The best of both worlds: Multi-billion word “dynamic” corpora

Mark Davies
Department of Linguistics
Brigham Young University
mark_davies@byu.edu

Abstract

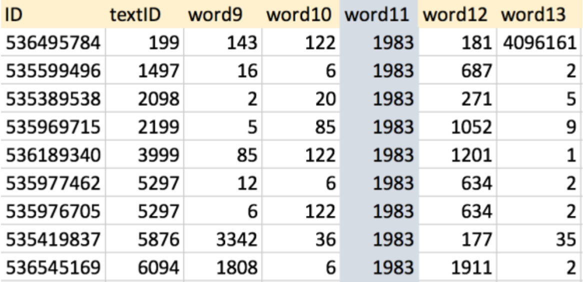
Nearly all of the very large corpora of English are “static”, which allows a wide range of one-time, pre-processed data, such as collocates. The challenge comes with large “dynamic” corpora, which are updated regularly, and where pre-processing is much more difficult. This paper provides an overview of the NOW corpus (News on the Web), which is currently 8.2 billion words in size, and which grows by about 170 million words each month. We discuss the architecture of NOW, and provide many examples that show how data from NOW can (uniquely) be extracted to look at a wide range of ongoing changes in English.

1 Corpus architecture

Multi-billion word corpora have become commonplace in the last 5-10 years. For example, there are several different 10-20 billion word corpora from Sketch Engine (Kilgarrif et al 2014; www.sketchengine.eu), Corpora from the Web (Schäfer 2015; corporafromtheweb.org), and English-Corpora.org (formerly the BYU Corpora).

Most of these corpora, however, are “static” corpora. The corpus texts are collected and annotated, and they are then indexed and pre-processed in other ways, which makes text retrieval very fast even on very large corpora. For example, the 14 billion word iWeb corpus (https://www.english-corpora.org/iweb), users can search by word form, lemma, part of speech, synonyms, user-defined wordlists, and more. A search for a complex string like *VERB _a =EXPENSIVE @CLOTHES* (verb + article + any form of any synonym of *expensive* + any form of any word in the user-defined *clothes* wordlist) will take just 2-3 seconds.

iWeb and all of the corpora from English-Corpora.org are based on highly-optimized relational databases, which yields corpora that are typically 5-10 times as fast as other large corpora (see www.english-corpora.org/speed.asp). The underlying architecture is similar to “columnstore” databases. In a 14 billion word corpus, for example, there would be 14 billion rows, each with a structure like the following:



ID	textID	word9	word10	word11	word12	word13
536495784	199	143	122	1983	181	4096161
535599496	1497	16	6	1983	687	2
535389538	2098	2	20	1983	271	5
535969715	2199	5	85	1983	1052	9
536189340	3999	85	122	1983	1201	1
535977462	5297	12	6	1983	634	2
535976705	5297	6	122	1983	634	2
535419837	5876	3342	36	1983	177	35
536545169	6094	1808	6	1983	1911	2

Figure 1: Corpus architecture

Each word / lemma / PoS combination is represented as an integer value, which is tied to an entry in the lexicon (and which is in a separate database). In Figure 1, for example, the integer value [1983] represents [best / best / jjt]. There is a clustered index on this “middle” column ([word11] in Figure 1), which means that all of the tokens of any word (*best* in this case) are stored *physically* adjacent to each other on the SSD, which increases access speed a great deal.

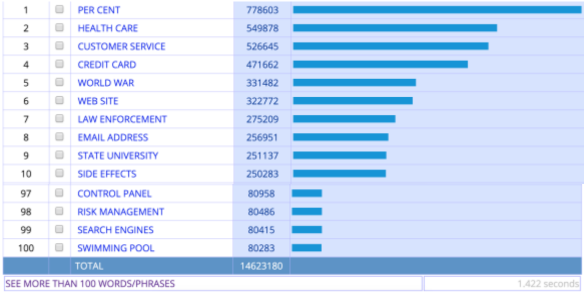
As it carries out the search, iWeb (or any of the corpora from English-Corpora.org) parses the search string to find the lowest-frequency, “weakest” part of the string. For example, in the search string *the best NOUN*, the word *best* occurs less than either *the* or all NOUNs. The search focuses first on the lemma *best*, and only when it finds those rows (all of the rows containing the value 1983 in column [word11]) does it narrow this to rows where the preceding column ([word10] in Figure 1) is the value for *the* and the following column ([word12] in Figure 1) is an integer value tied to a noun in the lexicon. (Note that in Figure 1 (for reasons of space), only the two columns to the left and to the right of the “node” column are shown, but – depending on the corpus – there are 5-10 columns each to the left and to the right).

