The Coronavirus Corpus
Design, construction, and use

Mark Davies
Brigham Young University

This paper discusses the creation and use of the Coronavirus Corpus, which is currently (March 2021) 900 million words in size, and which will probably be about one billion words in size by May–June 2021. The Coronavirus Corpus is a subset of the NOW Corpus (News on the Web), which is currently about 12.1 billion words in size and which grows by about two billion words each year. These two corpora are updated every night, with about 6–10 million words for NOW and 2–3 million words for the Coronavirus Corpus. The Coronavirus Corpus allows users to see the frequency of words and phrases over time (even by individual day), and users can find all words that are more frequent in one time period than another. Users can also see the collocates for words and phrases, and compare the collocates to see what is being said about particular topics over time.

Keywords: corpus design, NOW corpus, text archive, Coronavirus, COVID-19

1. Introduction

One of the challenges facing corpus creators is the need to make corpora relevant to current events – in politics, science, entertainment, or any other field of interest. The COVID-19 pandemic is a very good example of this. There are very few structured corpora that have been updated every month, week, or day during the last two or three years, which allow us to compare the “COVID-19” period to other recent time periods, or to see changes within different periods of the COVID-19 timeframe itself. One option is to simply use “text archives” (such as online newspapers, databases such as Lexis-Nexis, or even the entire Web itself) to look at very recent changes. But in this case, the “corpus” may not allow a full range of searches – such as word frequency, collocates, or concordance lines (see
Davies, 2015). Ideally, the corpus would be (i) very recent (ii) updated at least every few weeks, and (iii) allow a wide range of searches.

This paper will consider the design and creation of one of the very few structured corpora that fulfill all three requirements – the Coronavirus Corpus (https://www.english-corpora.org/corona), which is part of the suite of corpora from English-Corpora.org. In Section 2, I will consider the NOW Corpus, which is the underlying corpus on which the Coronavirus Corpus is based. Section 3 discusses how a subset of the NOW Corpus texts are transformed into the Coronavirus Corpus every night. Section 4 will provide a number of concrete examples of how the Coronavirus Corpus can be used to look at the impact of COVID-19 since early 2020 – culturally, economically, and even linguistically.

2. Creating and using the NOW Corpus

The NOW Corpus (“News on the Web”; https://www.english-corpora.org/now) – which was first released in 2015 – is (as of March 2021) composed of approximately 12.1 billion words, and it grows by about 6–10 million words each night, 200–250 million words each month, or 2–3 billion words each year. The NOW Corpus includes texts from the same twenty English-speaking countries as the GloWbE Corpus (https://www/english-corpora.org/glowbe; see Davies & Fuchs, 2015).

From 2015 through mid-2019, the NOW Corpus was based on links from Google News. Every hour of every day, Google News was queried (using a simple search, like all texts with the words the or to, which would find essentially all articles) to find online newspaper and magazine articles that had been released in the previous 60 minutes. This search would be repeated for each of the twenty different English-speaking countries, and the URLs from Google News would be stored in a relational database, along with all of the relevant metadata – country, source, URL, etc. Every night, scripts would then download the 15,000–20,000 articles, clean them (such as extracting the text from the raw HTML with tools like JustText), tag them (using CLAWS 7), remove duplicates (using a proprietary method involving 11-grams), and then merge the texts into the existing NOW Corpus.

In mid-2019, this procedure was modified. Changes in Google News had made it increasingly difficult to retrieve the 15,000–20,000 URLs each day without being blocked. As a result, I moved to Microsoft Azure Cognitive Services to collect the URLs. Every day, I retrieve a list of new magazine and newspaper articles (from any source) from the previous 24 hours, for each of the twenty English-speaking countries. In addition, each day I query Bing to find new articles (from the previous 24 hours) for 1,000 specific websites (the websites with the most
articles in NOW through mid-2019). Of course there are many duplicate URLs between these two sets of searches, but since everything is in a relational database, I can easily eliminate these duplicates.

