English-Corpora.org: a guided tour

Mark Davies, Professor of Linguistics

November 2020

English-Corpora.org is the **most widely used** collection of corpora (highly searchable collections of texts) anywhere in the world. The corpora are used by more than 130,000 people each month, from more than 140 countries. In addition, hundreds of universities worldwide have **academic licenses**, which provide their users with expanded access to the corpora.

The corpora have been used as the basis of **thousands of academic articles**, theses, and dissertations, and they form the backbone of **courses on language and linguistics** throughout the world, at all levels of instruction. Virtually every book on "teaching English with corpora" in the last 5-10 years has focused primarily on these corpora (which are also sometimes called the "BYU Corpora", for the university where they were created).

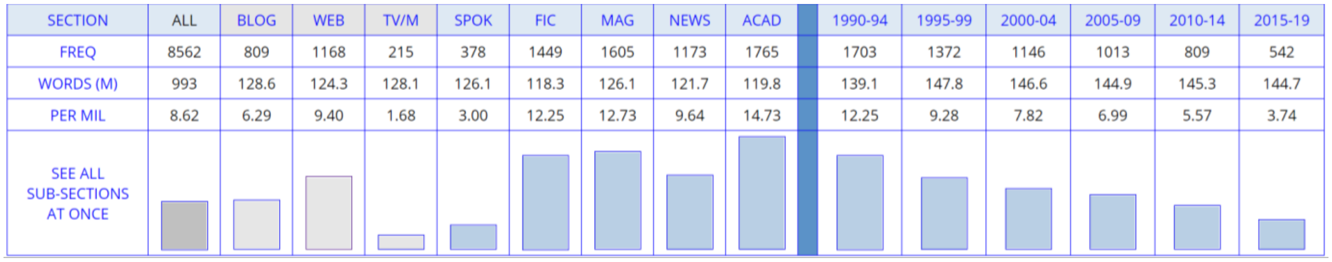
Since the first corpora were released in 2005, a total of seventeen corpora have been created:

	Corpus	**# words**	**Dialect**	**Time period**	**Genre(s)**
1	iWeb: The Intelligent Web-based Corpus	14 **billion**	6 countries	2017	Web
2	News on the Web (NOW)	11.3 **billion+**	20 countries	2010-yesterday	Web: News
3	Global Web-Based English (GloWbE)	1.9 **billion**	20 countries	2012-13	Web (incl blogs)
4	Wikipedia Corpus	1.9 **billion**	(Various)	2014	Wikipedia
5	Hansard Corpus	1.6 **billion**	British	1803-2005	Parliament
6	Corpus of Contemporary American English (COCA)	1.0 **billion**	American	1990-2019	Balanced
7	Early English Books Online	755 million	British	1470s-1690s	(Various)
8	Coronavirus Corpus	673 million+	20 countries	2020-yesterday	Web: News
9	Corpus of Historical American English (COHA)	400 million	American	1810-2009	Balanced
10	The TV Corpus	325 million	6 countries	1950-2018	TV shows
11	The Movie Corpus	200 million	6 countries	1930-2018	Movies
12	Corpus of US Supreme Court Opinions	130 million	American	1790s-present	Legal opinions
13	Corpus of American Soap Operas	100 million	American	2001-2012	TV shows
14	British National Corpus (BNC)	100 million	British	1980s-1993	Balanced
15	TIME Magazine Corpus	100 million	American	1923-2006	Magazine
16	Strathy Corpus (Canada)	50 million	Canadian	1970s-2000s	Balanced
17	CORE Corpus	50 million	6 countries	2014	Web

