SHORT PAPER

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

https://doi.org/10.1075/ijcl.21044.dav | Published online: 3 May 2021
International Journal of Corpus Linguistics 26:4 (2021), pp. 583–598. issn 1384-6655 | e-issn 1569-9811

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 ("**N**ews **o**n the **W**eb"; 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 JusText), 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.

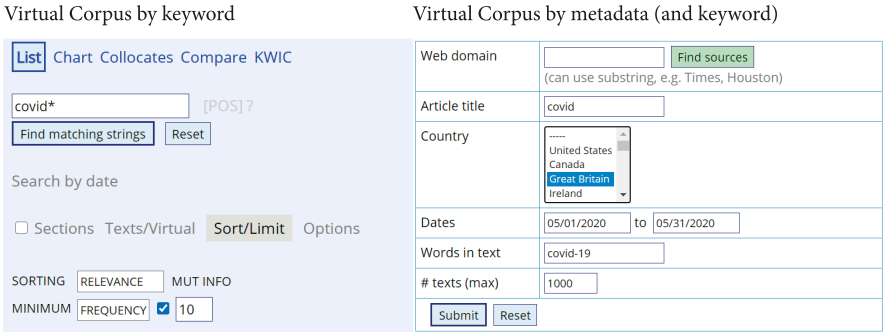
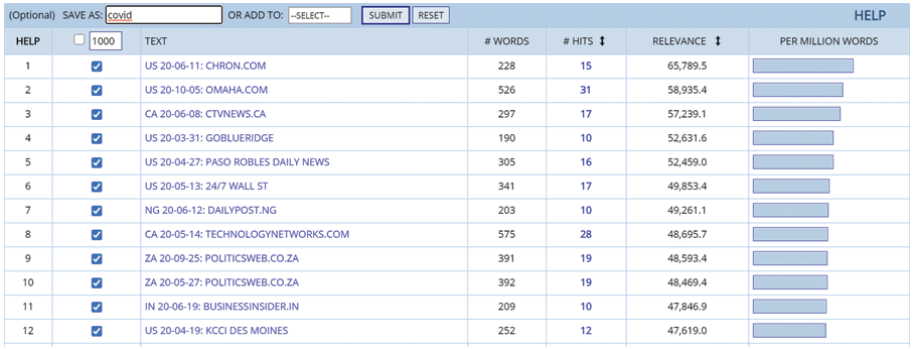


Figure 1. Creating Virtual Corpora in the NOW Corpus

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.



HELP	1000	TEXT	# WORDS	# HITS	RELEVANCE	PER MILLION WORDS
1	☑	US 20-06-11: CHRON.COM	228	15	65,789.5	
2	☑	US 20-10-05: OMAHA.COM	526	31	58,935.4	
3	☑	CA 20-06-08: CTVNEWS.CA	297	17	57,239.1	
4	☑	US 20-03-31: GOBLUERIDGE	190	10	52,631.6	
5	☑	US 20-04-27: PASO ROBLES DAILY NEWS	305	16	52,459.0	
6	☑	US 20-05-13: 24/7 WALL ST	341	17	49,853.4	
7	☑	NG 20-06-12: DAILYPOST.NG	203	10	49,261.1	
8	☑	CA 20-05-14: TECHNOLOGYNETWORKS.COM	575	28	48,695.7	
9	☑	ZA 20-09-25: POLITICSWEB.CO.ZA	391	19	48,593.4	
10	☑	ZA 20-05-27: POLITICSWEB.CO.ZA	392	19	48,469.4	
11	☑	IN 20-06-19: BUSINESSINSIDER.IN	209	10	47,846.9	
12	☑	US 20-04-19: KCCI DES MOINES	252	12	47,619.0	

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.

COVID [528,613 WORDS, 1000 TEXTS] (4.7% OF TOTAL) NOUN VERB ADJ ADV N+N ADJ+N [ALL CORPORA] SAVE LIST

HELP	WORD (CLICK FOR CONTEXT)	FREQ	# TEXTS	SPECIFIC FREQ 96 107 TEXTS	ENTIRE CORPUS	EXPECTED
1	CORONAVIRUS	1169	407	29.3	846	39.9
2	LOCKDOWN	242	112	1.3	4,010	189.3
3	QUARANTINE	177	108	0.7	5,614	265.0
4	VIRUS	1085	431	0.3	79,316	3,743.5
5	SYMPTOM	1038	283	0.2	93,271	4,402.2
6	INFECTION	960	361	0.2	89,635	4,230.6
7	OUTBREAK	470	186	0.2	49,361	2,329.7
8	SPREAD	365	223	0.2	44,127	2,082.7
9	ISOLATION	214	126	0.2	28,082	1,325.4
10	TESTING	551	239	0.2	75,188	3,548.7
11	PATIENT	2747	576	0.1	404,807	19,105.9
12	FATALITY	191	111	0.1	29,473	1,391.1