The NOW Corpus provides a wide range of searches. Users can search by word, phrase, substring (e.g. *icity), wildcards (as * as), lemma (FIND out whether = find / finds / finding / found out), part of speech (CONJ PRON BE like, | ” = “and he was like,” “but I’m like ‘”), synonym (=CLEAN the NOUN = “cleaned the car”, “rinsing the dishes”), customized wordlist (BUY * @CLOTHES = “bought some pants”, “buys expensive shoes”), and more.

Because the corpus is updated every day, many researchers use the corpus to find the frequency of words, phrases, or syntactic constructions over time. For example, users could find the frequency of virtue signal*, Brexit, gig economy, or trigger warning since 2010. It is even possible to see the frequency in 10-day increments. For example, users could see that the phrase fake news spikes immediately (within 1–2 days) after the 2016 US presidential elections.

The architecture of the NOW Corpus also allows users to quickly and easily compare the results in one section (e.g. a particular time period) to those of another section (or time period) (see Davies, 2017, 2018 for many more examples). For example, users could find two-word strings composed of climate + NOUN that are more frequent in 2019–2020 compared to 2010–2012, such as climate emergency, climate breakdown, climate strike, or climate warriors). Another example might be all new phrases with smart + NOUN that are at least 20 times as frequent in 2017–2020 as they were in 2010–2013 (if they occur back then at all): smart speaker, smart pole, smart airport, smart workplace, smart condom, smart coating, smart gas, smart doorbell, smart shower, smart park, smart waste, and smart fence. As can be seen, many of these provide evidence for the rise in “smart devices” during this time. In summary, the NOW Corpus is currently the only large (10+ billion words) “monitor corpus” of English that is updated every day, and which offers a wide range of searches to look at changes in the language.

3. Creating the Coronavirus Corpus

This section introduces the concept of Virtual Corpora in English-Corpora.org (3.1), followed by an explanation of the Coronavirus Corpus as a stand-alone corpus (3.2).
3.1 Virtual Corpora in NOW

With any of the corpora from English-Corpora.org, users can quickly and easily create 'Virtual Corpora', based on words in the texts or metadata about the texts. For our purposes in this paper, users could easily create Virtual Corpora dealing with Coronavirus / COVID-19. For example, as shown in Figure 1 (left), users could create a Virtual Corpus of those texts where the word *COVID-19* occurs at least 10 times in the text, sorted by the frequency of *COVID-19* (per 1,000 words) in the text. Or, as shown at the right in Figure 1, they could create a Virtual Corpus composed of texts from the UK (Great Britain, below) from 1–31 May 2020, which contain the word *COVID-19* in the text.

Users can then select from among the matching texts and can click on any text to see the original article online. And once they have created several Virtual Corpora, they can add to, delete from, or move texts between these corpora, as shown in Figure 2.

---

**Figure 1.** Creating Virtual Corpora in the NOW Corpus

**Figure 2.** Virtual Corpora in the NOW Corpus
Most importantly, after creating a Virtual Corpus, users can search within it, as though it were its own corpus. For example, they could find collocates for a given word or see concordance lines for a word—all limited to data from their COVID-19 Virtual Corpus. They can also quickly and easily generate a list of keywords from the Virtual Corpus, as shown in Figure 3.

**Figure 3.** Keywords from a “coronavirus” Virtual Corpus in the NOW Corpus

### 3.2 A stand-alone Coronavirus Corpus

This paper deals with the creation of the Coronavirus Corpus. But if it is already possible to create Virtual Corpora related to the Coronavirus / COVID-19 in NOW, why create a stand-alone corpus? First, while the architecture of English-Corpora.org allows users to easily create Virtual Corpora and then search within those corpora, it is not really possible to create a Virtual Corpus within a Virtual Corpus. So it would be difficult to limit searches to (for example) texts dealing with “maskers / anti-maskers” in a COVID-19 Virtual Corpus. Second, the NOW Corpus is currently about 12.1 billion words in size, and it will probably be 13–14 billion words in size by October 2021. Even when users limit their search to a much smaller Virtual Corpus, there is still a lot of overhead with the much larger complete corpus. By creating a stand-alone corpus dealing just with COVID-19, the corpus will be much smaller, and searches will be faster.