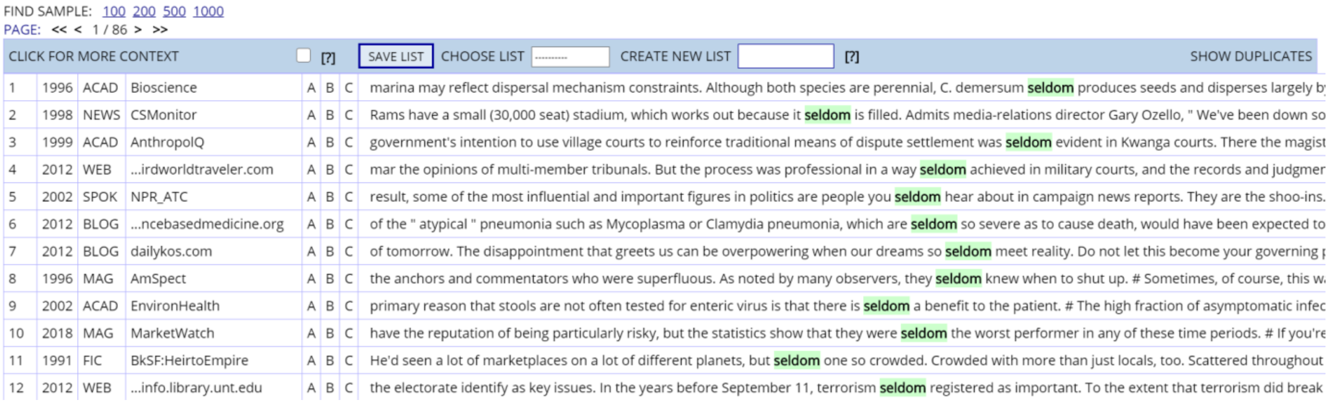
Why variation matters (a lot) (go to beginning)

What sets English-Corpora.org apart from all other corpora is the insight that they give into **variation in English** – between genres, historical periods, and dialects. Other corpora are just giant "blobs" of data, with little if any indication of variation. Why is this important? Consider the simple word *seldom*. As COCA (the one billion word Corpus of Contemporary American English) shows, this word is used much more in formal genres than in informal genres, and its use is sharply declining over time.

(Note: in the case of *seldom* and all other searches in this file, click on the blue link to run the search)



SECTION	ALL	BLOG	WEB	TV/M	SPOK	FIC	MAG	NEWS	ACAD	1990-94	1995-99	2000-04	2005-09	2010-14	2015-19
FREQ	8562	809	1168	215	378	1449	1605	1173	1765	1703	1372	1146	1013	809	542
WORDS (M)	993	128.6	124.3	128.1	126.1	118.3	126.1	121.7	119.8	139.1	147.8	146.6	144.9	145.3	144.7
PER MIL	8.62	6.29	9.40	1.68	3.00	12.25	12.73	9.64	14.73	12.25	9.28	7.82	6.99	5.57	3.74
SEE ALL SUB-SECTIONS AT ONCE															



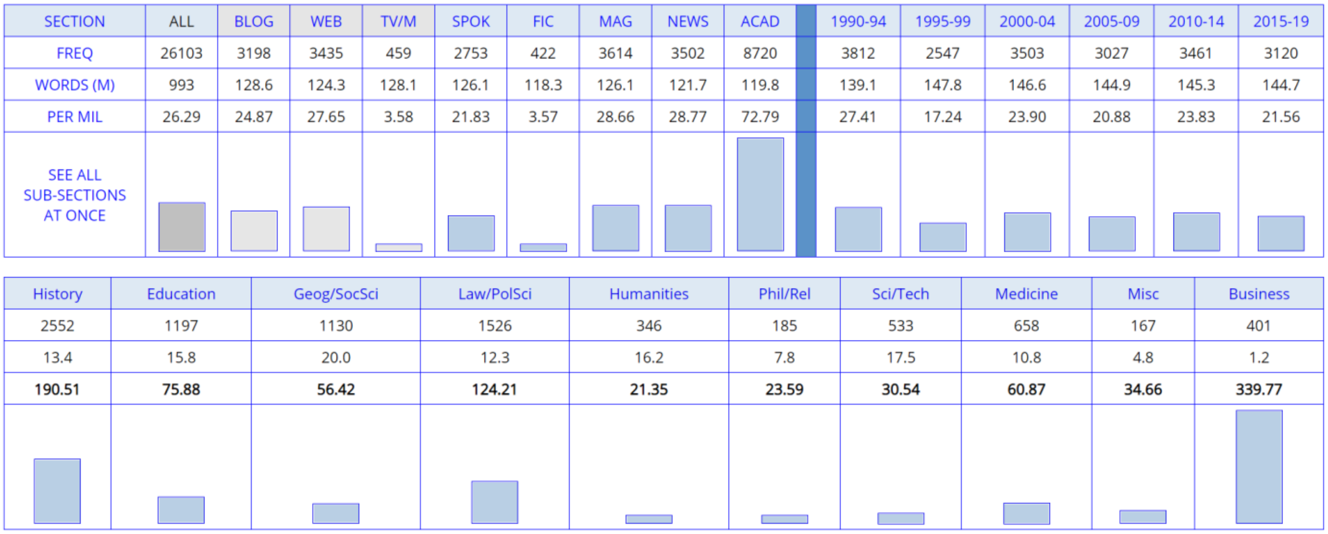
FIND SAMPLE: 100 200 500 1000
PAGE: << < 1 / 86 > >>

