Designing and Evaluating Language Corpora by Jesse Egbert, Douglas Biber, and Bethany Gray (Cambridge Univ Press, 2022)

Reviewed by Mark Davies
March 2023

All three of these authors have produced great research in the past. And in my opinion, that is what makes this book – comparatively – such a disappointment.

I write this review as the creator of the corpora from English-Corpora.org, such as COCA and COHA. I will discuss some of the **factual inaccuracies** of the book with regards to the corpora from English-Corpora.org (especially COCA), since these corpora are discussed at some length by the authors. I will also discuss the authors' approach to other corpora, such as the Longman Corpus or the British National Corpus. Finally, I will discuss some general methodological issues, such as "empirical validity", and exactly what we mean when we talk about a "corpus".

Before discussing more substantive issues, I would first note that in about a third of the 35+ **references to the corpora** from English-Corpora.org, the authors used the URL / web address corpus.byu.edu, which ceased to be used more than three years before the book was published. And the majority of these links are to the web address www.corpus.byu.edu, which has never existed and which would have never worked. Why would these three authors have made such a basic error so many times in the book?

One possibility might be that some of the book was written by graduate students and then never adequately reviewed by the authors, which would also explain some of the more serious issues that are discussed below. Another possibility is that none of the three authors has really used the corpora (much), even though the corpora are probably the most widely used online corpora. Of course that raises the question of why they would spend so much time critiquing resources about which they know so little.

1. Straw man arguments

In several sections, the authors argue that I have claimed X (which I haven't), and then they proceed to "knock down" this "straw man".

1. 1 Representativity #1: "The entirety of English". On page 62 the authors state (emphasis added):

> There were four corpora that set out to represent both an entire general language and individual language varieties within it. Two well-known examples of corpora that fell into this category were COCA and the BNC 1994. Most scholars agree that it would be a daunting task to *try to represent an entire language* in a single corpus. Indeed, the corpora in our survey that make this claim *have clearly fallen short*. For example, a close look at the

> contents of COCA reveals that there are substantial sub-domains of American English that are completely absent from COCA (e.g., unscripted conversation[1], emails, text messages, workplace discourse, etc.). In fact, it is clear that there are far more registers of American English that are not included in COCA than those that are. In reality, the texts in COCA represent a limited range of registers. These texts, in almost every case, are either edited writing or scripted speech produced by professionals. The overwhelming majority of American English speakers never write or speak professionally, and most never have their writing edited by a team of professionals or their speaking scripted by professional scriptwriters. Thus the actual sample included in **COCA falls short of its claims to *represent the entirety* of "contemporary American English"** (see, e.g., Davies 2009:176).

This quote is from Davies (2009). Look on page 176 of that article (and throughout the entire article), and you will see that I never claim that COCA "represents the entirety of contemporary American English". In this and many other articles, however, I do make the claim that COCA is a "balanced" corpus, in that it has texts from a number of genres, from informal (spoken, TV and movie subtitles) to formal (academic).

In fact, COCA is based on the Longman Corpus of Spoken and Written English, which was created by Biber and others in the 1990s. That corpus had four genres: spoken, fiction, newspaper, and academic. When COCA was first released in 2008, it had those same four genres, plus magazines (because of my view that the lexis (words) in magazines is sufficiently different from newspapers to warrant the inclusion of this genre). From 2008-2019 there was never any claim that COCA included more than these five genres. In late 2019, some new genres were added (TV and movie scripts, and blogs and other web pages), and the detailed online description of the corpus was updated to reflect those changes.

But I have never claimed that COCA had more genres that it actually did, and I have *never* claimed that COCA represents American English "in its entirety". For example, I have never claimed that there are any transcripts of people interacting with chiropractors, or menus from Tex-Mex restaurants, or freeway billboards, or warning labels from lawn care products, or EULA agreements for operating systems, or dishwasher repair manuals – all of which would presumably be included in a corpus of American English "in its entirety"[2].

---

[1] The authors state that COCA has no "unscripted conversation". This is not true. As this web page indicates (note: the page works best from within the COCA corpus), the 127 million words of text in the spoken part of COCA is from *unscripted* conversations – albeit *unscripted* conversation on national TV and radio programs.