Davies (2019) explains the underlying architecture in more detail, and provides a number of examples that show that the corpora with this architecture are typically 5-10 times as fast as the architecture of other very large corpora. Crucially, this is because these other corpora typically parse the search string left to right (e.g. with the word *the* first in the string *the best NOUN*), whereas we focus first on the "weakest link" in the search string.

Our approach also takes full advantage of relational database architecture, such as JOINs across any number of highly-optimized tables. For example, in the example of *VERB _a =EXPENSIVE @CLOTHES* shown above (verb + article + any form of any synonym of *expensive* + any form of any word in the user-defined *clothes* wordlist), the search will use lemma and part of speech information from the main [lexicon] table, as well as a separate [synonyms] table containing entries for more than 65,000 words, and another table containing user-defined lists such as clothing, emotions, or a particular class of verbs. Additional tables could contain pronunciation information or additional semantic information, and the search speed will not decrease much (if at all) no matter how many tables are involved.

Finally, there is a [sources] table that can contain any number of columns related to each of the texts in the corpus, and these are JOINed to the main corpus table (e.g. Figure 1) via the [textID] value. This allows users to quickly and easily create "virtual corpora" using any of the metadata from the [sources] table, such as author, date, website, or genre.

When the corpus sees that all of the "slots" in a search are very frequent, it defaults to using pre-processed n-grams, which are even faster than the previous approach. For example, a very high frequency search like "NOUN NOUN" takes less than two seconds, because it is only searching 10 or 100 million rows of data in the n-grams databases. (The downside of the n-gram tables is that they refer to the entire corpus, and not just particular sections, just as certain genres or texts.)



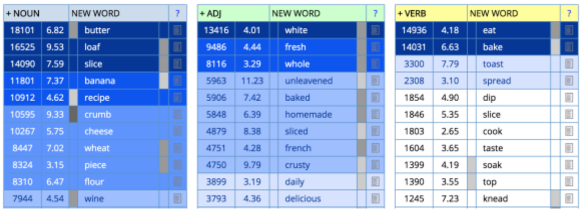
1	□	PER CENT	778603
2	□	HEALTH CARE	549878
3	□	CUSTOMER SERVICE	526645
4	□	CREDIT CARD	471662
5	□	WORLD WAR	331482
6	□	WEB SITE	322772
7	□	LAW ENFORCEMENT	275209
8	□	EMAIL ADDRESS	256951
9	□	STATE UNIVERSITY	251137
10	□	SIDE EFFECTS	250283
97	□	CONTROL PANEL	80958
98	□	RISK MANAGEMENT	80486
99	□	SEARCH ENGINES	80415
100	□	SWIMMING POOL	80283
		TOTAL	14623180

SEE MORE THAN 100 WORDS/PHRASES

1.422 seconds

Figure 2: iWeb high frequency: *NOUN + NOUN*

Finally, as with the Sketch Engine corpora, other data such as collocates are pre-processed in iWeb, which means they can be retrieved in just a second or two.



+ NOUN		NEW WORD	?
18101	6.82	butter	
16525	9.53	loaf	
14090	7.59	slice	
11801	7.37	banana	
10912	4.62	recipe	
10595	9.33	crumb	
10267	5.75	cheese	
8447	7.02	wheat	
8324	3.15	piece	
8310	6.47	flour	
7944	4.54	wine	

+ ADJ		NEW WORD	?
13416	4.01	white	
9486	4.44	fresh	
8116	3.29	whole	
5963	11.23	unleavened	
5906	7.42	baked	
5848	6.39	homemade	
4879	8.38	sliced	
4751	4.28	french	
4750	9.79	crusty	
3899	3.19	daily	
3793	4.36	delicious	

+ VERB		NEW WORD	?
14936	4.18	eat	
14031	6.63	bake	
3300	7.79	toast	
2308	3.10	spread	
1854	4.90	dip	
1846	5.35	slice	
1803	2.65	cook	
1604	3.65	taste	
1399	4.19	soak	
1390	3.55	top	
1245	7.23	knead	

Figure 3: iWeb collocates for *bread*

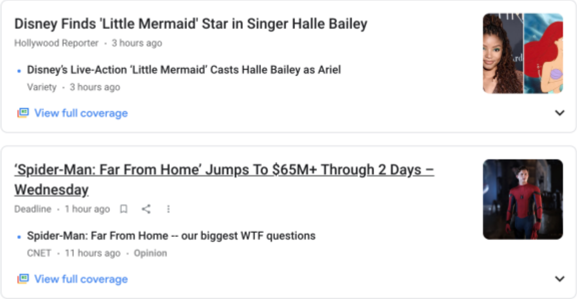
Pre-processing also allows for very fast retrieval (1-2 seconds for results from the 14 billion word corpus) for word clusters, related topics (words that frequently co-occur anywhere on the 22 million web pages), websites that use the word the most (which can be used to quickly and easily create "Virtual Corpora" on almost any topic), and sample concordance lines (see Davies 2019).

