Note: this page makes several references to Egbert, Biber, Gray (2022) (more). For convenience, we will sometimes refer to this as EBG 2022.

The British National Corpus: 2014 Written

English-Corpora.org (along with several other corpus websites) allows users to search the BNC **1994** corpus. Currently, BNC **2014** Written (which was actually released in 2021) is only available via the LancsBox X software (which is discussed in Section 5). But we would still like to provide an overview of the corpus – for those who have been using BNC 1994 at English-Corpora.org, and who might now be interested in BNC 2014 (Written).

1. Genres

The following is a comparison of the genres and sub-genres in the BNC 1994 and the BNC 2014 corpora, and shows the number of words in each genre (and sub-genre):

genre	BNC 2014	BNC 1994	genre [1]
informal speech	10,317,212	10,334,947	informal speech [1]
fiction	19,870,546	16,194,885	fiction [2]
magazines	14,979,505	16,190,916	magazine [3]
newspapers	19,996,923	9,345,878	newspapers [4]
academic prose	19,625,564	15,429,582	academic prose [5]
official documents: Parliament	2,000,778	1,156,171	Parliamentary [6]
official documents: business annual reports	4,996,104	119,808	Business reports [7]
written-to-be-spoken: drama	1,490,903		
written-to-be-spoken: TV scripts	1,500,611		
elanguage: email	58,582	213,045	Email [8]
elanguage: product reviews	1,407,408		
elanguage: social media	1,392,717		
elanguage: blogs	1,091,706		
elanguage: forums	1,000,805		
elanguage: SMS messages	220,180		
TOTAL	~100,000,000	~100,000,000	

As can be seen, the first five genres are roughly equivalent for BNC 1994 and 2014 (informal speech, fiction, magazine, newspapers, and academic prose), as well as Official Documents: Parliamentary proceedings. But BNC 2014 has three genres that were not present in 1994, or where there are many more texts in 2014 (2021 release) than in 1994 (shown in yellow above): written-to-be-spoken (dramas and TV scripts), elanguage, and annual business reports.

1.1 On the other hand, the following are genres and text types in BNC 1994, which do not find any direct equivalent in BNC 2014. The table shows the "Subject" in the spreadsheet of BNC 1994 texts, as well as the "Medium" (publication type):

[1] The following are the "Subject" codes for these genres from the spreadsheet that was created by David Lee: **1** S_ **2** W_fic **3** medium=periodical; not W_news **4** W_news; not medium="written_to_be_spoken" **5** W_ac **6** W_institut_doc+"annual report" **7** W_hansard **8** W_email

Subject	Publication type	# words
W_non_ac_humanities_arts	Book	3,605,171
W_non_ac_medicine	book	406,005
W_non_ac_nat_science	book	1,093,556
W_non_ac_polit_law_edu	book	1,412,429
W_non_ac_soc_science	book	2,822,101
W_non_ac_tech_engin	book	78,739
W_non_ac_humanities_arts	m_pub	123,638
W_non_ac_medicine	m_pub	30,232
W_non_ac_polit_law_edu	m_pub	61,357
W_non_ac_soc_science	m_pub	732,762
W_non_ac_polit_law_edu	m_unpub	9,770
W_non_ac_soc_science	m_unpub	353,249
W_admin	book	42,936
W_biography	book	3,507,358
W_commerce	book	2,687,257
W_institut_doc	book	69,733
W_instructional	book	197,401
W_misc	book	4,417,454
W_religion	book	1,065,595
W_admin	m_pub	1,362
W_advert	m_pub	530,460
W_biography	m_pub	16,820
W_commerce	m_pub	175,870
W_institut_doc	m_pub	134,013
W_instructional	m_pub	877
W_letters_prof	m_pub	492
W_misc	m_pub	1,209,403
W_religion	m_pub	16,449
W_admin	m_unpub	175,648
W_advert	m_unpub	27,673
W_biography	m_unpub	4,386
W_commerce	m_unpub	137,154
W_essay_school	m_unpub	146,530
W_essay_univ	m_unpub	55,717
W_institut_doc	m_unpub	222,707
W_letters_personal	m_unpub	52,480
W_letters_prof	m_unpub	65,539
W_misc	m_unpub	1,603,982
W_religion	m_unpub	17,846
W_news_script	to_be_spoken	1,292,156
W_misc	to_be_spoken	8,030
W_religion	to_be_spoken	21,742

Notice all of the genres and text types that were in BNC 1994, but which have been eliminated by BNC 2014 – texts such as “administrative documents”, letters, essays, instructional materials, advertisements, and more.

1.2 Why were these genres and sub-genres from BNC 1994 dropped in BNC 2014? It’s hard to say, but it is interesting that with these changes, BNC 2014 looks a lot more like COCA -- which also doesn’t have all of those text types either. Whether the creators of BNC Written 2014 have explicitly modeled that corpus on COCA (which was first released in 2008) is hard to say. But it is clear that both corpora have used a similar corpus design.