[2] This admittedly "snarky" comment reflects a more serious issue, which is how to determine exactly how "fine-grained" the domains should be, and just how wide our list of text types should be. For example, some would argue that you can't have a "general purpose" corpus without having poetry, but most people receive much more input in a given month from freeway billboards than from poetry. And people probably look at more restaurant menus every month than they look at "industry reports" or "Westerns (fiction)", which are genres that the authors mention approvingly in their discussion of the Brown Corpus (page 48). So which of these domains should be in a corpus?

In addition, if the overall corpus is tiny (for example, just 2-3 million words), then it might make sense to include some of these very narrow genres or domains – even if they have just 30,000-40,000 words of text. This is because even the "mainstream" genres (like spoken, fiction, magazines, newspapers, academic, or texts from the web), might have just

Also, why the focus on COCA, and just COCA? In BNC 2014 Written, there are now just 8-9 main genres and many fewer sub-genres than BNC 1994 Written. As is explained **here** in much more detail (Section 1), the creators of BNC 2014 Written have now eliminated sub-genres and text types like "administrative documents", letters, essays, instructional materials, advertisements, and much more. So at the end of the day, BNC 2014 Written actually looks quite similar to COCA in terms of genres and sub-genres.

Both Egbert and Biber have worked very closely with the creators of BNC 2014 Written, and they presumably knew – as they were writing their 2022 book – that virtually all of the criticisms that they had about COCA applied to BNC 2014 Written as well, including their complaint that "there are far more registers of American English that are not included in COCA than those that are. In reality, the texts in COCA represent a limited range of registers". Exactly the same thing now applies to BNC Written 2014. Why don't the author mention this in their book, and why do they single out COCA for criticism? What bias is at play here against COCA, and why?

1.2 Representativity #2. Another example of a "straw man argument" is when the authors claim (page 36) that I really only care about corpus size, but not representativeness:

> According to this conceptualization of a desirable corpus, the notion of **representativeness does not really matter**, because corpus size is the primary consideration

They are referring to this web page, in which I discuss the importance of corpus size. But nowhere on that page (or in any of my publications) have I ever suggested that corpus size is the only consideration, or even the most important consideration.

1.3 Representativity #3. A third example of a "straw man argument" is the following (page 37, emphasis added), in which they quote another author:

> [Patrick] Hanks goes on to applaud the efforts of Mark **Davies** because he was "***untroubled by*** *reservations about issues of* ***balance and representativeness*** — an approach that has enabled him to build large corpora … while others have floundered around worrying about theoretical obstacles"

First, this is a cheap shot on the part of Patrick Hanks, and it is unfortunate that the authors decided to include such an uninformed and biased quote in their book. Of course I care about corpus "balance" – that is something that I have written about repeatedly. But again, I have carefully avoided the "*r word*" (representativeness), simply because I know that this is such a "charged" term in corpus linguistics. I only use it in documents like this one (written after I had

200,000-300,000 words each. But what about larger corpora? COCA, for example, is one billion words in size, and each of the eight genres has between 120-130 million words. If we included the domain of restaurant menus or billboard signs, that would just be a tiny fraction of the corpus, and so these domains would essentially be "swamped" by the more "mainstream" genres, and comparison between the "mainstream" and the "niche" genres or domains would be problematic.

retired and after all of my other publications), to respond to the claim by Egbert, Larsson, and Biber (2020) about how COCA is supposedly so "unrepresentative".

2. Misrepresentations

There are also a number of misrepresentations of the corpora from English-Corpora.org – some of them quite elaborate.

2.1 COCA academic. For example, the authors dedicate several pages (201-206) to trying to show that COCA academic is not "true academic", because it is based only on **peer-reviewed academic journal articles**. (I have never claimed any more than that, so that argument is a straw man in and of itself; see the previous section).

But then in a strange twist, they focus (page 209 and several pages thereafter) on *one* supposed case of a non-peer-reviewed article in COCA academic (emphasis added):

> Academic Questions [the journal] is peer-reviewed, but *this particular article* certainly was not peer-reviewed: it is the verbatim transcript of a previously recorded speech.