2 Creating the dynamic NOW corpus

As we will discuss in Section 4. the challenge comes, however, when we create a corpus that is "dynamic. (We define "dynamic" as corpora in which texts are continually added, rather than corpora in which texts are both added and deleted – although our architecture would have the same advantages in this case as well.)

An example of a dynamic corpus is the NOW Corpus ("News on the Web"; www.english-corpora.org/now), which is – as far as we are aware – the only corpus larger than a billion words, and which is growing on a regular basis (at least every month). The NOW corpus debuted at 3.6 billion words in May 2016 (with texts going back to 2010) and is now (early July 2019) about 8.2 billion words in size. Every month 150-170 million words are added to the corpus, or about 1.5 billion words each year. Note that similar corpora for Spanish and Portuguese are also available (corpusdelespanol.org/now: 6.0 billion words in 21 Spanish-speaking countries since 2012, and corpusdoportugues.org/now: 1.3 billion words in 4 Portuguese-speaking countries since 2012), but the English NOW corpus will be the focus of this paper.

To create the NOW corpus, every hour five different machines search Google News to retrieve newly-listed newspaper and magazine articles, for 20 different English-speaking countries (the same 20 countries as GloWbE; see Davies 2013). For example, Figure 4 shows just two sample entries from Google News from 3 July 2019, and on average we gather the URLs for about 20,000 such articles each day.



Disney Finds 'Little Mermaid' Star in Singer Halle Bailey
Hollywood Reporter · 3 hours ago

- Disney's Live-Action 'Little Mermaid' Casts Halle Bailey as Ariel
  Variety · 3 hours ago

View full coverage

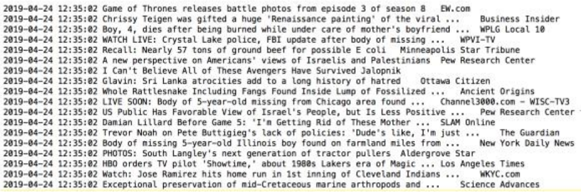
'Spider-Man: Far From Home' Jumps To $65M+ Through 2 Days – Wednesday
Deadline · 1 hour ago

- Spider-Man: Far From Home -- our biggest WTF questions
  CNET · 11 hours ago · Opinion

View full coverage

Figure 4: Sample Google News entries

The metadata for each of the 20,000 articles (URL, title, source, Google snippet) that appear each day are stored in a relational database. For example, the following is a small selection of the links from Google News from the US and Canada for the last hour on April 24, 2019, as the initial version of this paper was being written:

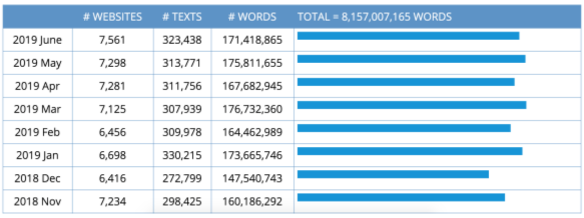


```
2019-04-24 12:35:02 Game of Thrones releases battle photos from episode 3 of season 8   EW.com
2019-04-24 12:35:02 Chrissy Teigen was gifted a huge 'Renaissance painting' of the viral ...    Business Insider
2019-04-24 12:35:02 Boy, 4, dies after being burned while under care of mother's boyfriend ...  WPLG Local 10
2019-04-24 12:35:02 WATCH LIVE: Crystal Lake police, FBI update after body of missing ...   WPVI-TV
2019-04-24 12:35:02 Recall: Nearly 57 tons of ground beef for possible E coli   Minneapolis Star Tribune
2019-04-24 12:35:02 A new perspective on Americans' views of Israelis and Palestinians  Pew Research Center
2019-04-24 12:35:02 I Can't Believe All of These Avengers Have Survived Jalopnik
2019-04-24 12:35:02 Glavin: Sri Lanka atrocities add to a long history of hatred    Ottawa Citizen
2019-04-24 12:35:02 Whole Rattlesnake Including Fangs Found Inside Lump of Fossilized ...   Ancient Origins
2019-04-24 12:35:02 LIVE SOON: Body of 5-year-old missing from Chicago area found ...   Channel3000.com - WISC-TV3
2019-04-24 12:35:02 US Public Has Favorable View of Israel's People, but Is Less Positive ...   Pew Research Center
2019-04-24 12:35:02 Damian Lillard Before Game 5: 'I'm Getting Rid of These Mother ...  SLAM Online
2019-04-24 12:35:02 Trevor Noah on Pete Buttigieg's lack of policies: 'Dude's like, I'm just ...    The Guardian
2019-04-24 12:35:02 Body of missing 5-year-old Illinois boy found on farmland miles from ...    New York Daily News
2019-04-24 12:35:02 PHOTOS: South Langley's next generation of tractor pullers  Aldergrove Star
2019-04-24 12:35:02 HBO orders TV pilot 'Showtime,' about 1980s Lakers era of Magic ... Los Angeles Times
2019-04-24 12:35:02 Watch: Jose Ramirez hits home run in 1st inning of Cleveland Indians ...    WKYC.com
2019-04-24 12:35:02 Exceptional preservation of mid-Cretaceous marine arthropods and ...   Science Advances
```