Egbert, Biber, Gray (2022:263) criticize COCA because it doesn't have the wide range of genres as BNC 1994 (in other words, all of the genres listed in the table above). They said that COCA claimed to "represent the entirety of (American) English", but that it had failed, because there were no advertisements, letters, instructional materials, etc.[2] But if COCA has failed, then the BNC 2014 has "failed" as well, since those text types are not in BNC 2014 Written either.

But actually, I don't think either corpus has failed. Quite frankly, **it is impossible to "represent the entirety" of a language in *any* corpus**. For example, that would require us to have texts or transcripts of people interacting with chiropractors, or menus from Korean restaurants, or freeway billboards, or warning labels from lawn care products, or EULA agreements for operating systems, or dishwasher repair manuals – or any one of thousands of other text types. "Corpus theorists" like EBG (2022) may wish for all of these things, but it's simply not realistic in most cases, and I think that the creators of BNC 2014 have (correctly and wisely) recognized that.

1.3 Note also that the **Longman Corpus** (which formed the basis of the Longman Grammar of Spoken and Written English) was created by Doug Biber and others, and it only had four genres – spoken, fiction, newspaper, and academic. COCA added to that the genres of magazines, TV and Movie Scripts, and texts from the Web (blogs and other web-based texts). While the BNC 2014 has relatively little from web-based materials (or TV and Movie scripts), it has added plays, Parliamentary debates, and annual business reports (as discussed above).

The goal should be to have texts from a number of genres from informal (such as informal speech or TV and Movie scripts) to formal (such as academic). This would be a "balanced" corpus of the type that I have tried to create with COCA, and which we find in the Longman Corpus and in BNC 2014. So while these corpora are balanced across genres from informal to formal, none of them accounts for the "entirety of English" in the way that ivory tower "corpus theorists" seem to want. And that's perfectly OK.

1.4 Some differences between BNC 2014 Written and COCA

Although the BNC 2014 Written and the COCA corpora are now quite similar, there are some important differences, as is shown in the table below. Genres that are much more common in BNC 2014 Written are in yellow in the COCA column, while those that are much more common in COCA are in blue.

genre	BNC 2014	COCA (1990-2019)	COCA / BNC	genre
informal speech	10,317,212	127,396,916	12.3	speech
fiction	19,870,546	119,505,292	6.0	fiction
magazines	14,979,505	127,352,014	8.5	magazine
newspapers	19,996,923	122,959,393	6.1	newspapers
academic prose	19,625,564	120,988,348	6.2	academic prose
official documents: Parliament	2,000,778			
official documents: business reports	4,996,104			
written-to-be-spoken: drama	1,490,903			
written-to-be-spoken: TV scripts	1,500,611	128,013,334	85.3	TV/Movie scripts
elanguage: blogs	1,091,706	125,496,215	115.0	
elanguage: other	4,021,110	129,899,426	32.3	
TOTAL	**99,949,544**	**1,001,610,938**	10.0	

1. Drama and TV scripts: COCA has about 85 times as much material from TV (and Movie) scripts as BNC 2014. We like these texts, because they reflect informal language very well. The BNC 2014 also has texts from parliamentary discourse and from plays, while these are not represented in COCA.

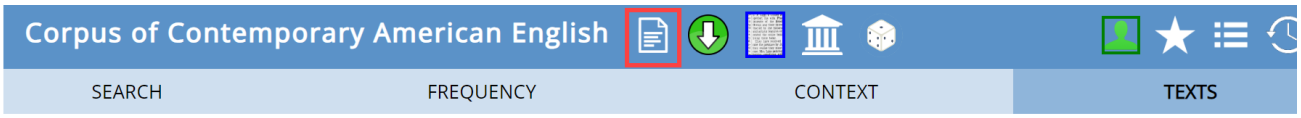
[2] As is explained here (1.1) I never claimed that COCA "represents the entirety of American English", nor anything like that. EBG (2022) have simply made up that quote.

2. Elanguage: This genre is about fifty times as large in COCA (250 million words vs 5 million in BNC 2014). But it's nice to finally have something in the BNC, no matter how small. BNC 1994 came out right when the Web was first becoming a think, and so naturally there were no texts from the Web in that corpus. That was understandable for 1994, but it became more awkward as people were using the BNC in 2000, or 2005 (when blogs were becoming a think), or 2010, or 2015, or even 2020 (BNC 2014 wasn't released until late 2021). I do question, though, **only** having **5 million words** in "elanguage". Surely more than 5% of the language input for an average speaker (at least the written language) comes from elanguage. I suspect that COCA (with 25% in "elanguage") represents much better what people encounter on a daily basis, rather than the small 5% in BNC 2014. (But again, anything at all from the web is an improvement over BNC 1994).