In early April 2020, as the COVID-19 pandemic had really begun to set in, I decided to create a stand-alone Coronavirus Corpus from the underlying NOW data. There were virtually no tokens of *coronavirus* in NOW before 1 January 2020 (less than 30 tokens from July–December 2019), but there were several thousand in January 2020, and nearly 20,000 tokens *per day* at the highest point in March 2020. So by April 2020, it seemed an opportune time to begin work on the corpus.
In order to extract data from the NOW Corpus for the stand-alone Coronavirus Corpus, I found texts from January–April 2020 that fulfilled one of the two conditions:

i. The text had at least three occurrences of the words \{coronavirus, COVID, or COVID-19\}.


The title words in (ii) were taken from keyword lists from texts where \{coronavirus, COVID, or COVID-19\} occurred at least three times in the text, as in (i). While there are undoubtedly some cases where a title would have one of these words (e.g. spread: the oil spill has spread to Alaskan beaches), informal tests of the data show that the keyword lists work very well. In 100 randomly selected texts in each of April, May, and June 2020, between 96% and 99% of the texts with at least one of these words in the title dealt (to at least some degree) with COVID-19, rather than being completely unrelated (as with the hypothetical example of an Alaskan oil spill). If humans manually approved each text for the Coronavirus Corpus every day, it would undoubtedly be an even “cleaner” corpus. But with 10,000–20,000 texts every day, that is probably not realistic.

Table 1 shows the size of the Coronavirus Corpus and the NOW Corpus by month from January 2020 through mid-March 2021 (this article was revised on 14 March 2021). As can be seen, very few of the texts in the NOW Corpus in January and February dealt with COVID-19, but this skyrocketed to about 40–50% of the articles in March through May, and has decreased somewhat since then.

As the table shows, at the time of writing, the Coronavirus Corpus has about 900 million words of text in about 1,192,413 texts from about 9,500 distinct websites. At the current rate of growth, the corpus should be about one billion words in size by May–June 2021.

Figure 4 shows the distribution by country. It shows that by far the largest portion of the corpus (about 44% of the entire corpus) comes from texts from the United States, followed by Great Britain, India, and Canada (7–9% each), and then Australia, Ireland, South Africa, Nigeria, and New Zealand, with about 14% from the other 11 countries in the corpus. This is a function of the texts in the
## Table 1. Size of NOW and Coronavirus corpora, by month (*through 14 March 2021*)

| Month | NOW Corpus | | | Coronavirus Corpus | | | | | | |
|-------|------------|------------|------------|-------------------|------------|------------|------------|------------|------------|
|       | # sites    | # texts    | # words    | # sites           | # texts    | # words    | # sites           | # texts    | # words    | % COVID-19 |
| 20–01 | 8,996     | 412,132    | 219,072,222 | 2,311             | 12,574     | 7,340,233  | 3.4%            |
| 20–02 | 8,445     | 329,921    | 180,724,244 | 2,637             | 24,178     | 14,494,437 | 8.0%            |
| 20–03 | 8,872     | 394,335    | 238,817,335 | 5,936             | 144,482    | 99,990,266 | 41.9%           |
| 20–04 | 9,142     | 361,876    | 212,275,353 | 6,227             | 160,599    | 107,974,419 | 50.9%           |
| 20–05 | 8,294     | 350,007    | 233,687,513 | 5,251             | 131,520    | 97,812,229 | 41.9%           |
| 20–06 | 2,429     | 352,956    | 240,053,628 | 1,847             | 104,294    | 83,275,665 | 34.7%           |
| 20–07 | 3,167     | 353,919    | 234,588,055 | 2,121             | 103,306    | 78,413,612 | 33.4%           |
| 20–08 | 3,824     | 373,629    | 255,999,268 | 2,034             | 89,230     | 74,161,690 | 29.0%           |
| 20–09 | 3,049     | 321,556    | 220,284,388 | 1,968             | 69,506     | 57,611,101 | 26.2%           |
| 20–10 | 3,555     | 298,449    | 198,094,127 | 2,061             | 68,357     | 57,135,670 | 28.8%           |
| 20–11 | 3,239     | 293,187    | 185,045,225 | 2,262             | 61,952     | 49,283,628 | 26.6%           |
| 20–12 | 3,375     | 301,602    | 189,600,860 | 2,420             | 65,883     | 50,801,864 | 26.8%           |
| 21–01 | 3,379     | 329,399    | 206,606,227 | 2,428             | 74,832     | 56,971,166 | 27.6%           |
| 21–02 | 3,789     | 288,012    | 184,824,841 | 2,411             | 56,103     | 45,191,176 | 24.5%           |
| 21–03*| 2,787     | 124,286    | 82,224,977  | 1,845             | 25,597     | 21,468,673 | 26.1%           |
| TOTAL | 14,112    | 4,900,266  | 3,081,498,263 | 9,498             | 1,192,413  | 901,925,829 | 29.3%           |