Figure 5: NOW sample list of articles

At the end of the month, we download the 250,000-300,000 articles using a custom program written in the Go language, which downloads all of the 250,000_ texts in about 30-40 minutes. We then use JusText (Pomikálak 2011; corpus.tools/wiki/Justext) to remove boilerplate material, and we tag the text with CLAWS 7 (for English; see Garside and Smith 1997), and a customized tagger based on Eckhard Bick's Palavras tagger for the Portuguese and Spanish corpora (Bick 1999). We then remove duplicate articles (always a problem in newspaper-based corpora) by looking for duplicate 11-grams across texts. For example, if a text has 68 11-grams starting with the word *the*, and 39 of these 11-grams are also found in any of the other 250,000+ texts from that month, then the text is tagged as a probable duplicate and it is removed from the corpus. (This process takes only 2-3 minutes for the 150-170 million words, because of the relational database architecture underlying the corpus).

Once we have done all of these steps, the new texts are then added to the existing corpus. As the Figure 6 shows (for Nov 2018 – June 2019), this results in about 150-175 million additional words of data each month:



	# WEBSITES	# TEXTS	# WORDS	TOTAL = 8,157,007,165 WORDS
2019 June	7,561	323,438	171,418,865	
2019 May	7,298	313,771	175,811,655	
2019 Apr	7,281	311,756	167,682,945	
2019 Mar	7,125	307,939	176,732,360	
2019 Feb	6,456	309,978	164,462,989	
2019 Jan	6,698	330,215	173,665,746	
2018 Dec	6,416	272,799	147,540,743	
2018 Nov	7,234	298,425	160,186,292	

Figure 6: NOW size by month (last 8 months)

Note that NOW contains just those articles that Google News links to, which are primarily newspaper and magazine sites. But there is an incredible variety in these sites – they are not just "staid" broadsheet newspapers. They include magazine and newspaper articles dealing not only with current events, but also technology, entertainment, and a wide variety of topics (as is evidenced by the 7,000+ "news" sites in a given month, as shown in Figure 6).

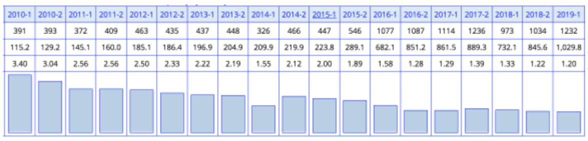
Evidence for the often informal nature of the texts comes from an investigation of the lexical creativity in the corpus. For example, there are more than 540 different *–alypse* words that are formed by analogy to the word *apocalypse*, such as *snarkpocalypse*, *snowpocalypse*, *chocopalypse*, *crapocalypse*, *kittiepocalypse*, *redditpocalypse*, *zombiepocalypse*, *and biebopalypse*. Likewise, there are more than 4,400 –fest words, including such innovative words as *gloomfest*, *testosterone-fest*, *brixfest*, *weep-fest*, *rant-fest*, *glumfest*, *oktemberfest*, *foul-fest*, and *raunchfest* (all of which occur at least five times in the corpus).

3 Examples from the NOW corpus

The advantage of a dynamic "monitor" corpus like NOW is that we are able to see what is going on with the language at the current time – not just 2 or 5 or 10 years ago.

At the most basic level, users can search for the frequency of a given word or phrase since 2010. For example, the following are just a few of the new words and phrases since 2010: *Brexit, trigger warning, catfishing, nomophobia, FOMO, birther, selfie stick, data lake, digital native, ransomware*. Some other cases of increase since 2010 include: (NOUN) *refugee*, *ransomware* (ADJ) *transgender**, *self-driving*, *on-demand*, *streaming*, *far-right* (VERB) *overreach*, *eventuate*, *intensify*, *text*, *retweet* (ADV) *effectively*, *programmatically*. Words showing a decrease in use during this time include: (NOUN) *waitress*, *disc*, *fax* (ADJ) *neat*, *old-fashioned*, *eco-friendly*, *eco-conscious*, *loopy*, *preppy*, *sullen*, *scanty* (VERB) *cream*, *clunk*, *flunk*, *gripe*, *murmur*, *foreclose* (ADV) *honorably*, *contentedly*, *frightfully*.