And then on page 211:

> For example, our qualitative inspections revealed that some texts in *COCA-Academic are not actually examples of peer-reviewed academic writing*, or even writing at all.

COCA has more than 120 million words of text in more than 26,000 articles from academic journals. Each and every one of these journals is in fact a peer-reviewed journal. The *one* article that they have fixated on is a talk that was written and then presented at an academic conference, and was then reprinted in this peer-reviewed journal.

The authors suggest that their "qualitative inspections" have shown systematic problems with COCA academic. If so, they (or anyone else) can try a simple experiment. Take 500 articles from COCA academic – absolutely at random[3]. Check to see how many of these are not peer reviewed, and especially how many are "not written" (such as the text of a keynote address given at an academic conference). If there are more than *two* texts (which would be less than one half of 1% of the 500 texts under investigation), then there might be grounds for criticizing COCA academic for not being peer-reviewed, and for not having only "written" texts. Until then, we should be suspicious of any claims of meaningful "qualitative inspections".

[3] This shouldn't be hard. All of the nearly 500,000 texts in COCA are listed in this spreadsheet, and the textID is assigned at random. Find ~500 texts where the [textID] ends in two different two digit numbers, e.g. [23] and [60]. That should be about 500 articles.

2.2 COCA updates. On page 260, the authors claim that (emphasis added):

> The ***COCA*** *has had a* ***spotty*** *history of* ***updates***. After several years of remaining static, it just underwent a major update with the addition of hundreds of millions of words and four new genres.

This is false. COCA was updated every 12-24 months from March 2008 (when it was first released) until January 2020 (when the final version was released). There was never a period of more than 24 months in which it was not updated. Compare this to the BNC. The BNC was first released in 1994, but then the written part (which is 90% of the corpus) was not updated until 2021 – a full 27 years later. (And this new portion is still, as of March 2023, only available as part of the LancsBox software, with no ability to check sources, etc). Why would the authors make this false claim about COCA, and then completely ignore the very real issues of "spotty updates" with the BNC?

2.3 COHA genres. The authors suggest that my decisions about corpus design were based on "convenience" and whatever I could easily get my hands on. For example, on page 265:

> This was the case in corpora such as COHA, which contains four register categories, the biggest of which (Fiction) makes up 51 percent of the corpus, and the smallest of which (Newspapers) comprises only 10 percent. Presumably, these **choices were made based on convenience** and text availability.

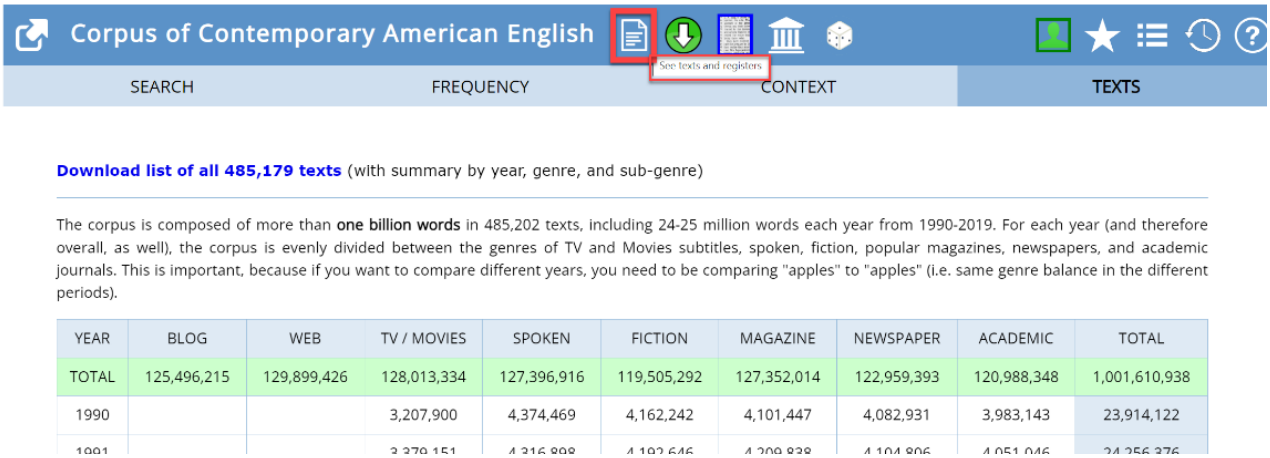
Actually, COHA (the Corpus of Historical American English) was meant to greatly expand on the Brown family of corpora, which have been used to look at language change from the 1960s (in e.g. the Brown and LOB corpora) to the 1990s (the FROWN and FLOB corpora). In the Brown family of corpora, 50% of the texts were from fiction and 50% from non-fiction. COHA was designed the same way, so that there could be comparisons between COHA and the Brown family of corpora.

