough that ignoring other genres will make a difference? It would be interesting to see. And what about collections for even earlier stages of English? For those stages, where so little writing has survived, will it be possible to compile a corpus large enough that it will unintentionally sweep in enough genres to be representative of all language? As we go further back in time, perhaps the only claims for representativeness we can make is that the corpus is representative of what has survived, and in that case, a corpus like the *Dictionary of Old English Corpus* (DOE) will essentially be the whole.

Finally, we might turn the basic conclusion regarding COHA and Google Books on its head. Some might argue that "smaller" corpora like the 400-million-word COHA corpus are now "obsolete", with the availability of massive data sets like Google Books n-grams. But we could also argue that – at least in terms of researching lexical frequency – it was perhaps not necessary to spend hundreds of millions of dollars (and perhaps millions of man-hours) to create Google Books, when similar data can be found in COHA – which was created on a much more modest budget and by just one person.

The results of these comparisons are intriguing. We know that language varies in noticeable ways from genre to genre, but what happens when we mix all the genres together and look at averages across genres? Does it make a difference whether the sample was carefully stratified by genre or not? This paper has provided evidence that a large collection of books will sweep in enough different kinds of language use that the average does not look very different from a corpus that is designed to sweep in many different types of language use. It would be interesting to see if these same kinds of similarities between corpora show up for other kinds of linguistic structures. Figures 13–18 have shown that several lexico-grammatical constructions behave similarly in COHA and Google Books. Will that be the case with more such lexico-grammatical constructions or with other grammatical constructions? Would it make a difference if a corpus were stratified by even more genres than are present in COHA?

For now, this present study has raised questions about the relative importance of widespread sampling and sample size. When it comes to examining lexical frequency over time, corpus size is crucial. Given a large enough corpus (such as Google Books), however, wide-spread sampling may not be as important as we have traditionally thought. Finally, we will eventually want to do more than look at the frequency of words over time, and in this case a real corpus like the *Corpus of Historical American English* (COHA) is invaluable.

References

Baker, Paul. 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.

Baker, Paul. 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 & Dawn Archer. 2009. Word frequency and key word statistics in corpus linguistics. *Anglistik* 20. 41–67.

Biber, Douglas. 1990. Methodological issues regarding corpus-based analyses of linguistic variation. *Literary and Linguistic Computing* 5. 257–269.

Biber, Douglas. 1993. Representativeness in corpus design. *Literary and Linguistic Computing* 8. 243–257.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad & Edward Finegan. 1999. *Longman grammar of spoken and written English*. London: Pearson ESL.

Davies, Mark. 2012a. Expanding horizons in historical linguistics with the 400 million word Corpus of Historical American English. *Corpora* 7. 121–157.

Davies, Mark. 2012b. The 400 million word Corpus of Historical American English (1810–2009). In Irén Hegedus & Alexandra Fodor (eds.), *English historical linguistics 2010*, 217–250. Philadelphia: John Benjamins.

Davies, Mark. 2012c. Examining recent changes in English: Some methodological issues. In Terttu Nevalainen & Elizabeth Closs Traugott (eds.), *The Oxford handbook of the history of English*, 263–287. Oxford: Oxford University Press.

Davies, Mark. 2014. Making Google Books n-grams useful for a wide range of research on language change. *International Journal of Corpus Linguistics* 19 (3). 401–416.

Hofland, Knut & Stig Johansson. 1982. *Word frequencies in British and American English*. Bergen: Norwegian Computing Centre for the Humanitiesm & London: Longman.

Leech, Geoffrey. 2006. New resources, or just better old ones? The holy grail of representativeness. In Marianne Hundt, Nadja Nesselhauf & Carolyn Biewer (eds.), *Corpus linguistics and the web*. Amsterdam: Rodopi.

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 & Erez Lieberman Aiden. 2011. Quantitative analysis of culture using millions of digitized books. *Science* 331. 176–182. [Published online ahead of print 12/16/2010].

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.