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*}.
ii. The text had at least one of the following words/strings in the title: *at-risk*, *cases*, *confirmed*, *contagious*, *containm**, *coronavirus*, *covid**, *curbside*, *curve*, *deaths*, *disinfect**, *distanc**, *epicenter*, *epidemic*, *epidemiol**, *flatten**, *flu*, *high-risk*, *hoard**, *hospital**, *hydroxychloroquine*, *infect**, *influenza*, *isolat**, *lock-down*, *lock-down*, *mask**, *nursing*, *outbreak*, *pandemic*, *panic*, *patient**, *pneumon**, *preventative*, *preventive*, *quarantin**, *re-open**, *reopen**, *respiratory*, *sanitiz**, *self-isolat**, *shelter**, *shutdown*, *spread*, *spreading*, *stay-at-home*, *stay at home*, *stockpil**, *testing*, *vaccine**, *ventilator**, *virus*.

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)

	NOW Corpus			Coronavirus Corpus			
Month	**# 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	365,007	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	2,824	373,629	255,599,268	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-*

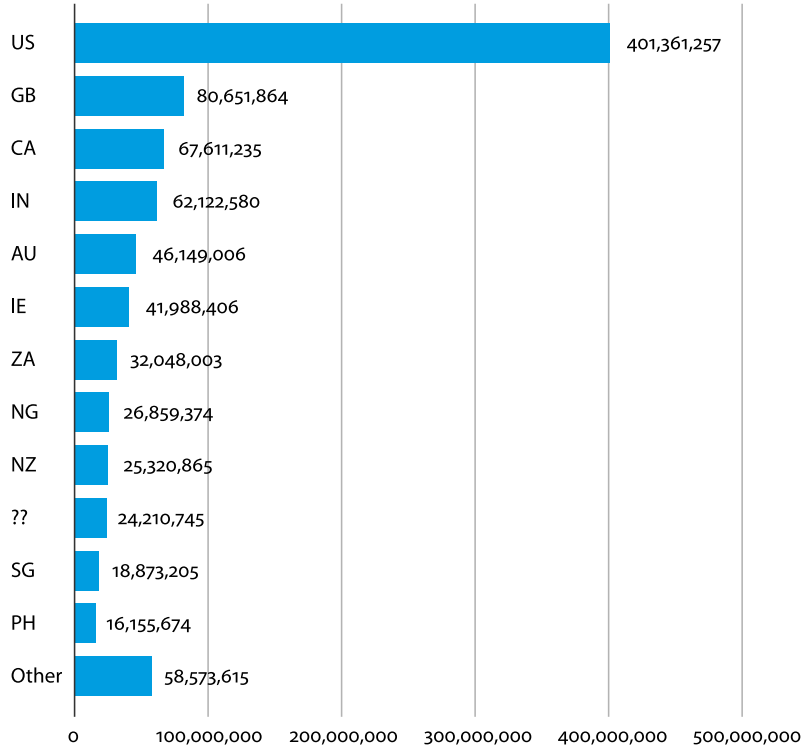


Figure 4. Composition of the Coronavirus Corpus, by country

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.

SECTION	ALL	20-01-01	20-02-01	20-02-11	20-02-21	20-03-01	20-03-11	20-03-21	20-04-01	20-04-11	20-04-21	20-05-01	20-05-11	20-05-21
FREQ	73907	14424	10516	4334	3074	4353	3325	6539	4548	4226	3097	2894	2441	1682
WORDS (M)	237	7.3	4.8	4.0	5.7	17.6	26.8	55.5	38.4	35.8	33.8	31.3	30.5	36.1
PER MIL	310.99	1,965.06	2,207.78	1,082.83	536.59	246.72	123.99	117.76	118.38	118.20	91.62	92.60	80.07	46.62
SEE ALL SUB-SECTIONS AT ONCE														

Figure 5. Frequency of *Wuhan* over time[1]

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

20-02-21	20-03-01	20-03-11	20-03-21	20-04-01	20-04-11	20-04-21	20-05-01	20-05-11	20-05-21	20-06-01	20-06-11	20-06-21	20-07-01	20-07-11
71	509	910	1924	798	654	482	184	79	36	60	107	90	169	64
5.7	17.6	26.8	55.5	38.4	35.8	33.8	31.3	30.5	36.1	29.3	27.1	26.8	27.0	22.5
12.39	28.85	33.93	34.65	20.77	18.29	14.26	5.89	2.59	1.00	2.05	3.95	3.35	6.20	2.84

