REPRESENTATIVITY (GENRES) See review of Egbert, Biber, Gray (2022)

The corpora from **English-Corpora.org** are probably the **best** source for looking at variation in English – whether historical, dialectal, or genre-based variation. As far as insights into genre-based variation, the Corpus of Contemporary American English (COCA) is the only corpus available from any source that:

1. contains texts from a wide range of genres (spoken, fiction, newspaper, academic, Web, etc)
2. is large (it contains one billion words), and
3. is recent (1990-2019), with essentially the same genre balance in each of these 30 years.

The British National Corpus[1], for example, has good genre distribution (although see below, regarding "missing" genres). But it is neither large (it has only 110 million words; about one tenth the size of COCA), nor is it recent (90% of the texts are from more than 30 years ago).

In terms of genres, some people **(more)** have criticized COCA for having poor "representativity", compared to the BNC (using the definition of "representativity" as a measure of how well the corpus represents language in the "real world"). We will show here that these arguments are wrong – COCA has at least as good of representativity as the current, publicly-available BNC (and actually better, we would argue).

The following are the major genres in COCA and the BNC. The table below is for the BNC from the early 1990s.[1]

COCA

Genre	# texts	# words
TV/Movies	23,975	129,293,467
Spoken	44,803	127,396,932
Fiction	25,992	119,505,305
Magazine	86,292	127,352,030
Newspaper	90,243	122,958,016
Academic	26,137	120,988,361
Web/Blog	98,748	125,496,216
Web/Genl	88,989	129,899,427
TOTAL	**485,179**	**1,002,889,754**

BNC (1994)

Genre	# texts	# words
Spoken	909	10,334,947
Non-Acad	501	15,429,582
Fiction	464	16,194,885
Magazine	211	7,376,391
Newspaper	518	10,638,034
Academic	534	16,634,076
Misc	916	21,011,396
TOTAL	**4,053**	**97,619,311**

[1] In this document we discuss the BNC 1994 release, not the BNC 2014 release. The written portion of BNC *2014* was supposedly released in late *2021*, but as of **July 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 still (**July 2023**) 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). Until something similar is available for BNC 2014, we will limit our discussion to the portion of the BNC that is truly publicly available – the 1994 data. If you are aware of a change in status for the BNC 2014 *Written* data, please email us and we will change this document accordingly.

In terms of **fiction, newspapers, and magazines** (and including the "Non-Acad" category in the BNC), these are not very controversial in terms of texts. It is fairly easy to collect texts from these genres, and both corpora do a good job in terms of the distribution of sub-genres, such as Sports, Finance, or Entertainment for Newspaper.

In terms of "**Spoken**", the BNC does a very good job – both in terms of context-determined texts (e.g. religious, educational, or judicial), as well as conversation. COCA-Spoken does not have as wide a range of text types for spoken – it is all from *unscripted* conversation on national TV and radio programs (although it is still quite informal).

But as we have argued, the TV/Movies texts more than make up for this. As our recent article in the *International Journal of Corpus Linguistics* (IJCL) shows (Davies, 2021 (pages 21-25); see also this page), the TV and Movies texts in COCA are linguistically at least as "spoken" as the "true spoken" in the BNC – both in grammatical and lexical terms.

In terms of "**Academic**", Egbert, Larsson, Biber (2020; hereafter ELB) has an extended case study (the longest case study in the book) regarding representativity, to supposedly show how a well-designed corpus (for them, the BNC) has better representativity than a "poorly designed" corpus (for them, COCA). In an attempt to prove this, they argue that COCA Academic (COCA-Acad) is not nearly as "academic" as BNC-Acad. They base their conclusions on three supposed pieces of evidence, none of which is supported by actual data.

1. They argue that **linking adverbials** (*however, thus, therefore*, etc) are a feature of academic English, but that these are not as frequent in COCA-Acad as in BNC-Acad. They therefore argue that COCA-Acad is somehow "defective".