NOW Corpus, which is in general a function of the links provided by Microsoft Azure Cognitive Services.

## 4. Using the Coronavirus Corpus

Because of its very large size and its granularity (the ability to see changes day by day), the Coronavirus Corpus can be used to look at a wide range of phenomena in ways that would probably not be possible with any other corpus. At the most basic level, the corpus shows the frequency of any word or phrase in ten-day increments from 1 January 2020 to the present. For example, the corpus shows (Figure 5) that Wuhan (China) (where the first cases of COVID-19 were reported) is frequent very early on, but that it declines dramatically by early March 2020, and that it has stayed at that low rate since then. The string hoard* (Figure 6) was very frequent in March 2020 (as the pandemic hit major Western countries and people were concerned about supply shortages). After having things “locked down” for a month or so in the US since March 2020, people were wondering by late April 2020 if perhaps Sweden (Figure 7) was not a better model, with more limited closures. And already by April 2020 people were talking about re-
opening (Figure 8) businesses, schools, and the economy. One of the most interesting charts is for flatten* the curve (Figure 9), which increased sharply in March 2020 as governments set this as a primary goal, and then the frequency “flattens” out very nicely from April 2020 on.

Figure 5. Frequency of Wuhan over time.

1. From January–December 2020, the corpus interface showed (at the top level) the frequency in ten day increments, and then (at a more detailed level) by individual day. The figures in this paper follow that format, since that was the format when the paper was initially written.
in October 2020. However, this format would have resulted in almost seventy (ten day) time periods by the end of 2021 (Jan 2020–Dec 2021), and this would have become increasingly unmanageable via the web interface. As a result, in early 2021 the web interface was changed to show (at the top level) frequency by month, and then by individual day.
In addition to seeing the frequency in ten-day increments, users can also see the frequency day by day. For example, many people in the US will remember March 9–15 as the week in which the seriousness of the pandemic became apparent, and that life as we knew it was going to be radically changed. The corpus shows that social distance* increased dramatically from March 1–10 to March 11–20 and then to March 21–31 (Figure 10), and it also shows an increase each day from March 11 through March 17 – the “week that changed everything”. Of course, the corpus would allow us to see the figures day by day during March 2020, in any of the 20 countries in the corpus.

![Figure 10. Frequency of social distance over time](image)

Users can also limit concordance searches to particular time periods. For example, Figure 11 (for the search re*open*) shows the relative naiveté of people in March 2020, as they viewed the pandemic as a short-term problem, and were looking forward to re-opening schools and businesses in April 2020.

![Figure 11. KWIC lines for re*open* in March 2020](image)

By April 21–May 20 (Figure 12), the corpus shows people thinking more long-term, and realizing that the shutdown had been much more severe than initially expected. The corpus also shows the collocates for any word or phrase. For example, Figure 13 shows the top noun collocates of BAN_v (all forms of ban as a verb). It is also possible to show the collocates in ten-day increments. For example, Figure 14 shows the verb collocates of MASK_n (all forms of mask as a noun) every ten days from March 21–31 through July 21–31. As one can see, early on there was emphasis on simply getting enough masks (making, donated, distributed). But
then in about June 2020, the discourse changes to having people actually wear the masks that they already have (wear, worn, requiring, require, recommended, encouraged).

**Figure 12.** KWIC lines for re*open* in April–May 2020

**Figure 13.** Noun collocates of BAN_v

**Figure 14.** Verb collocates of MASK_n, by 10-day period
Users can also search for collocates in a particular time period. For example, Figure 15 shows the noun collocates of re*open* from March 11–May 20, while Figure 16 shows the noun collocates from August 1–September 30.