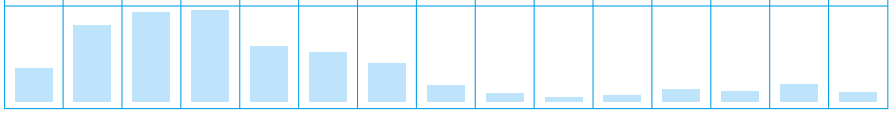


Figure 6. Frequency of *hoard** over time

20-03-11	20-03-21	20-04-01	20-04-11	20-04-21	20-05-01	20-05-11	20-05-21	20-06-01	20-06-11	20-06-21	20-07-01	20-07-11	20-07-21	20-08-01
264	563	601	613	921	724	658	818	897	644	523	470	333	427	324
26.8	55.5	38.4	35.8	33.8	31.3	30.5	36.1	29.3	27.1	26.8	27.0	22.5	28.7	25.6
9.84	10.14	15.64	17.14	27.25	23.17	21.58	22.67	30.59	23.75	19.49	17.25	14.79	14.90	12.65

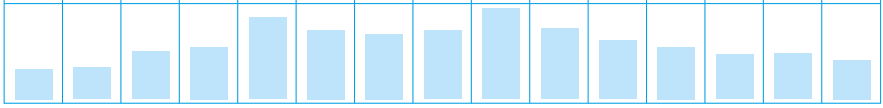


Figure 7. Frequency of *Sweden* over time

20-04-01	20-04-11	20-04-21	20-05-01	20-05-11	20-05-21	20-06-01	20-06-11	20-06-21	20-07-01	20-07-11	20-07-21	20-08-01	20-08-11	20-08-21
3266	11750	18827	22266	29448	29588	23180	21069	21988	20457	13339	13427	9323	9129	9051
38.4	35.8	33.8	31.3	30.5	36.1	29.3	27.1	26.8	27.0	22.5	28.7	25.6	24.5	24.1
85.01	328.63	556.98	712.48	965.98	820.16	790.44	776.94	819.45	750.91	592.36	468.62	364.05	373.14	375.76

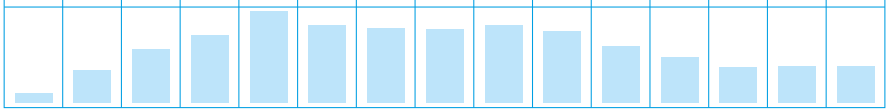


Figure 8. Frequency of *re*open** over time

20-02-01	20-03-01	20-03-11	20-03-21	20-04-01	20-04-11	20-04-21	20-05-01	20-05-11	20-05-21	20-06-01	20-06-11	20-06-21	20-07-01	20-07-11	20-07-21	20-08-01	20-08-11
0	35	533	2117	1566	1264	1029	773	647	531	350	333	352	317	284	252	217	206
5.7	17.6	26.8	55.5	38.4	35.8	33.8	31.3	30.5	36.1	29.3	27.1	26.8	27.0	22.5	28.7	25.6	24.5
0.00	1.98	19.88	38.12	40.76	35.35	30.44	24.73	21.22	14.72	11.94	12.28	13.12	11.64	12.61	8.80	8.47	8.42

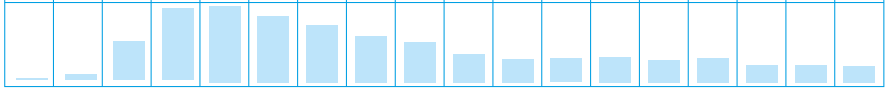


Figure 9. Frequency of *FLATTEN the curve* over time

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 distanc** 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.

20-03-01	20-03-11	20-03-21
910	5437	26948
17.6	26.8	55.5
51.58	202.74	485.29

20-03-11	20-03-12	20-03-13	20-03-14	20-03-15	20-03-16	20-03-17
253	299	780	561	637	1109	801
2.9	1.9	4.5	2.6	2.4	3.6	2.3
86.19	**156.24**	**174.56**	**215.40**	**264.26**	**311.82**	**355.09**

Figure 10. Frequency of *social distance* over time

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.