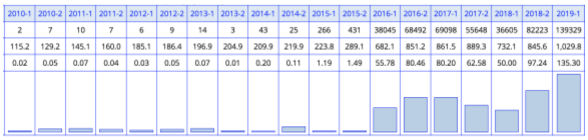
For any of these words or phrases, the NOW corpus shows the frequency in six month blocks (and with even more granularity, as we will soon see). For example, Figure 7 shows the decreasing frequency of *waitress* (which is viewed by some as being sexist, because of the feminine *–ess* ending) almost year by year since 2010:



2010-1	2010-2	2011-1	2011-2	2012-1	2012-2	2013-1	2013-2	2014-1	2014-2	2015-1	2015-2	2016-1	2016-2	2017-1	2017-2	2018-1	2018-2	2019-1
391	393	372	409	463	435	437	448	326	466	447	546	1077	1087	1114	1236	973	1034	1232
115.2	129.2	145.1	160.0	185.1	186.4	196.9	204.9	209.9	219.9	223.8	289.1	682.1	851.2	861.5	889.3	732.1	845.6	1,029.8
3.40	3.04	2.56	2.56	2.50	2.33	2.22	2.19	1.55	2.12	2.00	1.89	1.58	1.28	1.29	1.39	1.33	1.22	1.20

Figure 7: Frequency of *waitress*: every 6 months

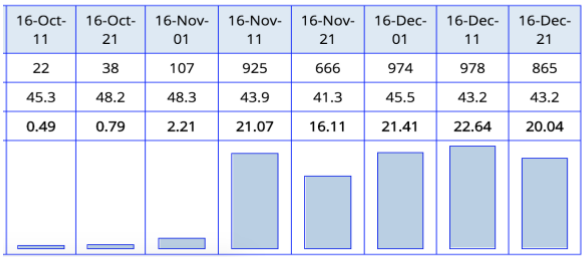
The 497,000+ tokens of *Brexit* show that it increased suddenly in the first half of 2016, and that (after a bit of a pause in late 2017 and early 2018) it has increased again in early 2019, to its highest level yet:



2010-1	2010-2	2011-1	2011-2	2012-1	2012-2	2013-1	2013-2	2014-1	2014-2	2015-1	2015-2	2016-1	2016-2	2017-1	2017-2	2018-1	2018-2	2019-1
2	7	10	7	6	9	14	3	43	25	266	431	38045	68492	69098	55648	36605	82223	139329
115.2	129.2	145.1	160.0	185.1	186.4	196.9	204.9	209.9	219.9	223.8	289.1	682.1	851.2	861.5	889.3	732.1	845.6	1,029.8
0.02	0.05	0.07	0.04	0.03	0.05	0.07	0.01	0.20	0.11	1.19	1.49	55.78	80.46	80.20	62.58	50.00	97.24	135.30

Figure 8: Frequency of *Brexit*: every 6 months

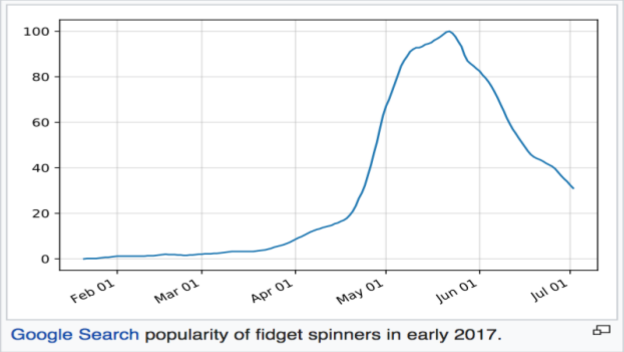
It is also possible to see the frequency of a word or phrase in 10-day increments. For example, the NOW corpus shows that the phrase *fake news* comes out of nowhere within a day or two of the 2016 US presidential elections (Nov 8, 2016):



16-Oct-11	16-Oct-21	16-Nov-01	16-Nov-11	16-Nov-21	16-Dec-01	16-Dec-11	16-Dec-21
22	38	107	925	666	974	978	865
45.3	48.2	48.3	43.9	41.3	45.5	43.2	43.2
0.49	0.79	2.21	21.07	16.11	21.41	22.64	20.04