And actually, collecting the 223+ million words of text in the nearly 15,000 texts from the fiction genre was very difficult, whereas collecting the newspaper texts was relatively easy. In the case of fiction, we had to scan in and then manually correct the data for thousands of books (tens of millions of words of data) from the 1820s to the 1980s, and that took many hundreds (perhaps thousands?) of hours. If we were looking for "convenience", we would have had hardly any fiction at all, and we would have had much more from newspapers.

2.4 COCA description / genres. As a final example, on page 261 the authors state that (emphasis added):

> In most cases, we resorted to inferring the target domain for a corpus from the name of the corpus or from ***vague statements*** *made in the* ***description of the corpus*** (see, e.g., WordBanks, COCA). For the general corpora in our survey, it was extremely *rare for corpus documentation to contain any mention of a target domain*, let alone a clear definition of its parameters.

Again, this 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:



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 nearly 500,000 texts (# words, source, title, genre, sub-genre, URL if available, etc). How does that constitute only "vague statements made in the description of the corpus"?

Also, as is explained **here** in detail (Section 2), it is BNC 2014 Written that has serious problems in terms of metadata, not COCA. Versions 1 and 2 of BNC Written 2014 (from 2021 and 2022) had no metadata whatsoever for the texts in the corpus – users had absolutely no idea what texts were in the corpus, or how many words of text there were in different genres and sub-genres. Some metadata was finally added in Version 3 (released in late 2023), but for more than 2/3 of the texts in the corpus, there is still not enough metadata (such as the title of the article or the URL) to identify the text in the "real world". As mentioned, COCA provides this data for all 485,179 texts. Why do Egbert, Biber, Gray focus (so incorrectly) on COCA, and then completely turn a blind eye to the very real problems in BNC 2014 Written?

3. "Empirical evidence"

3.1 Incorrect data from COCA and BNC academic (2020)

On page 30, the authors state the following (emphasis added):

> Take, for example, a widely used corpus like the Corpus of Contemporary American English (COCA). To our knowledge, *the compiler of this corpus has never offered empirical evidence* to support a claim that the corpus is "representative" of American English

This is both nonsensical and wrong. First, it is again a straw man argument, in the since that I have never argued that COCA is "representative", in the narrow sense in which they define the term.

More importantly, they suggest that there is no empirical evidence for claims made about COCA. This is not true. In 2020, Egbert and Biber (and Larsson) released a shorter version of what would become the 2022 book. In that book, they criticized COCA for not have a "true" academic component, at least compared to the British National Corpus (BNC), and they based their argument on three supposed pieces of linguistic evidence that show that BNC academic is more representative of "true" academic than COCA academic. **I provided detailed data** to show that in all three cases (#1-3 on pages 2-3) the truth is exactly the opposite of what they claim, and that COCA academic is actually more "academic" than BNC academic. I emailed the results of these studies to the three authors, and they emailed back to say that they received it. However, the authors chose not to include any of these three supposed pieces of "evidence" in their 2022 book. Apparently it was okay to include these tests when they supposedly favored the BNC, but they needed to be removed once it was shown that they favored COCA.

3.2 Incorrect data from COCA and BNC academic (2022)

Rather than re-use any of these three tests in the 2022 book, the authors found two *new* tests that (they claimed) showed that BNC academic is "better" academic than COCA academic. The first is NOUN+NOUN sequences (e.g. *data collection, case study*), and the second is "noun complements[4]. Unfortunately, the authors never provide a list of the head nouns that are involved in their BNC/COCA study of noun complements, and so it is impossible to confirm their claims regarding that construction.