29	20-03-18 US	Asheville Citizen-Times	A	B	C	Facebook page . # The post said the spa hopes to	reopen	by	April	10	; # Karen Ch ? vez is an award-winning
30	20-03-27 US	cbsnews.com	A	B	C	Trump 's hope that the U . S . economy could	reopen	by	Easter	and	whether we should expect for New Yo
31	20-03-24 US	cbsnews.com	A	B	C	: Trump says he " would love " to have U.S.	reopen	by	Easter	even	as nation 's death toll tops 700 # Pre
32	20-03-26 US	Poughkeepsie Journal	A	B	C	be . # President Trump has said he wants America to	re-open	by	Easter	(	April 12) . Health experts say we 'll
33	20-03-30 NZ	rnz.co.nz	A	B	C	have prompted President Donald Trump to push for businesses to	reopen	by	Easter	,	# Given rising infections and a mounting
34	20-03-21 US	Pensacola News Journal	A	B	C	in groups at all . " # Seacrest is hoping to	re-open	by	the	end	of March , but recognizes that 's looking
88	20-03-23 ZA	citizen.co.za	A	B	C	open from Wednesday 18 March . The school said they would	reopen	on	14	April	on whether at the school sites or digital
89	20-03-14 US	BroadwayWorld	A	B	C	time they will reevaluate the closure to determine if they can	reopen	on	April	1st	; # The Celebration of the Arts gala ,
90	20-03-14 US	WLWT	A	B	C	our resorts starting at 2pm on March 15 with plans to	re-open	on	April	2	, " The company will be offering full refun
91	20-03-14 US	Patch	A	B	C	, or COVID-19 . # Schools are tentatively scheduled to	reopen	on	April	6	; but will be re-evaluated as the date appr
92	20-03-07 US	HotNewHipHop	A	B	C	end of January due to concerns surrounding coronavirus , before	reopening	on	February	10th	; The Tesla headquarters are locat
93	20-03-11 IN	huffingtonpost.in	A	B	C	. # As the hill temple Lord Ayyappa at Sabarimala will	reopen	on	March	13	, monthly pujas can be held , but at
94	20-03-20 US	The Spectrum	A	B	C	slow the spread of the coronavirus . it is set to	reopen	on	March	31	; (Photo : Elizabeth Weise) # The

Figure 11. KWIC lines for *re*open** in March 2020

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*).

91	20-05-02 GB	leicestermercury.co.uk	A	B	C	public , he said : " We are not going to	re-open	schools	in it is n't safe . # " Of course ,
92	20-05-04 GB	spectator.co.uk	A	B	C	. It started at 1 p.m . # Iceland began to	reopen	schools	today . South Korea will begin to reopen school
93	20-04-26 AU	thenewdaily.com.au	A	B	C	a study cited by the federal government in its push to	reopen	schools	. # The National Centre for Immunisation Rese
94	20-04-29 US	Detroit News	A	B	C	GOP lawmakers and Whitmer have focused on how quickly to begin	reopening	sectors	of Michigan 's economy . # The governor has be
95	20-04-22 US	The Hill	A	B	C	the continued pandemic , Colorado will join other states in	reopening	select	stores . # By # Alexandra Kelley # Michael
96	20-04-21 US	News 96.5 - WDBO	A	B	C	EDT April 21 : Gov . Brian Kemp 's call to	reopen	shuttered	businesses in Georgia left many business ow
97	20-05-07 GB	bristolpost.co.uk	A	B	C	and recycling centres as soon as practicable . The decision to	reopen	sites	will be taken by individual councils based on risk
98	20-04-14 US	al.com	A	B	C	combined are the state 's largest private employer , suggests	reopening	small	retail stores with the same occupancy restrictions
99	20-04-27 CA	cbc.ca	A	B	C	Colorado , Montana , Texas and Tennessee were also set to	reopen	some	businesses to start reviving their battered econor
100	20-05-06 US	Penn Live	A	B	C	counties moving into the yellow phase Friday will be allowed to	reopen	some	businesses . with some restrictions . # For counti
101	20-04-12 GB	eadt.co.uk	A	B	C	, while Austria and the Czech Republic are looking to gradually	re-open	some	shops to the public . # There are still concerns ab
102	20-04-29 IE	thejournal.ie	A	B	C	failure to ramp up testing has hindered Ireland 's hopes of	reopening	sooner	. # The Irish Times reports today that according
103	20-04-30 US	Abilene Reporter-News	A	B	C	think it was a joke . " Some question delay in	reopening	spas	. salons in Abilene # Some see Gov . Greg Abbott
104	20-04-16 --	Herald and News	A	B	C	said on Monday that his " authority is total " to	reopen	states	but has since backed away from that claim after