3. Business annual reports: It seems a bit **strange to have almost 5 million words** from this (sub-)genre in BNC 2014. Ideally, a corpus should reflect what native speakers see or hear "in the real world", but I suspect that few of us have ever read a single "annual business report" in our entire life. Does the average speaker of British English really read as much material from annual business reports (5% of the corpus) as they do material from the Web (5% as well)? COCA doesn't have anything from annual business reports. But again, since most people never encounter this genre in the "real world", that's probably not a problem.

2. Documentation and metadata

Egbert, Biber, and Gray (EBG) (2022: 261) discuss the importance of providing rich metadata for the texts in a corpus. But they claim that COCA fails in this respect, since it only provides "vague statements" about the texts in the corpus (2022:261). This claim is false. At **COCA** (and each of the 17 corpora from English-Corpora.org), users can click on a clearly marked link at the top of the corpus, which takes them to a **page that provides a great deal of information about the corpus**:



Download spreadsheet for all 485,179 texts (with summary by year, genre, and sub-genre)

The corpus is composed of more than **one billion words** in 485,202 texts, including 24-25 million words each year from 1990-2019. For each year (and therefore overall, as well), the corpus is evenly divided between the genres of TV and Movies subtitles, spoken, fiction, popular magazines, newspapers, and academic journals. This is important, because if you want to compare different years, you need to be comparing "apples" to "apples" (i.e. same genre balance in the different periods).

YEAR	BLOG	WEB	TV / MOVIES	SPOKEN	FICTION	MAGAZINE	NEWSPAPER	ACADEMIC	TOTAL
TOTAL	125,496,215	129,899,426	128,013,334	127,396,916	119,505,292	127,352,014	122,959,393	120,988,348	1,001,610,938
1990			3,207,900	4,374,469	4,162,242	4,101,447	4,082,931	3,983,143	23,912,132
1991			3,379,151	4,316,898	4,192,646	4,209,838	4,104,806	4,051,046	24,254,385
1992			3,183,858	4,523,054	3,893,956	4,288,694	4,092,031	4,028,147	24,009,740

This page has been available since the day that COCA was first released in 2008, as well as the spreadsheet (updated with each update of COCA), which lists information on each one of the 485,179 texts (# words, source, title, genre, sub-genre, URL if available, etc). How do these detailed entries constitute only "vague statements" about the content of the corpus, as suggested by Egbert, Biber, and Gray?

While the metadata for COCA is very robust, this is **unfortunately not (yet) the case for BNC 2014 Written**[3]. The version of BNC 2014 Written that was available in 2021 and 2022 had no documentation whatsoever for the 88,000+ individual texts in the corpus, nor any information about the size of sub-genres, etc. We just didn't know what was in the corpus. But things are improving.

In the 2023 release, some of this information was finally added. The following table show which genres (and sub-genres) in BNC 2014 have sufficient metadata to find the original texts (highlighted in green), which is probably a "baseline" in terms of what we should expect for metadata.[4]

# words	genre	sub-genre	available metadata	missing metadata
10,317,212	informal speech		(detailed)	
19,870,546	fiction		author, title	
1,490,903	written-to-be-spoken	drama	author, title	
1,500,611	written-to-be-spoken	TV scripts	series, episode	
1,091,706	elanguage	blogs	URL	title
14,979,505	magazines		magazine	date, title
19,996,923	newspapers		newspaper, date	title
19,625,564	academic prose		journal, author	date, title
2,000,778	official documents	parliamentary	date	speaker, title
4,996,104	official documents	business reports		(no information)
1,407,408	elanguage	product reviews	"Amazon"	(no other information; no URL)
1,392,717	elanguage	social media	"Facebook"	(no other information; no URL)
1,000,805	elanguage	forums	"Twitter"	(no other information; no URL)
58,582	elanguage	email		(no information)
220,180	elanguage	SMS		(no information)

Note that this **data is only available for about one third of the corpus** (34 million words in the 100 million word corpus). For example, the following are a few entries for the [magazines] genre in the metadata spreadsheet, and similar screenshots could be provided for the other genres listed above. There are no columns that provide the date or the title of the article (including the columns that were "hidden" for the screenshot to fit in this document). There is just no way to know what texts in the "real world" were used for about two thirds of BNC 2014. [5]

ID	Name	date	genre	id	inSample	mode	sample	section	source	subgenre	words
49755	MagCla1870.xml	NA	magazines	MagCla1870		writing	whole	NA	Classic Rock	magazines: music	3390
80214	MagCla1871.xml	NA	magazines	MagCla1871		writing	whole	NA	Classic Rock	magazines: music	853
85227	MagCla1872.xml	NA	magazines	MagCla1872		writing	whole	NA	Classic Rock	magazines: music	231
36725	MagCla1873.xml	NA	magazines	MagCla1873		writing	whole	NA	Classic Rock	magazines: music	599
27539	MagCla1874.xml	NA	magazines	MagCla1874		writing	whole	NA	Classic Rock	magazines: music	191
62689	MagCla1875.xml	NA	magazines	MagCla1875		writing	whole	NA	Classic Rock	magazines: music	349