But in the case of **NOUN+NOUN**, actual **data from COCA and BNC disproves the claim made in the book** (pages 211-215). The following is the frequency of NOUN+NOUN in the academic portion of both COCA and the BNC. (Click on the links for a list of NOUN+NOUN in the two corpora.[5]

	# words in academic	# NOUN+NOUN	NOUN+NOUN pmw
COCA	119,790,456	4,225,693	0.035
BNC	15,331,668	360,362	0.023

As can be seen, NOUN+NOUN is more than 50% more common (per million words: pmw) in COCA than in the BNC. Following their criteria, COCA has "better" academic.

Related data also shows that **the lexis (words) in BNC academic seems to be quite "skewed"** towards medical texts. For example, each of the following phrases is in the top 35 NOUN+NOUN sequences in the BNC: *gall bladder, bowel disease, biopsy specimens, acid secretion, bile acid*, and *plasminogen activator*[6]. The academic genre should have texts from a wide range of domains – like

[4] In noun complements, (for example, *the claim* [*that this research is valid*]), the head noun (e.g. *claim*) does not play a role in the subordinate clause. Compare this to a sentence like *the man* [*that you saw ~~the man~~*], where the head noun does play a role in the subordinate clause (as object in this case)

[5] These lists show all NOUN+NOUN strings that occur five times or more in the 120 million word academic portion of COCA corpus, and two times or more in the 15 million word academic portion of the BNC.

[6] As this document shows, many of these medical terms are found in just a handful of texts – maybe just 5-10 texts total. This is one of the downsides of having an academic component like the BNC, where 1% of the texts is comprised of only

science, law, philosophy, humanities, history, education, etc. (**COCA is balanced** between these domains.[7]) Why the serious skewing towards the medical domain in the BNC? Egbert, Biber, and Gray (2022) focuses almost entirely on syntactic issues rather than lexical issues, so perhaps these seriously skewed texts escaped their notice.

And just to summarize, the following are the results of the five tests that the authors mention in the 2020 and 2022 books, in their comparison of BNC and COCA academic:

Book	Phenomenon	BNC vs COCA: which is more "academic"
2020	Linking adverbials	They say BNC, but it is COCA when adjusted for time/dialect (#1)
2020	Nominalizations	They say BNC, but it is COCA when adjusted for time/dialect (#2)
2020	Lexical	Clearly COCA (#3)
2022	NOUN+NOUN	Clearly COCA (see discussion above)
2022	Noun complements	Possibly BNC (but no clear explanation of data in book)

Based on the data that the authors themselves provide in their books, **COCA academic** is probably **more representative of** what we would expect from **academic than** is the academic of **the BNC.**

3.3 No corpora are "empirically evaluated"?

I must admit that I am very confused by Appendix B: Survey of Corpus Design and Compilation Practices (pages 226-270). In this extended appendix, the authors list thirty corpora (most of the major corpora of English), the first four of which are corpora from English-Corpora.org: COCA, COHA, GloWbE, and NOW. For each of the thirty corpora, they ask whether the corpus has been "Empirically Evaluated", and for each one of the thirty corpora, they say that the answer is "No". From page 267:

> To the best of our knowledge, none of the corpora in our survey has ever been evaluated empirically for representativeness. This was true for both the domain characteristics and the distribution of linguistic features. If this evaluation has been carried out, it was not mentioned in the documentation for the corpus.

What does this even mean? For the authors, the only corpora that are "empirically validated" are apparently just those that discuss this in just the right way (so that it can be "counted" by the authors) – but only in the "official" documentation for the corpus. In academia, however, **empirical validation** usually comes from **similar results** across different studies, and especially from **different researchers**, and using **different approaches**.

To take a simple example, suppose that a corpus (like GloWbE) claims to be "representative" of the countries in the corpus (20 different English-speaking countries in the case of GloWbE).

about 5 texts. In COCA, on the other hand, 1% would be about 260 texts. So having a very small number of texts is nice because then you can carefully account for. each. and. every. text. The downside, however, is that just a handful of texts can really skew the data.

[7] See the ACAD(ademic) entries in the [sub-domains] tab from the list of nearly 500,000 texts in COCA.