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

HELP		CONTEXT	FREQ	ALL	%	MI
1		GATHERINGS	4059	36911	11.00	8.87
2		TRAVEL	1095	118007	0.93	5.30
3		SALE	888	18315	4.85	7.68
4		ENTRY	823	17204	4.78	7.67
5		EVENTS	733	73667	1.00	5.40
6		FLIGHTS	709	35871	1.98	6.39
7		USE	665	191431	0.35	3.88
8		VISITORS	585	28477	2.05	6.45
9		GATHERING	538	19344	2.78	6.88

Figure 13. Noun collocates of *BAN_v*

HELP		CONTEXT	ALL	20-03-21	20-04-01	20-04-11	20-04-21	20-05-01	20-05-11	20-05-21	20-06-01	20-06-11	20-06-21	20-07-01	20-07-11	20-07-21
1		WEAR	68540	1527	2709	2511	2484	2994	3355	4371	2890	3636	5302	5145	5097	5252
2		WEARING	67722	2538	2610	2383	2417	2579	3050	4275	2880	3224	4338	4300	4142	4773
3		FACE	12382	749	763	679	695	597	674	660	596	562	712	636	693	697
4		REQUIRED	8077	54	139	193	271	397	492	479	367	513	637	664	624	682
5		WORE	6013	227	158	171	179	223	283	506	308	246	360	376	477	324
6		WEARS	5177	372	282	312	250	244	239	315	151	157	271	234	251	270
7		PROTECT	4419	350	378	276	313	222	195	266	177	251	176	193	199	252
8		WORN	4080	75	154	121	107	156	167	212	180	227	295	331	290	309
9		MAKING	3808	459	446	384	299	218	222	177	183	112	124	125	180	178
10		REQUIRING	2930	8	26	66	84	102	108	127	87	190	331	348	345	331
11		REQUIRE	2866	13	37	54	67	111	90	103	72	251	419	344	381	292
12		PREVENT	2517	83	205	132	115	105	84	113	79	123	152	155	117	189
13		COVERING	1854	18	52	125	122	114	128	100	78	67	100	123	149	176
14		DONATED	1405	212	219	133	152	94	88	76	34	47	35	34	33	35
15		WASH	1387	21	59	49	52	42	63	81	51	52	101	119	90	120
16		WALKS	1367	153	85	89	77	68	69	57	60	33	48	30	50	45
17		DISTRIBUTED	1316	128	123	155	133	102	82	55	40	44	69	39	32	53
18		REMOVE	1241	31	52	44	38	58	46	69	43	68	53	78	77	96
19		RECOMMENDED	1232	47	109	59	57	59	58	70	65	52	96	77	70	70
20		PRODUCE	1143	198	226	118	86	61	41	41	37	27	28	25	20	18
21		SELLING	1129	63	76	61	80	67	59	47	56	62	44	33	39	42
22		ENCOURAGED	1111	12	42	49	35	57	94	73	74	62	89	84	65	69

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.

HELP		CONTEXT	FREQ		ALL	%	MI
1	☐	ECONOMY	4965		93453	5.31	5.56
2	☐	BUSINESSES	2380		116541	2.04	4.18
3	☐	SCHOOLS	1789		62918	2.84	4.66
4	☐	ECONOMIES	1068		13377	7.98	6.15
5	☐	PLAN	841		63264	1.33	3.56
6	☐	STORES	674		35330	1.91	4.08
7	☐	RESTAURANTS	525		33930	1.55	3.78
8	☐	PARTS	504		25156	2.00	4.15
9	☐	PARKS	497		19134	2.60	4.53
10	☐	SHOPS	417		19159	2.18	4.27
11	☐	BEACHES	409		6436	6.35	5.82

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

HELP		CONTEXT	FREQ		ALL	%	MI
1	☐	SCHOOLS	5481		61197	8.96	6.04
2	☐	PLAN	1447		43393	3.33	4.61
3	☐	ECONOMY	1207		50890	2.37	4.12
4	☐	PLANS	802		32498	2.47	4.18
5	☐	BUSINESSES	741		50220	1.48	3.44
6	☐	FALL	608		33492	1.82	3.74
7	☐	DOORS	434		5900	7.36	5.75
8	☐	BORDERS	405		5112	7.92	5.86
9	☐	CAPACITY	359		26337	1.36	3.32
10	☐	PUBS	299		4817	6.21	5.51
11	☐	BARS	296		9051	3.27	4.58
12	☐	ECONOMIES	294		6090	4.83	5.15
13	☐	INSTRUCTION	280		6484	4.32	4.99

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).