[3] There is a **file** that is downloaded for the BNC 2014 from within LancsBox X, which contains **information on the 88,000+ texts** in the corpus. When I downloaded the corpus, the file [texts-whole corpus-1569214244.tsv] was put in the same folder as the corpus itself. Because this is a tab-separated file (.TSV), you can open this file directly in Excel (or another spreadsheet).

[4] The idea here is **replicability**, which is the foundation of empirical research. Researchers (and corpus creators) should provide enough data and explain their methodology well enough that subsequent researchers can replicate the findings of a study (or re-create the corpus). But if we have no idea where the texts for a corpus have come from, then this a significant blow to the goal of replicability.

[5] There might be a way around this, if a person had enough time. A person could enter the text name (e.g. MagCla1870.xml for the first entry above) in **LancsBox X** and then see the title of the article (although still not the date or page number). But assume that it takes 20 seconds to copy the filename from the spreadsheet, paste this into the right window in LancsBox X, find the title in the document, copy the title, and then paste it into a spreadsheet. To insert all of this information for the 88,000+ texts in the corpus, it would take approximately **4.5 months**, working 7.5 hours a day, five days a week. In the case of COCA, users can simply download one file from the corpus website, which already has all of this information.

As was mentioned, the meager metadata shown above was not even available in Version 1 or Version 2 of LancsBox X (the program that is required in order to use the corpus). It was only made available in Version 3, which was released in 2023 – about two years after BNC 2014 Written was released in 2021. So it is possible that the creators of BNC 2014 might at some point start providing more robust metadata, although it is unclear why they haven't done so before now.

Finally, we return again to the (incorrect) claim in Egbert, Biber, and Gray (2022:261) that COCA provides only a "vague" description of what is in the corpus. As we have seen, **COCA provides much more metadata for the texts in the corpus than does BNC 2014**, and with extensive data on all 485,179 texts in the corpus, COCA alone provides enough metadata for other researchers to re-create the corpus. Because at least two of these authors (Egbert and Biber) have close contacts with the BNC 2014 creators, they were probably aware of the difference between COCA and BNC 2014 Written as their book was being published in 2022, and so their incorrect claim is perhaps something that they should correct in any future editions of their book.

3. Corpus size [6]

When BNC 1994 was released 30 years ago, 100 million words was absolutely massive. But in the intervening three decades, things have changed. It is now common for corpora to be billions of words in size. For example, there are seven corpora at English-Corpora.org that are at least a billion words in size, and at sites like Sketch Engine there are even more. Virtually all of those large corpora (more than a billion words) at Sketch Engine are composed of just web pages, however. COCA is the only corpus anywhere that is balanced across genres (informal to formal) with at least a billion words of text.

For most studies of high and medium-frequency words, phrases, and syntactic constructions, 100 million words is adequate. For **lower frequency constructions and words**, however – and especially for the study of collocates (nearby words) – **100 million words may not be enough**.

For example, consider how corpus size affects the number of meaningful **collocates** for a word. The following chart shows 7-8 words chosen completely at random from different frequency levels in COCA – nouns near the 12,000th most common word, near #25,000, near #37,000, and near #50,000. [7] (Click on the two links for *prowess* to see the results from the BNC and COCA, and replace *prowess* with any of the other word to see the results for those words.

#12,000	BNC	COCA	**#25,000**	BNC	COCA	**#37,000**	BNC	COCA	**#50,000**	BNC	COCA
prowess	11	315	cirrhosis	22	79	chemise	0	13	bricolage	0	7
abstinence	15	363	tungsten	17	115	verbena	1	36	eyestrain	0	8
flair	25	304	anemone	6	31	zirconium	0	28	poppet	0	3
shard	1	60	slurry	20	70	logician	0	4	condyle	0	3
toxicity	15	400	vocali[sz]ation	0	24	ornithology	4	32	plasmacytoma	0	12
downfall	9	219	codex	3	68	duopoly	4	12	sailplane	0	4
seedling	3	99	cistern	20	37	farmyard	10	16	cryptogram	0	2
fiasco	8	274	rumination	0	27	self-fulfillment	0	28	bricolage	0	7

[6] This section is based on data from BNC 1994 rather than BNC 2014. This is due (in part) to the fact that we can't link to search results from BNC 2014 Written, since it is not online (as is BNC 1994). But because the two corpora are the same size, the results should be fairly comparable.