Researchers can search GloWbE for words or phrases that are commonly accepted as being more common in a particular country (like *soft days* or *banjaxed* in Irish English). If the data from GloWbE shows that these words and phrases are in fact more common in the 100+ million words in the Ireland portion of the corpus, and if that **data agrees with the findings of other corpora** like ICE-Ireland, and if it agrees with any number of other studies of Irish English, how is this not "empirical validation"? And in the case of GloWbE and many of the other corpora from English-Corpora.org, there are detailed studies in respected academic journals, which discuss this empirical data in detail (see this as one of the early studies of GloWbE, but hundreds of other articles have been published since then).

And this type of "empirical validation" has been done – tens, hundreds, and perhaps even thousands of times – for corpora like COCA and COHA. In the case of COHA, for example, the lexical, syntactic, and semantic data agrees very well with the data from other historical corpora of English, such as the Brown family of corpora (or even Biber's ARCHER corpus). And again, in the case of COHA, there is ample **"empirical validation" in journal articles** (like this one), which shows that the COHA data agrees with many other studies (on lexical, phraseological, syntactic, and semantic change in English). And of course, there are even more detailed studies (such as this one for the "verb someone into V-ing" construction), which shows that the COHA data agrees extremely well with the data from other historical corpora of English.

And in the case of COCA, the data regarding different genres agrees extremely well with data from the Longman Corpus (and Biber's Longman Grammar). I know this, because I used the Longman Grammar as I taught English Grammar for 17 years – probably 30+ classes total (including the one that Jesse Egbert took as an undergraduate student). In each class from 2008 (when COCA was created) until 2020 (when I retired), we compared the Longman data to COCA data. In chapter after chapter in the student grammar, I would show how Biber et al predicted genre-based variation X, and then we would find that same variation in COCA. In addition, the COCA data has been compared to results from many **other studies that use other methodologies**, such as response times on Lexical Decision Tasks, or training of language models for Natural Language Processing, or surveys of native speakers.

Again, how does all of this corroborating data ***not*** provide "empirical validation" for the corpora? How can it possibly matter whether the authors did or did not find a specially crafted proof of "empirical validation" in the official corpus documentation? And how is it that other journal articles that were written by the corpus creator(s) soon after the release of a corpus do not "count" for "empirical validation"? Why only the special statement in the "official documentation"? Clearly, the authors are using an exceptionally narrow and non-standard definition of "empirical validation".

Comparison #1: Longman Corpus

As 1.1 above shows, the authors criticize COCA, because it only has the five genres of spoken, fiction, magazines, newspapers, and academic, rather than genres/registers/domains like transcripts of people interacting with chiropractors, or menus from Tex-Mex restaurants, or

freeway billboards, or warning labels from lawn care products, or EULA agreements for operating systems, or dishwasher repair manuals.

If COCA is "problematic" in this respect, then so are many other corpora, such as the Longman Corpus of Spoken and Written English that created by Biber and others in the 1990s, and which served as the basis for the Longman Grammar of Spoken and Written English. Personally, I think the Longman Grammar is an incredible resource, and I used the student version of this book for nearly twenty years as a teacher of English grammar.

But are Biber's Longman Corpus and the Longman Grammar fundamentally flawed, because they don't include every possible genre of American English? The authors don't say one way or the other. But I would argue that the Longman Corpus is not flawed, in the same way that COCA is not flawed – simply because they "'only" have four or five of the major genres of English.

In addition, the authors have a bit of a methodological issue when it comes to evaluating corpora. This is because they often provide data from the Longman Corpus to suggest what we "should" see in different genres like spoken, fiction, newspaper, or academic. (Examples of this would be the NOUN + NOUN and noun complement constructions discussed above). But if the Longman Corpus has such an "impoverished" set of genres, then it probably wouldn't be very applicable to evaluate the type of corpora that the authors prefer, where there is a much wider range of genres.