SEC 1 (20-03-11, 20-03-21, 20-04-0...): 221,572,548 WORDS

	WORD/PHRASE	TOKENS 1	TOKENS 2	PM 1	PM 2	RATIO
1	ECONOMY	4965	1207	22.4	9.2	2.4
2	BUSINESSES	2380	741	10.7	5.6	1.9
3	STATE	1948	454	8.8	3.4	2.6
4	SCHOOLS	1789	5481	8.1	41.6	0.2
5	COUNTRY	1279	387	5.8	2.9	2.0
6	ECONOMIES	1068	294	4.8	2.2	2.2
7	PLAN	841	1447	3.8	11.0	0.3
8	WEEK	829	540	3.7	4.1	0.9
9	BUSINESS	825	346	3.7	2.6	1.4
10	STORES	674	215	3.0	1.6	1.9
11	RESTAURANTS	525	269	2.4	2.0	1.2

SEC 2 (20-08-01, 20-08-11, 20-08-2...): 131,772,791 WORDS

	WORD/PHRASE	TOKENS 2	TOKENS 1	PM 2	PM 1	RATIO
1	SCHOOLS	5481	1789	41.6	8.1	5.2
2	PLAN	1447	841	11.0	3.8	2.9
3	ECONOMY	1207	4965	9.2	22.4	0.4
4	PLANS	802	248	6.1	1.1	5.4
5	BUSINESSES	741	2380	5.6	10.7	0.5
6	FALL	608	194	4.6	0.9	5.3
7	WEEK	540	829	4.1	3.7	1.1
8	STATE	454	1948	3.4	8.8	0.4
9	DOORS	434	404	3.3	1.8	1.8
10	BORDERS	405	172	3.1	0.8	4.0
11	COUNTRY	387	1279	2.9	5.8	0.5

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*).

SEC 1 (20-02-01, 20-02-11, 20-02-2...): 32,137,941 WORDS

	WORD/PHRASE	TOKENS 1	TOKENS 2	PM 1	PM 2	RATIO
1	POSTPONED	34	1	1.1	0.0	171.1
2	SURROUNDING	30	1	0.9	0.0	150.9
3	CORONAVIRUS-HIT	29	1	0.9	0.0	145.9
4	SELF-ISOLATE	26	1	0.8	0.0	130.8
5	SHUTTERED	20	1	0.6	0.0	100.6
6	AFFECTED	77	4	2.4	0.0	96.8
7	44-YEAR-OLD	27	0	0.8	0.0	84.0
8	14-DAY	26	0	0.8	0.0	80.9
9	FOREMOST	16	1	0.5	0.0	80.5
10	FLU-LIKE	24	0	0.7	0.0	74.7
11	RADICAL	42	3	1.3	0.0	70.4
12	MARINE	14	1	0.4	0.0	70.4
13	EPICENTER	226	17	7.0	0.1	66.9
14	ISSUING	13	1	0.4	0.0	65.4

SEC 2 (20-07-11, 20-07-21, 20-06-0...): 161,689,277 WORDS

	WORD/PHRASE	TOKENS 2	TOKENS 1	PM 2	PM 1	RATIO
1	PRO	80	0	0.5	0.0	49.5
2	GALWAN	31	0	0.2	0.0	19.2
3	TERRITORIAL	76	1	0.5	0.0	15.1
4	DISPUTED	24	0	0.1	0.0	14.8
5	INTACT	22	0	0.1	0.0	13.6
6	NUCLEAR	21	0	0.1	0.0	13.0
7	INFLUENTIAL	20	0	0.1	0.0	12.4
8	UNLAWFUL	19	0	0.1	0.0	11.8
9	ASYMPTOMATIC	59	1	0.4	0.0	11.7
10	ACCOUNTABLE	55	1	0.3	0.0	10.9
11	DETAINED	15	0	0.1	0.0	9.3
12	UNFAIR	14	0	0.1	0.0	8.7
13	UNJUSTIFIED	14	0	0.1	0.0	8.7
14	EXPANSIVE	11	0	0.1	0.0	6.8

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).

SEC 1 (20-03-11, 20-03-21, 20-04-0...): 221,572,548 WORDS

	WORD/PHRASE	TOKENS 1	TOKENS 2	PM 1	PM 2	RATIO
1	PALLETS	32	1	0.1	0.0	22.0
2	PRINTERS	60	2	0.3	0.0	20.7
3	NUNS	30	1	0.1	0.0	20.7
4	COFFIN	28	1	0.1	0.0	19.3
5	ACCESSING	40	0	0.2	0.0	18.1
6	GRAPHIC	26	1	0.1	0.0	17.9
7	PLANTS	24	1	0.1	0.0	16.5

SEC 2 (20-07-11, 20-07-21, 20-08-0...): 152,575,302 WORDS

	WORD/PHRASE	TOKENS 2	TOKENS 1	PM 2	PM 1	RATIO
1	FREEDOM	68	1	0.4	0.0	98.8
2	REPUBLICAN	66	1	0.4	0.0	95.8
3	ORDINANCES	86	2	0.6	0.0	62.4
4	MANDATES	1155	27	7.6	0.1	62.1
5	DRINK	42	1	0.3	0.0	61.0
6	BYLAW	93	0	0.6	0.0	61.0
7	MANDATE	3336	83	21.9	0.4	58.4

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.