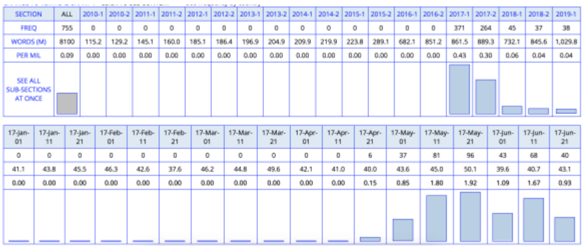
Figure 9: Frequency of *fake news* by 10 day period

The NOW corpus can also be used to examine cultural shifts. For example, Google Trends (which measures the frequency of searches, but not the actual frequency of a word or phrase in texts), shows that people started searching for *fidget spinner* in April 2017, that it reached its peak in mid-May 2017, and that it largely disappeared by June/July 2017. The NOW corpus (Figure11; based on actual occurrences in texts) shows the same thing:



Google Search popularity of fidget spinners in early 2017.

Figure 10: *fidget spinner* in Google Trends

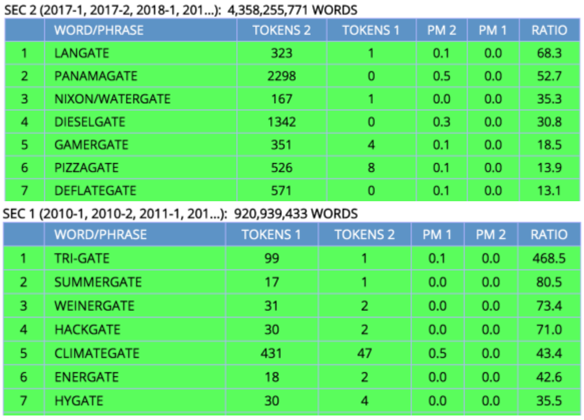


SECTION	ALL	2010-1	2010-2	2011-1	2011-2	2012-1	2012-2	2013-1	2013-2	2014-1	2014-2	2015-1	2015-2	2016-1	2016-2	2017-1	2017-2	2018-1	2018-2	2019-1
FREQ	755	0	0	0	0	0	0	0	0	0	0	0	0	0	0	371	264	45	37	38
WORDS (M)	8100	115.2	129.2	145.1	160.0	185.1	186.4	196.9	204.9	209.9	219.9	223.8	289.1	682.1	851.2	861.5	889.3	732.1	845.6	1,029.8
PER MIL	0.09	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.43	0.30	0.06	0.04	0.04
SEE ALL SUB-SECTIONS AT ONCE																				

17-Jan-01	17-Jan-11	17-Jan-21	17-Feb-01	17-Feb-11	17-Feb-21	17-Mar-01	17-Mar-11	17-Mar-21	17-Apr-01	17-Apr-11	17-Apr-21	17-May-01	17-May-11	17-May-21	17-Jun-01	17-Jun-11	17-Jun-21
0	0	0	0	0	0	0	0	0	0	0	6	37	81	96	43	68	40
41.1	43.8	45.5	46.3	42.6	37.6	46.2	44.8	49.6	42.1	41.0	40.0	43.6	45.0	50.1	39.6	40.7	43.1
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.15	0.85	1.80	1.92	1.09	1.67	0.93

Figure 11: *fidget spinner* in NOW by 10 day period

3.1 The corpus architecture 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, the following chart shows words ending in **gate* (sometimes indicating "scandal") that are more frequent in 2017-2019 (top; e.g. *Panamagate*, *dieselgate*, *deflategate*) compared to 2010-2013 (bottom; e.g. *hackgate*, *cablegate*, *climategate*):



SEC 2 (2017-1, 2017-2, 2018-1, 201...): 4,358,255,771 WORDS

	WORD/PHRASE	TOKENS 2	TOKENS 1	PM 2	PM 1	RATIO
1	LANGATE	323	1	0.1	0.0	68.3
2	PANAMAGATE	2298	0	0.5	0.0	52.7
3	NIXON/WATERGATE	167	1	0.0	0.0	35.3
4	DIESELGATE	1342	0	0.3	0.0	30.8
5	GAMERGATE	351	4	0.1	0.0	18.5
6	PIZZAGATE	526	8	0.1	0.0	13.9
7	DEFLATEGATE	571	0	0.1	0.0	13.1

SEC 1 (2010-1, 2010-2, 2011-1, 201...): 920,939,433 WORDS

	WORD/PHRASE	TOKENS 1	TOKENS 2	PM 1	PM 2	RATIO
1	TRI-GATE	99	1	0.1	0.0	468.5
2	SUMMERGATE	17	1	0.0	0.0	80.5
3	WEINERGATE	31	2	0.0	0.0	73.4
4	HACKGATE	30	2	0.0	0.0	71.0
5	CLIMATEGATE	431	47	0.5	0.0	43.4
6	ENERGATE	18	2	0.0	0.0	42.6
7	HYGATE	30	4	0.0	0.0	35.5

Figure 12: Comparison of **gate* words 2017-2019 (top) vs 2010-2012 (bottom)

And of course researchers can compare new phrases as well (rather than just words). For example, the following are all new phrases with smart NOUN that are at least 20 times as frequent in 2017-2019 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*.

