Designing and Evaluating Language Corpora by Jesse Egbert, Douglas Biber, and Bethany Gray (Cambridge Univ Press, 2022) is a **disappointing read**. All three of these authors have produced great research in the past. But in my opinion, this book is not one of them.

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 Egbert, Biber, and Gray in their book.

Before discussing more substantive issues, I would first note that in about a third of all **references to the corpora** from English-Corpora.org, the authors used the web address corpus.byu.edu, which ceased to be used more than four years ago (and 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. I only bring this up to suggest that perhaps some of the book was written by graduate students and then never adequately reviewed by Egbert, Biber, or Gray, which might explain some of the more serious problems discussed below.

1. Straw man arguments

The authors repeatedly argued that I have claimed X or Y (which I haven't), and then they proceed to "knock down" this "straw man".

1. 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

[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 spoken component of COCA is from *unscripted* conversations, albeit unscripted conversation on national TV and radio programs.

> 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 article), and you will see that I never claim that COCA "represents the entirety of contemporary American English". In this and many other articles, 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)[2]. But I have never claimed that COCA had more than five genres from 2008-2019 (spoken, fiction, magazine, newspaper, academic), and a few more since then (TV and movie subtitles, and blogs and web pages). And I have *never* claimed that COCA represents American English "in its entirety". (And incidentally, if COCA is problematic because it only contains these five genres, then so is Biber's Longman corpus, which only had spoken, fiction, newspaper, and academic).

Finally, a short note about a **serious double standard** regarding different corpora. Certainly any modern corpus that fails to have any texts at all from the Web is completely unrepresentative of the language. But this was the case with the **BNC** from 1994 until very recently (and in a certain respect, it's still an issue[3]). 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 in 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

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 these books 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 last 20-25 years. Why is it that those who claim to care so much about genre balance and representativity **have given the**

[2] In fact, COCA is based on the Longman Grammar of Spoken and Written English that was created by Biber and others in the 1990s. That corpus had spoken, fiction, newspaper, and academic. When COCA was first released in 2008, it had those same five 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 not then, nor has there ever been, 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).

[3] The written portion of BNC *2014*, which does contain data from the Web, was supposedly released in late *2021*, but as of March 2023 it is still not really publicly available. It is only available via 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. As Egbert et al (2020) notes, 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 (such as we have made available for COCA since the day it was released, and for every update since then). So without any information on the actual texts in the web portion of BNC 2014 Written, we will ignore it for the time being.

BNC a "pass" on this for at least 15-20 years now – with only the promise that at some point it would all be fixed?

(Note that COCA does have material from the Web -- 125 million words from blogs, and 130 million words from other web genres. All of the 187,737 texts are nicely sub-categorized into sub-genres (thanks, Serge Sharoff!), and with rich metadata on each of the texts.)

1.2 Representativity #1. Another example of a "straw man argument" is when they 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. They are creating a straw man by putting words into my mouth, and that is poor academic scholarship.

1.3 Representativity #2. A third example of "straw man" arguments 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 unprofessional for Egbert, Biber, and Gray to include this quote in their book (and thus tacitly endorse the sentiment). 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 retired and after all of my other publications), to respond to previous critiques of how COCA is supposedly so "unrepresentative".

2. Mischaracterizations

Closely related to the practice of "straw man arguments" is mischaracterization. This is when the authors try to convince others that someone believes X (when s/he doesn't), or (in the case of the Egbert et al book), that a corpus is like X, when it is not.

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 bizarre 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 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 – absolutely at random – from COCA[4]. 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 "qualitative inspections".

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 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

[4] 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.

make up this information about COCA, and then completely ignore the very real issues of "spotty updates" with the BNC?

2.3 COHA genres. At times the authors intuit what was in the mind of the corpus creator when s/he created the corpus, and then to impute less than admirable motives. 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.

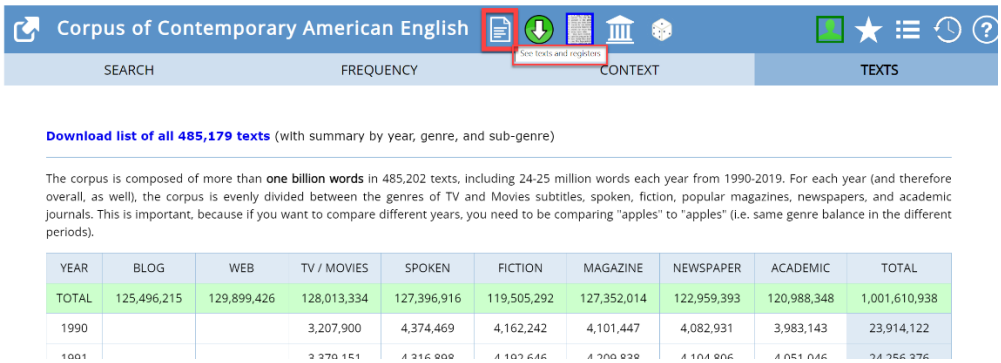
Actually, COHA (the Corpus of Historical American English) was meant to greatly expand on the Brown family of corpora. In the Brown corpora, 50% of the texts were from fiction and 50% from non-fiction. COHA was designed the same way. And actually, collecting the 223+ million words of texts in the nearly 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 from the 1820s to the 1980s, and that took many hundreds of hours. If we were looking for "convenience", we would have hard hardly any fiction at all.

Rather than guessing what was in my mind as I created a corpus, the authors might consider sending a quick email to ask. I have worked very closely with two of the three authors in the past, and they are more than welcome to just reach out and ask – that's how things should work in academia.

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 any of the other 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 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"?

Finally, a small issue about **double standards**. The written portion of **BNC *2014*** was supposedly released in late *2021* (while the Egbert book was still being written). But to date (as of March 2023), there is 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. Why was there no uproar from Egbert, Biber, and Gray about this egregious lack of documentation about what is in the BNC 2014 Written. Why the double standard?

3. "Empirical evidence". 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 deceptive 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. Second, COCA has the same four genres as the **Longman grammar** that was created by Biber in the 1990s (plus magazines). If COCA **isn't representative** of anything, because it only has these genres, then neither is the Longman corpus, or the resulting Longman Grammar of Spoken and Written English. Following this logic, that book might be a nice discussion of English grammar, but it doesn't really "represent" actual phenomena from English in any meaningful way.

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 "academic" than COCA academic. **I provided detailed data** to show that in all three cases (#1-3 on pages 2-3) the data showed that COCA academic is actually more "academic"

than BNC academic. I emailed the results of these studies to the three authors, and they emailed to say that they received it. It is interesting, however, that the authors chose not to include any of these three supposed pieces of evidence in their 2022 book, perhaps because people would have then found the very data that they claim does not exist.

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 is publicly available, and which can actually be 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[5]. Egbert is working on the LANA Corpus, which is currently being discussed on Twitter, YouTube, and TikTok, and which will hopefully see the light of day at some point. Gray has created a very nice corpus of academic English, but this is not publicly available.

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

In summary, all three researchers (Egbert, Biber, and Gray) are great scholars overall, and that is what makes this particular book such a disappointment.

[5] 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.