Figure 15. Noun collocates of re*open* in March–May 2020

Figure 16. Noun collocates of re*open* in August–September 2020

Note the emphasis on businesses and the economy in general in the earlier period, and the emphasis on re-opening schools (and planning out re-openings, perhaps better than before) as the school year approached in August and September. Users can also see the results side-by-side, as in Figure 17 (the higher the “ratio” column, the more it is used in that time period than in the other, and a score less than 1.0 means that it is used less in that period than in the other).

Figure 17. Noun collocates of re*open* in two different time periods
A comparison of collocates can provide useful insight into what is being said about particular topics over time. For example, Figure 18 shows adjectival collocates of China from Feb 1–March 10 (left) vs June 1–July 31 (right). Note that early on, the newspaper and magazine articles simply talk in matter-of-fact terms about the existence of COVID-19 and measures that were being taken in China (postponed, self-isolate, shuttered, affected, flu-like, epicenter). By June–July, however, articles about China were much more negative in tone, and focused on disputes with China (territorial, disputed, nuclear, expansive) or they simply had a more negative tone overall (accountable, unlawful, unfair, unjustified).

Figure 18. Adjectival collocates of China in two different periods

Figure 19 shows noun collocates of MASK_n in March 11–May 10 (left) vs July 1–August 31 (right). There is little of interest in March–May, but by late summer people were rebelling against the wearing of masks, and we see collocates like freedom and Republican (presumably anti-mask groups) and ordinances, mandate(s), and bylaws (as the government was trying to encourage people to wear masks).

Figure 19. Adjectival collocates of MASK_n in two different periods

In Figures 13–19 we focused on collocates of given words and phrases. But to return to the general issue of word frequency, the corpus allows us to find interesting examples of how life changed over time, simply by searching for all words that were used more after the start of the pandemic than before. For example, Figure 20 is the result of searching for all adjectives that are much more common in January 1–March 10 (left; pre-pandemic) compared to March 21–May 20 (right; after the pandemic really hit Western countries). Again, there is little of interest...
in the “pre-pandemic” earlier period (left), but nearly all of the adjectives that are more common in the later period (right) do relate in some way to the pandemic.

Finally, just as we could create Virtual Corpora in the NOW Corpus, we can also do the same in the Coronavirus Corpus. For example, we could create Virtual Corpora from “progressive” vs “conservative” newspapers, or texts from different countries, or different time periods, or any combination of these (e.g. “progressive” British newspapers in March 2020 and then later in July–August 2020). We can also just find the texts that use a particular word or phrase the most, and then select from those texts to create a Virtual Corpus.

For example, we might want to see what the texts say about the approach taken by Sweden, which has been quite different from most other countries (very limited lockdowns, but more focus on vulnerable populations). In just a couple of seconds, we can create a [Sweden] Virtual Corpus, and then look for keywords in these texts, as with the adjectives in Figure 21.

Not all of these are immediately obvious, but once we take a more qualitative approach by looking at the keywords in context, most of these make more sense. For example, relaxed, stringent, voluntary, and mandatory refer to the approach that Sweden has taken. The phrase per capita compares death rates in Sweden and other countries (and this is also shown with misleading, comparable, and controversial). A word like successful may refer to different sides of the same coin (i.e. has Sweden’s approach in fact been more successful overall than that of other countries), as in Figure 22 (where, again, all of the concordance lines are taken just from our [Sweden] Virtual Corpus).
The COVID-19 pandemic will likely be something that we will remember and discuss for years to come. The topics for future research will include a wide range of issues that were affected by the pandemic – health, education, society, culture, the economy, and more. As we have seen with many concrete examples in this paper, the Coronavirus Corpus – because of its size, granularity, and corpus architecture – allows researchers to look at a wide range of phenomena related to the pandemic, in ways that would be difficult or impossible to study with any other resource.

All rights reserved
References


Address for correspondence

Mark Davies
Department of Linguistics
Brigham Young University
Provo, UT, 84602
USA
mark_davies@byu.edu

Publication history

Date received: 22 October 2020
Date accepted: 24 March 2021
Published online: 3 May 2021