SEC 1 (20-01-01, 20-02-01, 20-02-1...): 39,478,174 WORDS

	WORD/PHRASE	TOKENS 1	TOKENS 2	PM 1	PM 2	RATIO
1	-FULL	352	0	8.9	0.0	891.6
2	BLOCK-TIME	350	0	8.9	0.0	886.6
3	DELUSIONARY	156	2	4.0	0.0	445.0
4	SCANDALSEVEN	156	0	4.0	0.0	395.2
5	PROVINCIAL-LEVEL	141	9	3.6	0.0	89.4
6	NCOV	359	27	9.1	0.1	75.9
7	BEAKED	143	11	3.6	0.0	74.2
8	HOLI	291	29	7.4	0.1	57.3
9	PAN-DEMOCRATIC	122	17	3.1	0.1	40.9
10	TAAL	218	36	5.5	0.2	34.5
11	PRO-ESTABLISHMENT	144	25	3.6	0.1	32.9
12	SHINCHEONJI	525	100	13.3	0.4	30.0
13	JINYINTAN	149	32	3.8	0.1	26.6
14	VIRUS-STRICKEN	485	106	12.3	0.5	26.1
15	LUNAR	3273	748	82.9	3.3	25.0
16	COLONIAL-ERA	100	23	2.5	0.1	24.8
17	SELF-FULFILLING	169	47	4.3	0.2	20.5
18	HAINAN	141	42	3.6	0.2	19.2
19	2019-NCOV	1250	382	31.7	1.7	18.7
20	FLU-RELATED	206	64	5.2	0.3	18.4

SEC 2 (20-03-21, 20-04-01, 20-04-1...): 225,240,320 WORDS

	WORD/PHRASE	TOKENS 2	TOKENS 1	PM 2	PM 1	RATIO
1	STAY-AT-HOME	19480	62	86.5	1.6	55.1
2	ANTI-MALARIAL	636	4	2.8	0.1	27.9
3	21-DAY	3484	23	15.5	0.6	26.5
4	DISTANCING	3892	28	17.3	0.7	24.4
5	FURLOUGHED	826	6	3.7	0.2	24.1
6	CONTACT-TRACING	835	7	3.7	0.2	20.9
7	COVID	13870	137	61.6	3.5	17.7
8	SHELTER-IN-PLACE	3291	33	14.6	0.8	17.5
9	TO-GO	1013	11	4.5	0.3	16.1
10	INTER-STATE	1079	12	4.8	0.3	15.8
11	NONVIOLENT	535	6	2.4	0.2	15.6
12	UNORGANISED	428	5	1.9	0.1	15.0
13	STAY-HOME	1298	16	5.8	0.4	14.2
14	SOCIAL-DISTANCING	3982	52	17.7	1.3	13.4
15	LAID-OFF	747	10	3.3	0.3	13.1
16	PRE-PANDEMIC	774	13	3.4	0.3	10.4
17	NON-COVID	277	5	1.2	0.1	9.7
18	NEAR-TOTAL	220	4	1.0	0.1	9.6
19	DRIVE-THRU	5163	96	22.9	2.4	9.4
20	LONG-TERM-CARE	429	8	1.9	0.2	9.4

Figure 20. Comparison of all adjectives in two different periods

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).

SWEDEN [218,854 WORDS, 100 TEXTS] (0.0% OF TOTAL) NOUN VERB ADJ ADV N+N ADJ+N [ALL CORPORA] SAVE LIST

HELP	WORD (CLICK FOR CONTEXT)	FREQ	# TEXTS	SPECIFIC FREQ 22 4 TEXTS	ENTIRE CORPUS	EXPECTED
1	CAPITA	43	18	141.0	900	0.3
2	PER	43	18	138.9	914	0.3
3	RELAXED	40	23	70.8	1,667	0.6
4	BANNED	24	9	36.7	1,931	0.7
5	ELDERLY	242	53	36.3	19,653	6.7
6	STRICT	164	43	30.7	15,762	5.3
7	MISLEADING	23	7	29.4	2,310	0.8
8	COMPARABLE	23	12	27.5	2,469	0.8
9	SOFT	31	17	27.3	3,351	1.1
10	FAR	38	19	26.9	4,176	1.4
11	SUCCESSFUL	92	18	24.3	11,193	3.8
12	STRINGENT	38	14	24.2	4,630	1.6
13	CONTROVERSIAL	24	20	21.4	3,306	1.1
14	QUIET	36	10	20.2	5,270	1.8
15	VOLUNTARY	41	25	19.4	6,251	2.1
16	MANDATORY	72	10	19.0	11,194	3.8