CLICK FOR MORE CONTEXT				☐ [?]		SAVE LIST CHOOSE LIST --------- CREATE NEW LIST [?] SHOW DUPLICATES
1	1996	ACAD	Bioscience	A	B C	marina may reflect dispersal mechanism constraints. Although both species are perennial, C. demersum **seldom** produces seeds and disperses largely b
2	1998	NEWS	CSMonitor	A	B C	Rams have a small (30,000 seat) stadium, which works out because it **seldom** is filled. Admits media-relations director Gary Ozello, " We've been down so
3	1999	ACAD	AnthropolQ	A	B C	government's intention to use village courts to reinforce traditional means of dispute settlement was **seldom** evident in Kwanga courts. There the magist
4	2012	WEB	...irdworldtraveler.com	A	B C	mar the opinions of multi-member tribunals. But the process was professional in a way **seldom** achieved in military courts, and the records and judgmer
5	2002	SPOK	NPR_ATC	A	B C	result, some of the most influential and important figures in politics are people you **seldom** hear about in campaign news reports. They are the shoo-ins.
6	2012	BLOG	...ncebasedmedicine.org	A	B C	of the " atypical " pneumonia such as Mycoplasma or Clamydia pneumonia, which are **seldom** so severe as to cause death, would have been expected to
7	2012	BLOG	dailykos.com	A	B C	of tomorrow. The disappointment that greets us can be overpowering when our dreams so **seldom** meet reality. Do not let this become your governing
8	1996	MAG	AmSpect	A	B C	the anchors and commentators who were superfluous. As noted by many observers, they **seldom** knew when to shut up. # Sometimes, of course, this w
9	2002	ACAD	EnvironHealth	A	B C	primary reason that stools are not often tested for enteric virus is that there is **seldom** a benefit to the patient. # The high fraction of asymptomatic infec
10	2018	MAG	MarketWatch	A	B C	have the reputation of being particularly risky, but the statistics show that they were **seldom** the worst performer in any of these time periods. # If you're
11	1991	FIC	BkSF:HeirtoEmpire	A	B C	He'd seen a lot of marketplaces on a lot of different planets, but **seldom** one so crowded. Crowded with more than just locals, too. Scattered throughout
12	2012	WEB	...info.library.unt.edu	A	B C	the electorate identify as key issues. In the years before September 11, terrorism **seldom** registered as important. To the extent that terrorism did break

If a large online corpus simply says that *seldom* occurs 87,000 times in a 17 billion word corpus, that is not very useful. Students would never know that if they use this word, they will sound like 1) a 70-80 year old person and/or 2) someone in a formal setting. This is just one simple example, dealing with word frequency. But this applies to thousands of words (frequency, meaning, and usage) and many grammatical constructions as well. Variation matters a great deal, and English-Corpora.org has the **only corpora that show this variation** in such detail.

Word frequency (go to beginning)

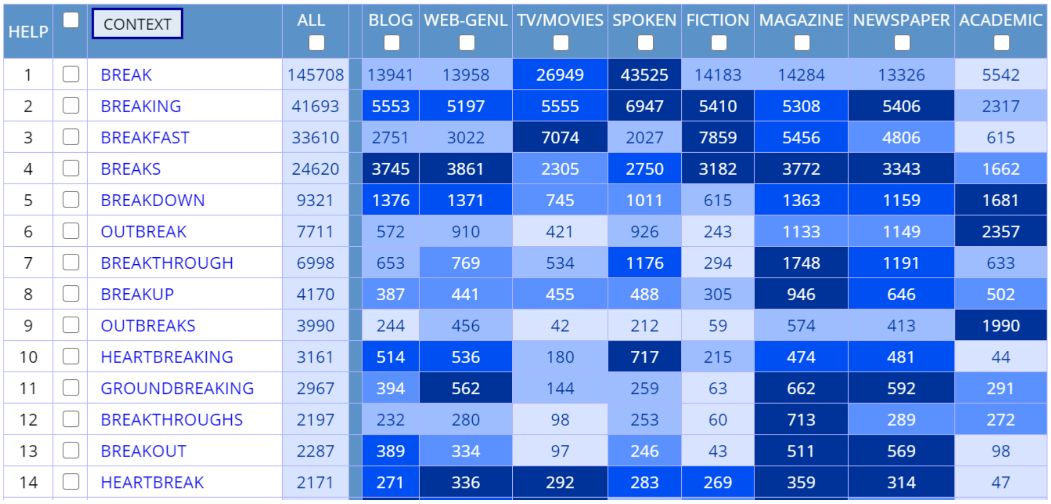
At the most basic level, users can see the **frequency of any word or phrase** in the different sections of the corpus, as well as sub-sections (in certain corpora). For example, they can see that *strategic* occurs most frequently in academic texts in COCA, and within the academic genre, it is the most frequent in business, history, and law / political science.