Comparison #2: British National Corpus (BNC)

As 1.1 above shows, the authors take COCA to task, because it does not contain the range of genres as the British National Corpus (BNC). What they fail to mention is that from at least 1994 until 2021 (nearly a complete generation), the **BNC was massively *un*-representative** of English, because it didn't have any texts – any at all – from the Web. Because the BNC was completed in 1994, some might say that "well, **the Web** (and certainly blogs) weren't a thing in the 1980s and early 1990s when the BNC was being created". *But that's just the point*. The BNC did a very good job representing genres from 1990, but less so for 1995 (when the Web was already becoming popular), or in 2000, . . . or in 2005 (when blogs existed), or definitely in 2010 . . . or 2015 . . . or 2020 (when the BNC Written update was still not available).[8]

It makes absolutely no sense to quibble about presumed issues with COCA in terms of representativity, and then ignore "the elephant in the room" – the fact that as Egbert, Larsson, and Biber (2020) and Egbert, Biber, and Gray (2022) were being written, the BNC was completely missing material from an *entire genre*, and one that has been a huge part of people's language input for the previous 20-25 years. Why is it that those who claim to care so much about

[8] BNC 2014 (released in 2021) does finally have some texts from the Web, but it is less than 5 million words of text, compared to 250+ million words of "web text" in COCA. In fact, there are actually more words of text from "annual business reports" in BNC 2014 Written than in texts from the Web. People in the "real world" undoubtedly read more material every day from material on the Web than they do material from "annual business reports". Why is BNC 2014 Written so "out of sync" with the "real world" in this case, and why (in a book that supposedly deals so much with "representativity") is there absolutely no mention of this by Egbert, Biber, and Gray?

"representativity" **have given the BNC a "pass" on this** for several decades – with only the promise that at some point it would all be fixed? Why is there no discussion of this incredibly important "missing genre" in these books?

And some might say, well it's all good now, since the written portion of the **BNC 2014 was released in 2021**, and it finally does have some web texts. But (as of the writing of this document in March 2023), the BNC Written 2014 is *still* not really publicly available. It is only available from within the proprietary LancsBox X software, but not via the Web or as downloadable files. Apparently there is also no publicly-accessible information on sub-genres (size, etc), and especially no publicly-accessible metadata for the 88,000+ texts in the written portion of the corpus. In essence, we have no idea what is in the web portion of BNC 2014 Written – or any other part of that corpus for that matter.[9]

As Egbert et al (2020) note, one of the first things that should be available for a corpus is information about what is in the corpus, including metadata about the texts. As discussed in 2.4 above, we have made this type of information available for COCA since the day it was released, and for every update since then.

So why is it that both Egbert, Larsson, and Biber (2020) and Egbert, Biber, and Gray (2022) are so critical of COCA for having supposedly "**poor documentation**" regarding the composition of the corpus and documentation about the corpus (arguments that of course aren't supported by actual data), and yet they completely fail to mention the very real problems of the BNC 2014 Written corpus in this regard? At least two of the authors of the 2022 book have great contacts with (and have had much collaboration with) the people at Lancaster University, who created BNC Written 2014. They could certainly have received information about the composition of the corpus update before they published their 2020 or 2022 books. Why no discussion at all of the very real shortcomings of that corpus, in terms of documentation?

4. A very narrow definition of "corpus"

Finally, we should note that the authors have a **very narrow definition** of what a corpus is. For them, a corpus is only the actual texts in the corpus – no more, no less. It does not include the corpus architecture or interface in any meaningful way. As we will see, this is a very problematic approach.

Maybe the best example of the problems inherent in ignoring the corpus architecture and interface would be the example of LancsBox X, which is the program that is required to use BNC 2014 Written. As is discussed in Section 5 **here**, LancsBox X is extremely limited, compared to the architecture and interface that is used at English-Corpora.org, or that of Sketch Engine, or even

[9] See Section 2.4 above for an update from January 2024, based on Version 3 of BNC 2014 Written, which was released in late 2023. While there is now some metadata for the texts in that corpus, more than two thirds of the texts still do not have enough metadata (such as the title of the article, or a URL for web texts) to find the text "in the real world".

that of CQPWeb. It was not until Version 3 of LancsBox X (released in late 2023) that it was even possible to search for collocates.