Our response: ELB is not comparing "apples to apples", in terms of time period or dialect. Data from COHA shows that these linking adverbials were much more common in American English generally 30-35 years ago (when the BNC was created) than they are now, and COCA shows a sharp decline from 1990-2019. In addition, in GloWbE these adverbials are more common generally in British than American English. When we take into account the historical changes and the dialectal differences, these linking adverbials are actually *more common in COCA*-Acad than BNC-Acad.

2. They also look at **nominalizations** (**tion, *ism, *ence*, etc), which are also a feature of academic English. The frequency in BNC-Acad is 35,613 tokens (per million words) vs 33,636 in COCA-Acad. Again, they argue that COCA-Acad is somehow "defective".

Our response: Is this difference of just 6% really significant? Also, nominalizations have decreased generally in American English in the 30-40 years since the 1980s, when the BNC was created (COHA, COCA). If that overall change in American English (not just in academic) is taken into account, nominalizations are actually *more common in COCA-Acad* than in BNC-Acad.

3. Perhaps the strangest of the three pieces of "evidence" for the supposed weakness of COCA is their argument that **one word** – ***intestine*** –is more common (per million words) in BNC-Acad than

in COCA-Acad. They therefore argue that on lexical grounds, the BNC has better / more representative "academic" than COCA.

Our response: Frankly, it is silly to focus on just one word *(intestine*) and – based on that one word– to suggest that BNC-Acad is somehow better.

In order to look at this systematically, we took all 568 of the words in the Academic Word List (AWL; Coxhead, 2000) – a "neutral" word list that would favor neither the BNC nor COCA – and compared the frequency of each of these words in the AWL list in both COCA and the BNC. The results are found in this spreadsheet (ZIP, TXT).

If we look at all 568 words in AWL, there are 310 words that are more frequent in COCA; in other words, the normalized frequency (per million words) is at least 1% more in COCA than the BNC. There are 254 words that are more frequent in the BNC; i.e. where COCA is at least 1% less frequent than the BNC.[2]

So using the criteria of word frequency -- but going beyond the one word (*intestine*) that Egbert, Larsson, and Biber use – to look systematically at *all words* in the AWL – the relative frequency of academic words is higher in COCA-Academic than in BNC-Academic.[3]

The "elephant in the room" [1]

But as long as we're talking about representativity – how well a corpus reflects language in the "real world" – let's look at the most **glaring problem with the BNC** – its complete lack of texts from **blogs and the Web in general**. 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[1]).

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 **(more)** 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 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?

And even beyond this general issue of no Web / blog texts in the BNC, there is the issue of **"missing lexis" in other genres** (magazines, newspapers, etc.) for the last 20-30 years. Go ahead and search

[2] But a difference of 1% either way probably doesn't matter much. So let's limit it to words that are at least 50% more frequent in either COCA or the BNC (e.g. 12 tokens per million words in COCA, but 7 tokens pmw in the BNC), but which are "still in the ballpark" in terms of frequency (in other words, not more than 20x as frequent in one of the two corpora). In this case, there are 142 words (yellow) that are more frequent in COCA, and 97 words (blue) that are more frequent in the BNC.

[3] To be fair, though, the authors' confusion about lexical issues ("content" words) may be due to the fact that virtually all of their work on representativity has dealt with grammatical differences between genres. As a result, they may not be familiar with looking at words (but which are, of course, part of language as well).

the BNC for words relating to technology or societal and cultural change since the 1990s in these other genres – they're just not there. As long as researchers simply limit themselves to grammatical differences between genres (which is the approach of many "experts" in corpus design), there's no problem. But if they were to look at lexis (words) as well, then all of the sudden the BNC massively fails to be "representative" of the language during the last 20-30 years. And yes, words are part of the language too.

A more systematic bias and problem

This bias when it comes to looking at "representativity" in COCA and the BNC reflects a more systematic bias among corpus linguists. Stated very bluntly, the bias is that "if another corpus doesn't look like the BNC, then it's somehow "defective" and "suspect". According to these researchers **(more)**, different isn't just different – it's "bad". And here's the worst aspect of this attitude – ***it seriously limits progress in the field of corpus linguistics***.