SECTION	ALL	BLOG	WEB	TV/M	SPOK	FIC	MAG	NEWS	ACAD		1990-94	1995-99	2000-04	2005-09	2010-14	2015-19
FREQ	26103	3198	3435	459	2753	422	3614	3502	8720		3812	2547	3503	3027	3461	3120
WORDS (M)	993	128.6	124.3	128.1	126.1	118.3	126.1	121.7	119.8		139.1	147.8	146.6	144.9	145.3	144.7
PER MIL	26.29	24.87	27.65	3.58	21.83	3.57	28.66	28.77	72.79		27.41	17.24	23.90	20.88	23.83	21.56
SEE ALL SUB-SECTIONS AT ONCE																

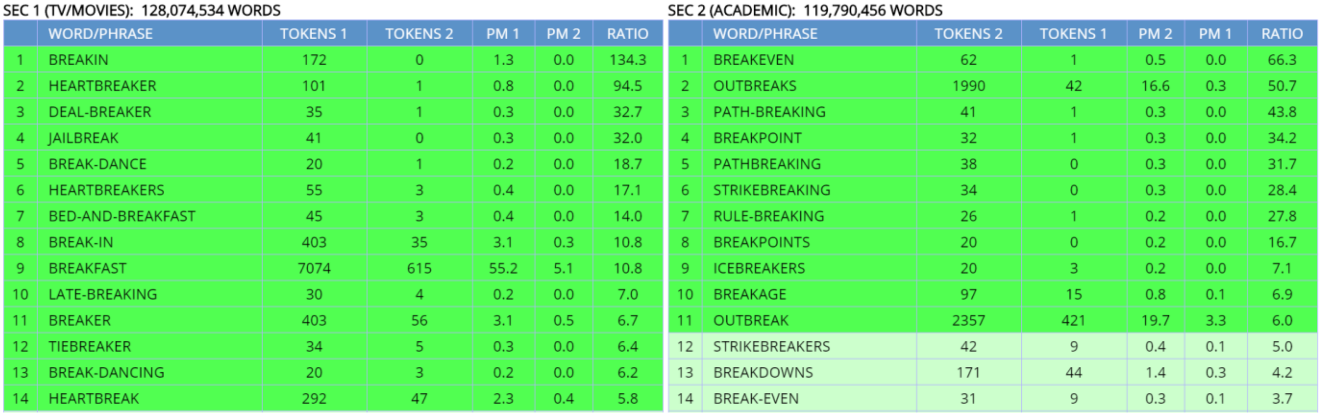
History	Education	Geog/SocSci	Law/PolSci	Humanities	Phil/Rel	Sci/Tech	Medicine	Misc	Business
2552	1197	1130	1526	346	185	533	658	167	401
13.4	15.8	20.0	12.3	16.2	7.8	17.5	10.8	4.8	1.2
190.51	**75.88**	**56.42**	**124.21**	**21.35**	**23.59**	**30.54**	**60.87**	**34.66**	**339.77**

Users can search for any word, phrase, or substring (e.g. words with *_break_*), and see **all matching forms** in the different sections of the corpus. For example, COCA shows the frequency in blogs, other web pages, TV/Movie subtitles, unscripted spoken TV and radio programs, fiction, magazines, newspapers, and academic journals.