Figure 21. Keywords in "Sweden" Virtual Corpus

CLICK FOR MORE CONTEXT [?] SAVE LIST CHOOSE LIST —— CREATE NEW LIST [?] SHOW DUPLICATES

1	20-09-11 US	Reason	A	B	C	Per Capita The trends suggest that Sweden's less restrictive policy has been more successful at reducing fatal outcomes. # I argued last week that it wa
2	20-06-22 US	New York Times	A	B	C	. Its economy, which is far more globally integrated than the others, includes successful brands like Volvo, Ikea and H&M, as well as the band Abba.
3	20-06-27 GB	theguardian.com	A	B	C	world, to gain clinical knowledge. The anti-inflammatory drug, Dexamethasone, has proved successful for critical patients, leading to a reduction in fata
4	20-05-20 US	Business Insider	A	B	C	Organization, told Business Insider that while Sweden's relaxed policy had not yet been successful in curbing its COVID-19 outbreak, it would in the futu
5	20-04-19 IE	thejournal.ie	A	B	C	million people and where 13,271 cases are recorded -- it might suggest Sweden has been successful in slowing the pace of the virus spreading more tha
6	20-05-05 NZ	newstalkzb.co.nz	A	B	C	several weeks had been stable, " so in that aspect it has to be successful. " # But he added: " What it failed at, I think
7	20-06-27 GB	theguardian.com	A	B	C	world, to gain clinical knowledge. The anti-inflammatory drug, Dexamethasone, has proved successful for critical patients, leading to a reduction in fata
8	20-06-27 GB	theguardian.com	A	B	C	". # I think the country that will be proven to be the most successful in its dealing with Covid will be by economic measure. Which country will get
9	20-06-27 GB	theguardian.com	A	B	C	Either the Swedish Government and the Folkh? Isomyndigheten have a really poor understanding of successful lockdown policies or the Swedish Gover
10	20-06-27 GB	theguardian.com	A	B	C	covid level the quickest. That country will be seen to have been the most successful # As someone who has always tended to a positive view of Sweden
11	20-08-17 NZ	newsroom.co.nz	A	B	C	dichotomy. # " All of this suggests that if New Zealand goes into another successful lockdown, it will suffer only incremental additional economic damag
12	20-06-27 GB	theguardian.com	A	B	C	responses based on the same data? " It depends what you mean by " successful ". # I think the country that will be proven to be the most
13	20-07-09 US	Business Insider on MSN.com	A	B	C	their recent spell of good news does not mean the country's no-lockdown plan was successful, or that other countries should follow its path. # Jon Talli
14	20-09-11 US	Reason	A	B	C	surpassed Sweden's, and the trajectory of deaths suggests that Sweden has been more successful at reducing mortality, despite (or perhaps because o
15	20-08-17 NZ	newsroom.co.nz	A	B	C	of the virus, compared to countries that do not. Countries that have implemented successful lockdowns are generally doing much better economically

Figure 22. KWIC lines for the keyword *successful* in a "Sweden" Virtual Corpus

5. Conclusion

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.

References

Davies, M. (2015). Corpora: An introduction. In D. Biber & R. Reppen (Eds.), *Cambridge Handbook of English Corpus Linguistics* (pp. 11–31). Cambridge University Press. https://doi.org/10.1017/CBO9781139764377.002

Davies, M. (2017). Using large online corpora to examine lexical, semantic, and cultural variation in different dialects and time periods. In E. Friginal (Ed.), *Studies in Corpus-Based Sociolinguistics* (pp. 19–82). Routledge. https://doi.org/10.4324/9781315527819-2

Davies, M. (2018). Corpus-based studies of lexical and semantic variation: The importance of both corpus size and corpus design. In C. Suhr, T. Nevalainen, & I. Taavitsainen (Eds.), *From Data to Evidence in English Language Research* (pp. 34–55). Brill.

Davies, M., & Fuchs, R. (2015). Expanding horizons in the study of World Englishes with the 1.9 billion word Global Web-Based English Corpus (GloWbE). *English World-Wide*, *36*(1), 1–28. https://doi.org/10.1075/eww.36.1.01dav

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