[7] Anyone who questions whether these words were really chosen at random can consult a frequency list of English words (for example this one), and find their own words at each of these four frequency bands.

The two numbers to the right of each word are the number of different noun collocates in the BNC and COCA that occur at least three times near the node word (4 words left / right). For example, there are 11 different nouns that occur near *prowess* at least three times in the BNC, and 315 different nouns in COCA.

Notice that for words near #12,000 the BNC is doing OK, although there are many more collocates in COCA. By the time we get to words near #25,000, the difference is even more apparent. Near #37,000, more than half of the words have 0 or 1 collocates in the BNC, but seven of the eight words have at least 12 collocates in COCA. And near #50,000, the BNC is failing to provide much at all in terms of collocates.

And let's take just one example of **phraseology / syntax** (although we could discuss many more). Consider the phrase [VERB her way PREP], such as *made her way through (the crowd)*. There are 14 different strings that occur at least three times in the BNC, and 82 different strings in COCA. That difference is significant, but it's the verbs themselves that are so much richer in COCA. In the BNC, there are five different verbs: (*made, pushed, found, worked, pick*ed}. In COCA, there are 32 different verbs: {*ate, became, clawed, danced, edged, elbowed, felt, forced, fought, found, fucked, groped, inched, knew, made, navigated, nudged, paid, prayed, pushed, sang, shouldered, shoved, slept, talked, threaded, weaved, wended, worked, wound, wove*}. It is this "lexical richness" that characterizes large corpora like COCA.

These are just a couple of examples of where size really does matter, and where the COCA data is much richer than that of the BNC, and many more examples could be given. And just to reiterate, for high and medium frequency words, phrases, and constructions, the 100 million word BNC should be just fine. But for lower frequency words, phrases, constructions – and especially for collocates (which allow us to look at the meaning and usage of a word) and for searches where there is interesting interplay between specific words and a particular syntactic construction – a much larger corpus like COCA is invaluable.

4. Historical change

As the creators of BNC 2014 Written have explicitly mentioned, one of the primary purposes in creating BNC Written 2014 was to provide the ability to look at changes in British English from 1994 to 2014:

> The British National Corpus 2014 is a major project led by Lancaster University to create a 100-million-word corpus of present day British English. This corpus has been constructed as a **comparable counterpart** of the original British National Corpus (referred to as the BNC1994 in this article), which was compiled in the early 1990s. . . . In sum, social changes and changes in technology over the past twenty or so years have transformed, among other things, communication and access to language data. For this reason, it is important to create a comparable counterpart to the BNC1994 **reflecting these changes** and taking advantage of new methods of data collection

Or, from this comment by one of the creators of the corpus:

> The British National Corpus 2014 is a project led by home of corpus linguistics, Lancaster University that will be open to all and used in the next 20 years by researchers and anyone interested in describing how language, 'real-life' language, is used and **how it changes over time**.

It is nice to finally have two data points for British English, in terms of different time periods. But as nice as this is, we should recognize **the limitations of having just two data points in the BNC, two decades apart**. To understand what these limitations are, consider the following data from COCA, which has 20-25 million words each year from 1990-2009, and where the genre balance is the same from year to year (so that we are "comparing apples to apples"). In the following table, the chart on the left shows the frequency of the word *Afghanistan* in the corpus each year from 1990-2019 (only part of the chart is shown here).

Afghanistan

SECTION NAME	# PER MILLION	# TOKENS	# WORDS
1990	10.88	258	23,707,141
1991	3.70	89	24,047,702
1992	7.65	182	23,803,545
1993	4.11	101	24,545,643
1994	1.92	48	25,002,985
1995	4.09	105	25,649,369
1996	6.63	163	24,585,307
1997	8.07	199	24,644,381
1998	8.51	213	25,037,330
1999	4.78	121	25,293,970
2000	7.53	189	25,095,795
2001	115.89	2833	24,446,361
2002	84.08	2085	24,798,572
2003	50.43	1270	25,185,875
2004	47.93	1203	25,097,903
2005	34.42	865	25,130,658
2006	50.79	1273	25,064,550
2007	39.33	971	24,690,690
2008	41.61	1002	24,082,614
2009	95.92	2310	24,082,494
2010	96.83	2332	24,082,934
2011	75.36	1932	25,636,361

Y2K

SECTION NAME	# PER MILLION	# TOKENS	# WORDS
1990	0.00	0	23,707,141
1991	0.00	0	24,047,702
1992	0.00	0	23,803,545
1993	0.00	0	24,545,643
1994	0.00	0	25,002,985
1995	0.00	0	25,649,369
1996	0.00	0	24,585,307
1997	1.34	33	24,644,381
1998	4.31	108	25,037,330
1999	30.96	783	25,293,970
2000	6.77	170	25,095,795
2001	1.39	34	24,446,361
2002	0.48	12	24,798,572
2003	0.48	12	25,185,875
2004	0.24	6	25,097,903
2005	0.36	9	25,130,658
2006	0.60	15	25,064,550
2007	0.32	8	24,690,690
2008	0.21	5	24,082,614
2009	0.33	8	24,082,494
2010	0.12	3	24,082,934
2011	0.35	9	25,636,361