The BNC was a great corpus for the time in which it was developed (30-35 years ago). This was due in large part to a budget of millions of dollars, amazing support from organizations like the Oxford University Press, and a large team of researchers. But that situation is **simply not the reality** for most corpus creators today. These other researchers don't have millions of dollars; they don't have an institution that is willing to give them tons of copyrighted texts; and they probably don't have a large research team.

Are the creators of these other corpora just supposed to pack their bags and go home, unless and until they have the amazing institutional support of the BNC? If so, we can expect very few additional corpora in the next 15-20 years, in the way that we have (quite bluntly) had **very few "genre-balanced" corpora during the last 15-20 years**. People are just too intimidated by pressure from the BNC-philes, who demand that new corpora look and act just like the BNC.

Fortunately, a few other corpus creators have refused to be intimidated by the BNC-first crowd. **Sketch Engine** has created a number of incredibly useful corpora, which have had a huge impact on lexicographical research and practice during the last 15-20 years. And yet if you go to a conference like ICAME (one of the two most important conferences for English corpus linguistics), Sketch Engine is almost completely ignored there – simply because these corpora don't look like the BNC, and therefore the studies are not even accepted for presentation at ICAME. (Or maybe they're not even submitted in the first place, since it's clear that ICAME isn't a friendly place for such presentations.)

And fortunately, we weren't intimidated by this BNC-first crowd either, when we created COCA in 2008 and as we have updated it since then. COCA was created by just one person, in just 3-4 months, without any "free" texts from publishers, and with a budget of approximately $0. The message here is that you don't need tons of money and a huge research team and lots of support from publishers. There's **no reason to be intimidated by others** and to not even try.

And COCA is now the most widely-used online corpus in the world, and it is joined by many other corpora at English-Corpora.org. These corpora don't look exactly like the BNC, and that's OK. And we expect and hope that 5-10 years from now, there will be other corpora that are even better than

COCA, and that's OK too. **That's how a field progresses** – continual change and improvement – and not by being tied down and limited by the way things were done 30-35 years ago.

"The proof of the pudding is in the eating"

And who should be final arbiter of the "usefulness" of a corpus anyway? Actual users of the corpora, or just corpus linguists in their ivory towers? (And yes, I was a corpus linguist for many years, and I published as much as anyone.) If we have to choose, we'll go with actual users.

As mentioned, **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 provided such "poor" data.

And although the following argument doesn't have to do with "representativity" per se, it definitely has to do with "**usability**" – which is a **pretty big deal for actual users** of corpora (even if it's not for "corpus theoreticians"; **more**) [4]. COCA provides an extremely wide range of search types, it is much faster than any other corpus architecture, its "association measures" and collocates and topics are great, it 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 corpus data to analyze entire texts.

The bottom line is that **COCA is the only large, recent, representative corpus of English**, and it is joined by many other extremely useful corpora at English-Corpora.org. Taken as a whole, the corpora allow more insight into variation in English and more features for teachers, learners, and researchers than any other collection of corpora anywhere in the world.

---

[4] It seems quite strange that the book *Designing and Evaluating Language Corpora* by Egbert, Biber, and Gray (Cambridge Univ Press, 2022) should completely **ignore the perspective and experience of the actual users** of corpora, since "evaluation" is presumably one of the goals of the book. This is probably due to some corpus linguists' tendency to separate the "pure" textual corpus (which is an ideal, Platonic entity – unsullied by contact with real humans) from the actual corpus that people use, and which takes into account the architecture into which the texts are placed, and the interface with which the humans interact. I believe that this approach is misguided. The best corpus in the world (looking only at the texts in the corpus, and from a strictly purist point of view) can be rendered almost completely useless by a bad architecture and interface. This is kind of like hiring great players for a sports team (the best talent, like just the right texts in a corpus), but then never bothering to buy equipment, or to practice as a team, or to design actual plays or create a strategy for the team to work together (which is like the architecture and interface for a corpus).
One possible explanation for this skewed focus is that **none of these three authors has ever created a corpus that has been used by a large number of other researchers**, and they are therefore perhaps unaware of some of the real-world, practical issues that are involved with corpus design and creation.