3.2 In addition to looking at changes in lexis and phraseology, researchers can also use NOW to look at very recent changes in syntax. The impression has often been that syntax changes so slowly that a corpus with just a ten year time span (as with NOW; 2010-2019) wouldn't show much change during this short period. But cases of syntactic change during just the last ten years are not hard to find

For example, the frequency of the perfect progressive (HAVE+been+VERB-ing: *has been working*) has increased about 10% during the last ten years, from less than 260 tokens per million words in 2010-2011, to 280-290 tokens per million words in 2017-2019.

Likewise, there have been changes in verbal subcategorization during just the last few years. For example, Figure 13 shows an increase in the "bare infinitive" with *help* (e.g. *they helped me -- clean the room*) compared to the "to infinitive" (*they helped me* ***<u>to</u>*** *clean the room*) since 2010. (The figure shows the percentage of all tokens that are the bare infinitive. For more on the construction, which has been a favorite of corpus linguistics, see Kjellmer 1985, Mair 2002, Rohdenburg 2009, and Callies 2015.)

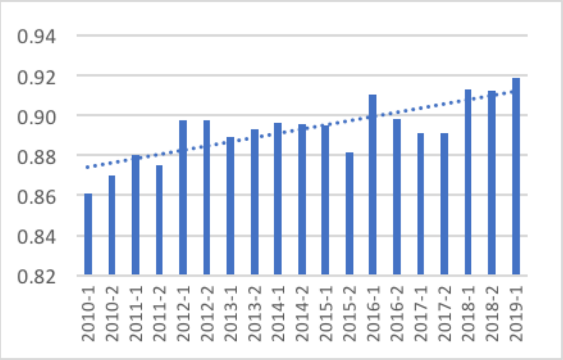
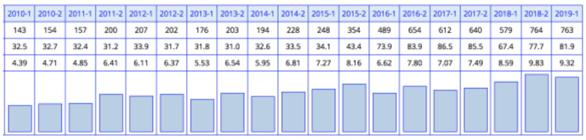


Figure 13: % HELP PRON -- VERB

Finally, it is possible to see change in just a given variety (or group of varieties) of English, such as British, American, or Singaporean English. For example, Figure 14 shows the increase in *gotten* as a past participle (e.g. *I've <u>gotten</u> over the guilt*) compared to the more common *got* (*I've <u>got</u> over the guilt*) in British English.



2010-1	2010-2	2011-1	2011-2	2012-1	2012-2	2013-1	2013-2	2014-1	2014-2	2015-1	2015-2	2016-1	2016-2	2017-1	2017-2	2018-1	2018-2	2019-1
143	154	157	200	207	202	176	203	194	228	248	354	489	654	612	640	579	764	763
32.5	32.7	32.4	31.2	33.9	31.7	31.8	31.0	32.6	33.5	34.1	43.4	73.9	83.9	86.5	85.5	67.4	77.7	81.9
4.39	4.71	4.85	6.41	6.11	6.37	5.53	6.54	5.95	6.81	7.27	8.16	6.62	7.80	7.07	7.49	8.59	9.83	9.32

Figure 14: (HAVE+) *gotten* in British English

Whereas the normalized frequency was less than five tokens per million words in 2010-2011, it is nearly twice that (8.6 to 9.8 tokens per million words) in 2018-2019. Because we can focus on both different time periods and different varieties in NOW, we can use the corpus to see how linguistic changes spread from one dialect to another over time.

In summary, NOW allows us to look at ongoing changes in English in ways that are not possible with any other corpus. This is due to two features that NOW has, which are not found together in any other corpus – its very large size and the fact that it has been updated on a regular basis (every month), up to the current time.

4 Problems and challenges

In spite of the possibilities with a continually updated corpus like NOW, there are also some challenges – compared to "static" corpora like iWeb.

First, as was explained in Section 2, the SQL Server database relies heavily on "clustered" indexes for search speed. This means that data is physically stored on the SSD – one row next to another – according to whatever column we choose. Therefore, when new data is added to the corpus (for example, 170-180 million words each month for NOW), the new rows of data need to be placed (on the SSD) adjacent to the existing rows. For example, all of the rows for the word *market* need to be physically placed between *market* and the next word (such as *marketable*). If the "fill factor" is not set high enough, millions of rows of data will need to be moved on the SSD to make room for the new rows of data. This can be very slow, even for SSDs.