Suppose we had only data from 1991 and from 2011 (two data points, 20 years apart, as with the BNC). While we would see an increase in the use of ***Afghanistan*** (on the left) between 1991 and 2011, we wouldn't have any idea when the largest increase occurred (2001, which was when the US invaded Afghanistan). Perhaps an even more striking example comes from the data for ***Y2K*** on the right (if you recall, Y2K was a computer bug that some people thought would cause havoc as the new millennium started in 2000). Suppose again that we had data from 1991 and 2011 (20 years apart, as with the BNC). The huge spike in frequency from 1997 to 2001 would be almost completely invisible, since the frequency in 2011 was just about the same as 1991.

The proceeding examples are with lexis (words and phrases), where things can change very quickly. But even with **lexico-grammatical phenomena**, 20 years apart might be too much. For example, consider the following data from the "like construction" (e.g. *and I was like, I'm not going to eat it*). If we had data just from the mid-1990s and the mid-2010s, we could see that the construction had increased dramatically during this time. But was it a gradual increase, or did it "spike" more in a particular decade, or was there more of an "S-Curve" (as is common with syntactic change)? With COCA, we can see that it followed more or less an "S-curve". But without data for the periods between the mid-1990s and the mid-2010s, there would be no way to verify this. That's **the problem with the BNC – there's no way to accurately track the changes from 1994 to 2014**.

SECTION	ALL	BLOG	WEB	TV/M	SPOK	FIC	MAG	NEWS	ACAD	1990-94	1995-99	2000-04	2005-09	2010-14	2015-19
FREQ	6489	196	283	2109	2931	65	591	306	8	36	127	248	607	1049	1728
WORDS (M)	993	128.6	124.3	128.1	126.1	118.3	126.1	121.7	119.8	121.1	125.2	124.6	123.1	123.3	122.8
PER MIL	6.53	1.52	2.28	16.47	23.24	0.55	4.69	2.51	0.07	0.30	1.01	1.99	4.93	8.50	14.08
SEE ALL SUB-SECTIONS AT ONCE															

Having "continuous" data (20-25 million words each year 1990-2009 in the case of COCA) is also helpful for looking at **discourse** – what is being said about a given topic. The following data from COCA shows the collocates of *crisis* in each five year period from 1990-2019 (and we could also see this year by year, if we wanted to). Notice the sudden decline of *gulf* and *oil* after the Gulf War in the early 1990s, and the *health* (*care*), *housing*, *refugee*, and *climate* crises during the last 15 years.

	ALL	BLOG	WEB	TV/M	SPOK	FIC	MAG	NEWS	ACAD	1990-94	1995-99	2000-04	2005-09	2010-14	2015-19
DEBT	1513	370	419	4	264	2	76	203	175	115	63	25	29	420	72
CRISIS	1088	142	146	64	268	24	140	128	176	194	128	76	116	162	124
HEALTH	917	65	92	36	209	7	214	177	117	149	84	84	133	115	195
ENERGY	771	87	96	32	129	10	135	196	86	94	43	245	123	48	35
GULF	656		5	3	267	1	122	194	64	629	11	3	2	4	2
HOUSING	549	107	112	4	119	5	53	116	33	12	13	17	101	104	83
RESPONSE	482	76	60	14	52	9	51	92	128	45	42	44	73	79	63
OIL	437	58	38	8	76	8	89	69	91	132	46	64	53	28	18
REFUGEE	381	10	16	6	122	6	57	65	99	47	48	19	15	27	199
SOLUTION	381	49	49	4	97	7	42	67	66	116	31	33	36	40	27
MIDLIFE	348	32	33	115	62	24	40	21	21	25	31	45	26	104	52
CREDIT	385	86	82	2	46	1	50	94	24	3	9	2	138	50	15
CLIMATE	396	114	80	2	65	4	86	14	31	2	1	12	48	33	106

5. LancsBox X (or, why a corpus is more than just the texts in a corpus)

Egbert, Biber, and Gray (2022) has as its title "Designing and Evaluating Language Corpora". But as we have discussed elsewhere (Section 4), they have a very narrow definition of "corpus". For them, **a corpus is only composed of the texts in the corpus. Nothing else matters** (or at least nothing else is discussed at any length in the book) – including the search engine to access the corpora, or other features that make the corpus helpful for the end user.