And 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 you can't search for collocates in that corpus. 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, 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, and Gray never consider **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. [10]

I would argue that a corpus is best seen as the texts in the corpus, along with the architecture and interface for the corpus. **The corpora from English-Corpora.org provide great architecture and interface, to go along with great textual data.** For example, they have an extremely wide range of search types, they are much faster than any other corpus architecture, their "association measures" and collocates and topics are great, they provides in-depth word-oriented features that aren't available anywhere else (including searching and browsing), and users can quickly and easily create Virtual Corpora and they can use the corpora data to analyze entire texts.

And talking just about COCA, **COCA is used more than any other corpus of English**, including the BNC. Each year it is used by hundreds of thousands (perhaps millions) of researchers, teachers, and students at thousands of universities throughout the world. Data from COCA has been used in more than 3,500 articles in the last few years. And its data (full-text, word frequency, collocates, and n-grams) has been used by most of the large tech companies in the world during the last 5-10 years, because of their belief that COCA does a great job of representing language in the "real world". It's hard to imagine that all of these teachers, students, researchers, and companies would use COCA so much, if it was as "poor" of a corpus (in terms of representativity and "empirical validation") as the authors claim.

To understand the importance of a "holistic" view of corpora (texts + architecture + interface), consider the following **analogy**. Suppose that a new owner of a professional sports team wants to have the **best "team" possible**. He buys up 10-15 great players and then announces that the

[10] Some might argue that it makes sense to separate "textual" corpora (which is what the authors focus on) from "overall" corpus, which would include architecture and interface. This is because there are some corpora (such as the BNC or the Brown family of corpora) where the same textual data might exist within multiple architecture / interfaces / websites. In this view, it would just "muddy the waters" to consider the two halves together. But again, why not even *consider* the point of view of the end user, for whom Corpus A in architectures / interfaces X, Y, and Z are almost different corpora, simply because the architecture and interface make available the data in such different ways? I understand the desire to be a "purist" about corpora – with an almost Platonic worship for just the texts themselves. But perhaps there should be a better balance.

"team" is ready to play the next day. The team, however, doesn't have any uniforms, any equipment, a stadium in which to play, or even a coach. And the team has never actually practiced together, and there is no strategy or any plays that have been designed around the special abilities of the players. When this "team" goes out to compete against other teams (which do have a coach, and which have practiced certain plays together), the "wonder team" may not perform as well as expected. And the fans will notice.

If a team has good players, and if the players work well together (just as the corpus texts, architecture and interface work well together), and especially if the team is at the top of its league in terms of wins and losses (just as COCA is used more than any other corpus in the world), then why not admit that this is a very good team? **Any other** unnecessarily **definition** of what a "team" is would look **very narrow and biased.**

Summary

The purpose of Egbert, Biber, and Gray (Cambridge, 2022) is to enlighten others about how to create quality, "representative" corpora. But it is interesting that none of the three authors has ever created a corpus that has been broadly accessible to or widely used by others, and which can therefore be readily examined by others.

Biber created the Longman Corpus in the 1990s, but only a small handful of researchers have ever had access to that corpus[11]. Gray has created a very nice corpus of academic English, but this is not publicly available. And Egbert[12] is working on the LANA Corpus, which is being actively discussed on Twitter, YouTube, and TikTok. If/when the LANA Corpus is released, it can be *evaluated* according to the exacting guidelines that Egbert has set forth.

I suspect that **if any of the authors had actual experience with creating corpora** that large numbers of researchers use on a regular basis (as is the case with English-Corpora.org), then they would have had a more *holistic* and *realistic* perspective of the issues that are involved in corpus design and creation.

[11] Biber also one of the creators of the ARCHER corpus. According to https://www.projects.alc.manchester.ac.uk/archer/, this has been made available to 14 "consortium universities". It is also supposedly available via the web, but I have not been able to find any links to a publicly-available online interface.

[12] Egbert and Biber were also involved with the CORE Corpus (Corpus of Online Registers of English). On page 48 of Egbert et al (2022), they give as a reference for CORE the book: [Biber, D., & Egbert, J]. (2018) *Register Variation Online*. Cambridge University Press]. Even though Egbert and Biber didn't mention me in that reference, I was actually the one that collected the texts for CORE, and I alone designed the architecture for the corpus, as well as the web interface. But again (as discussed in the previous section), the corpus architecture and interface don't really matter much for Egbert and Biber, and so that might explain their "oversight".