Second, in iWeb we could create n-gram databases to handle very high frequency searches, like "VERB the NOUN" or "NOUN NOUN". With the NOW corpus, we would need to rebuild these every time the corpus is updated, such as every month. Because the corpus is now so large (more than 8 billion words), this would be computationally quite expensive to do each month. As a result, we do not use n-grams for NOW, which means that some very high frequency search strings (e.g. NOUN NOUN) are disallowed.

Third, there is other data that is pre-processed in iWeb that would be expensive to pre-process every month in NOW, such as collocates. The only reason that collocates are even doable in iWeb or the Sketch Engine corpora is because they *are* pre-processed. But the collocates would need to be pre-processed again for all 60,000 lemmas whenever new data is added to the corpus, and that can take a full day or two. And unless the collocates are re-generated each month, the collocates data will gradually become more and more outdated until they are updated again.

One might claim that in principle other architectures that are designed for "static" corpora *should* be able to use preprocessing strategies for incrementally updated values (such as ngram indices or term frequencies). But we are not aware of any other very large corpora that *actually employ* such an approach, for corpora that are updated every day or even every month. And while term frequencies can be easily updated, other data such as collocates and n-grams will take a significant amount of time, to say nothing of the basic "clustered" data, as explained above.

5 Conclusion

In summary, the NOW corpus provides at least two important advantages. First, it is very large – currently more than 8 billion words in size. Second, unlike most other large corpora, it is continually updated – by about 150-170 million words each month, or 1.5 billion words each year. The combination of these two features allows it to model ongoing linguistic change in English in ways that are not possible with any other corpus.

Due to its relational database architecture (which uses an architecture similar to sharding in columnstore databases, including clustered indexes), most searches (words, substrings, phrase, and even grammatical constructions; cf. "HELP PRON (to) VERB" shown above) are only 4-5% slower in an 8 billion word corpus (the current size of NOW) than in a 3-4 billion word corpus (the size of NOW in 2015).

But some searches (such as very high frequency strings like NOUN NOUN, which are based on n-grams), or queries that use pre-processed data (such as collocates) can still present a challenge in these dynamic corpora.

References

Bick, Eckhard. 1999. *The parsing system Palavras,* Aarhus: Aarhus Univ. Press.

Callies, Marcus. 2013. Bare infinitival complements in Present-Day English. In *The Verb Phrase in English: Investigating Recent Language Change with Corpora*. Ed. Bas Aarts, et al. Cambridge: CUP, 239-255.

Davies, Mark and Jong-Bok Kim. 2019. The advantages and challenges of 'big data': Insights from the 14 billion word iWeb corpus. *Linguistic Research* 36: 1-34.

Davies, Mark. 2018. Corpus-based studies of lexical and semantic variation: The importance of both corpus size and corpus design. In *From data to evidence in English language research*. Ed. Carla Suhr, et al. Leiden: Brill. 34-55.

Davies, Mark. 2017. Using Large Online Corpora to Examine Lexical, Semantic, and Cultural Variation in Different Dialects and Time Periods. In *Corpus-Based Sociolinguistics*. Ed. Eric Friginal et al. London: Routledge. 19-82.

Davies, Mark and Robert Fuchs. 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: 1-28.

Garside, R., and Smith, N. 1997. A hybrid grammatical tagger: CLAWS4/ *Corpus Annotation: Linguistic Information from Computer Text Corpora*. Ed. Roger Garside, et al. Longman, London, pp. 102-121

Kilgarriff, Adam, Pavel Rychlý, Pavel Smrž, David Tugwell. Itri-04-08 the sketch engine. *Information Technology*, 2004.

Kjellmer, Göran. 1985. Help to/help – revisited. *English Studies* 66: 156-61.

Mair, Christian. 2002. Three Changing Patterns of Verb Complementation in Late Modern English: A Real-Time Study Based on Matching Text Corpora/ *English Language and Linguistics* 6: 105-131.

Pomikálek, Jan. 2011. Removing Boilerplate and Duplicate Content from Web Corpora. Ph.D. thesis, Univ. Masaryk.

Rohdenburg, Gunter. 2009. Grammatical Divergence between British and American English in the Nineteenth and Early Twentieth Centuries. *Linguistic Insights - Studies in Language and Communication* 77:301-329.

Schäfer, Roland. 2015. Processing and querying large web corpora with the COW14 architecture. In *Proceedings of the 3rd Workshop on Challenges in the Management of Large Corpora (CMLC-3)*, Ed. Piotr Banski, et al.