This approach is unfortunate, especially in a case like BNC 2014 Written. As we discussed in Section 1, BNC 2014 Written is – from the point of view of the texts in the corpus – quite a nice corpus. In many respects, BNC 2014 looks much more like COCA than it looks like BNC 1994, especially since it has a much more reduced set of genres and text types (and yet is still well-balanced from informal to formal texts). But – just as EBG 2022 runs into problems when they ignore things like the search engine for a corpus – the same is true of BNC 2014 (Written).

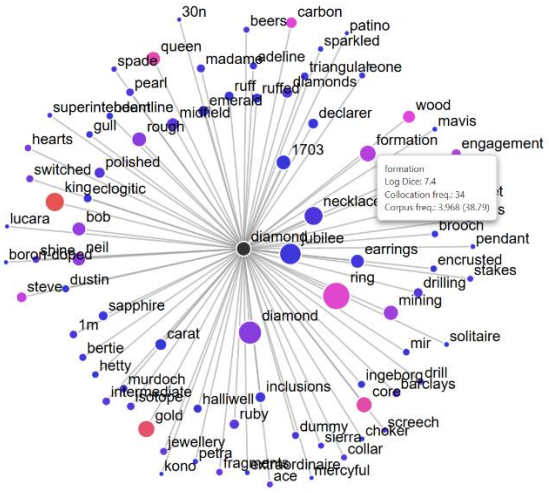
At the current time, BNC 2014 (Written) is only available via **LancsBox X**, a proprietary piece of software created at Lancaster University. This is very **similar to** the situation with **BNC 1994** in its first decade. For about ten years after BNC 1994 was released, it was only available via the **SARA** (and then **Xaira**) programs, which were very slow and very limited in what they could do. (I'm not aware of anyone who still uses SARA or Xaira to access the BNC.) But in the early 2000s, the BNC folks "open sourced" the texts, so that they could be used with other architectures and interfaces. As a result, there are now several incarnations of BNC 1994 (most of them available via the Web), and these different incarnations provide incredible functionality that was not available in SARA/Xaira.

Perhaps one of the biggest **limitations of LancsBox X** (even Version 3, which was released in 2023) is just **how slow it is**. For example, the following table shows how long different searches would take in the version of the BNC 1994 at English-Corpora.org, and in LancsBox X.[8]

[8] The architecture for English-Corpora.org is also very scalable. The same searches in the table above only take 10-15% longer to run in COCA, which is ten times as large as the BNC. And they only take about 2-3 times as long as that to run in the 14 billion word iWeb Corpus or the 18+ billion word NOW Corpus (which grows by about 4-5 million words each day, or 120-140 million words per month – and which is larger in size than either of the two BNC releases). There's no way that we can compare this speed to other corpus architecture like CQPWeb, since they don't allow corpora larger than about 2 billion words in size.

English-Corpora (click to run)	Time	Time [9] [10]	LancsBox X (CQP)
green N	0.5	20	[word="green"] [pos="NN.*"]
the green N	0.5	20	[word="the"] [word="green"] [pos="NN.*"]
more ADJ	0.7	32	[word="more"] [pos="J.*"]
BE * more ADJ	0.8	> 200	[hw="be"] []{1,1} [word="more"] [pos="J.*"]
PUT the NOUN PREP	1.3	22	[hw="put"] [word="the"] [pos="NN.*"] [pos="I.*"]
I VERB PRON BE	1.2	> 200	[word="I"] [pos="V.*"] [pos="P.*"] [hw="be"]
ADJ IDEA	0.4	21	[pos="J.*"] [hw="idea"]
VERB the IDEA	0.4	> 200	[pos="V.*"] [word="the"] [hw="idea"]

The version at English-Corpora.org is about 40-50 times as fast as LancsBox X. If a user is doing just one or two searches, that probably doesn't matter. But if they were doing 50 searches, this might take a combined total of less than one minute at English-Corpora.org, but probably 15-20 minutes or more with LancsBox X, with the user just sitting there, looking at the screen, waiting for the searches to finish.



And it's not just speed, either. Recent versions of LancsBox X have added in new features, like the ability to find **collocates**, such as the collocates of *diamond*, shown to the left. The ability to find collocates is a pretty basic feature that should be in any corpus analysis tool. It wasn't available in LancsBox X until 2023 (with Version 3), but it now does a great job of showing collocates with the Graph Coll display.

But this raises a very **important question about what a corpus is**. Suppose that someone is talking about BNC 2014 Written in May 2023, and complaining about how it can't find collocates. But the same person in December 2023 might comment how nice the GraphColl feature is in BNC Written 2014. Did the *corpus* change during these few months? According to Egbert, Biber, Gray (2022; see discussion in Section 4 here), the corpus did not change, since all that matters are the texts in the corpus. But from the point of view of the *end user*, the corpus (at least the data that can be gathered from the corpus) has changed dramatically. It is strange that **Egbert, Biber, Gray** (2022) **never considers the viewpoint of the end users** as they talk about "evaluating corpora" (part of the title of their book) – but rather, only how the texts in the corpus can be evaluated by "corpus theorists" such as themselves.

Coming back to BNC 2014 Written and LancsBox X, there are still many, **many features** that are available from other corpus site like Sketch Engine, CQPWeb, and English-Corpora.org, which make life much **better** for the end user **than the LancsBox X program** (but again, which would never be discussed by Egbert, Biber, Gray). To provide just a partial list from English-Corpora.org, we might note the more robust and much faster architecture, improved association measures, topics (which go way beyond what collocates can do, extremely detailed word sketches, powerful browsing of the top 60,000 words in the corpus (including by meaning and pronunciation), "analyzing" texts that users input (and output from KWIC displays), saved words and phrases, customized word lists (which can then be integrated into

[9] Three of these searches didn't finish within 200 seconds, which is when I cancelled the query. I'm not sure if they ever would have finished. Also, these searches were done on a machine with Windows 11, i7 CPU, 16 GB RAM, and an SSD drive, which is probably similar to the machines that many other people would be using for LancsBox X.

[10] The times shown above are for Version 3 of LancsBox. In April 2024, Version 4 was released, and it is now advertised as being "lightning fast". However, English-Corpora.org is still about 10 times as fast as Version 4. It's not clear what adjective could best define an architecture that is ten times as fast as "lightning fast". It would also be interesting to see if LancsBox X is scalable, and if it can efficently handle corpora that have billions of words of data, which is common now (compared to the small 100 million word BNC).

searches), reviewing and annotating search history, integration with external resources on the Web, and extremely powerful "Virtual Corpora". And there are many additional features at web-based corpus sites like Sketch Engine and CQPWeb[11] as well.

Just like BNC 1994 **data was essentially "trapped" inside** the SARA/Xaira program until the texts were "open sourced" in the early 2000s, the same is true of the **BNC 2014 Written** texts **with** the current **LancsBox X** program. Perhaps LancsBox X will be improved in ways that SARA/Xaira never was 25-30 years ago. But if not, let's hope that the creators of BNC 2014 Written allow the texts to be "open sourced" to other publicly-accessible web-based sites, to allow the full value of BNC 2014 Written to be available to researchers, teachers, and students worldwide.

6. Conclusion

The BNC 2014 Written corpus is a marked departure from BNC 1994, in that (like COCA) it has much fewer **genres and sub-genres** – about 8-9 genres in total. But we believe that as long as a corpus has the "major" genres from informal (e.g. speech) to formal (e.g academic), it can provide very useful data. In other words, we (and apparently the creators of BNC 2014 Written as well) disagree with Egbert, Biber, Gray (2022), who criticize corpora that "only" have 5-10 genres, and which therefore don't represent the "entirety" of the language[12] (which is an impossible goal anyway, see Section 1.1 above).

The main differences between BNC 2014 Written and the written genres in COCA are that COCA has about 50 times as much data from Web texts and about 85 times as much data from TV and Movie scripts (which represent informal language very well). BNC 2014 Written, on the other hand, has more from plays, parliamentary debates, and especially annual business reports.

Unlike COCA (which has robust metadata for all 485,179 texts), BNC 20214 Written only has enough **metadata** to identify these texts "in the real world" for about one third of the texts. In terms of **size**, the 100 million word BNC is probably large enough for high and medium-frequency words, phrases, and constructions. But a much larger corpus like COCA provides much richer data for lower frequency words and constructions, and especially for collocates (to examine the meaning and usage of a word). In terms of **historical change**, the BNC can provide data at historical Point 1 (1994) and Point 2 (2014). But it can't show intermediate changes in between these two data points the way that COCA can, which is especially problematic when looking at lexical change.

Finally, **access to the rich BNC 2014 Written data is limited by the LancsBox X software**. If the BNC 2014 Written ever becomes available via a modern architecture and interface, then users will truly be able to take advantage of its potential to look at different genres in present-day British English – in much the way that they can already do this [13] for American English with the Corpus of Contemporary American English (COCA).

---

[11] BNC 2014 Spoken is available from Sketch Engine and CQPWeb, but as of now (Jan 2014) BNC 2014 **Written** is only available via LancsBox X.

[12] They specifically single out COCA; see 1.1 here. Again, they don't mention BNC 2014 Written at all, even though it had already been released when their 2022 book was published.

[13] Needless to say, we fundamentally disagree with the creators of the forthcoming LANA Corpus, who argue that only with LANA will we *finally* have a corpus that provides rich, reliable data for contemporary American English. That corpus already exists, and it is the one billion word Corpus of Contemporary American English (COCA). We sincerely hope that the creators of LANA carry out detailed comparisons of LANA and COCA if/when LANA is finally released, and we definitely plan on doing that as well.