HELP	□	CONTEXT	ALL □	BLOG □	WEB-GENL □	TV/MOVIES □	SPOKEN □	FICTION □	MAGAZINE □	NEWSPAPER □	ACADEMIC □
1	□	BREAK	145708	13941	13958	26949	43525	14183	14284	13326	5542
2	□	BREAKING	41693	5553	5197	5555	6947	5410	5308	5406	2317
3	□	BREAKFAST	33610	2751	3022	7074	2027	7859	5456	4806	615
4	□	BREAKS	24620	3745	3861	2305	2750	3182	3772	3343	1662
5	□	BREAKDOWN	9321	1376	1371	745	1011	615	1363	1159	1681
6	□	OUTBREAK	7711	572	910	421	926	243	1133	1149	2357
7	□	BREAKTHROUGH	6998	653	769	534	1176	294	1748	1191	633
8	□	BREAKUP	4170	387	441	455	488	305	946	646	502
9	□	OUTBREAKS	3990	244	456	42	212	59	574	413	1990
10	□	HEARTBREAKING	3161	514	536	180	717	215	474	481	44
11	□	GROUNDBREAKING	2967	394	562	144	259	63	662	592	291
12	□	BREAKTHROUGHS	2197	232	280	98	253	60	713	289	272
13	□	BREAKOUT	2287	389	334	97	246	43	511	569	98
14	□	HEARTBREAK	2171	271	336	292	283	269	359	314	47

They can also compare any set of sections in a corpus, such as words with *_break_* that occur much more in (very informal) TV/Movies subtitles (left), compared to much more formal academic texts (right).



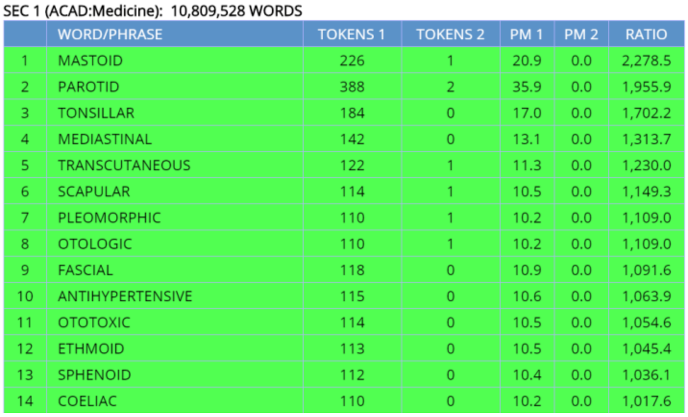
SEC 1 (TV/MOVIES): 128,074,534 WORDS

	WORD/PHRASE	TOKENS 1	TOKENS 2	PM 1	PM 2	RATIO
1	BREAKIN	172	0	1.3	0.0	134.3
2	HEARTBREAKER	101	1	0.8	0.0	94.5
3	DEAL-BREAKER	35	1	0.3	0.0	32.7
4	JAILBREAK	41	0	0.3	0.0	32.0
5	BREAK-DANCE	20	1	0.2	0.0	18.7
6	HEARTBREAKERS	55	3	0.4	0.0	17.1
7	BED-AND-BREAKFAST	45	3	0.4	0.0	14.0
8	BREAK-IN	403	35	3.1	0.3	10.8
9	BREAKFAST	7074	615	55.2	5.1	10.8
10	LATE-BREAKING	30	4	0.2	0.0	7.0
11	BREAKER	403	56	3.1	0.5	6.7
12	TIEBREAKER	34	5	0.3	0.0	6.4
13	BREAK-DANCING	20	3	0.2	0.0	6.2
14	HEARTBREAK	292	47	2.3	0.4	5.8

SEC 2 (ACADEMIC): 119,790,456 WORDS

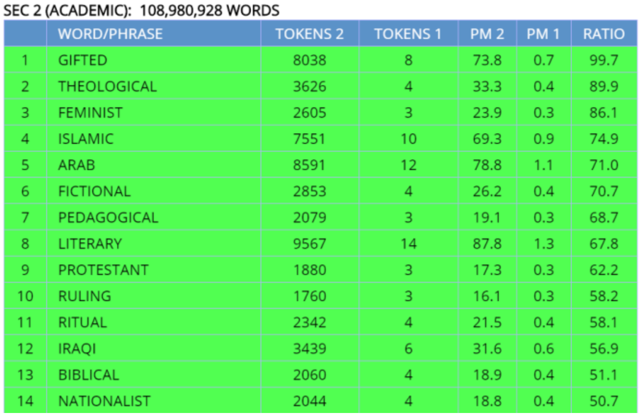
	WORD/PHRASE	TOKENS 2	TOKENS 1	PM 2	PM 1	RATIO
1	BREAKEVEN	62	1	0.5	0.0	66.3
2	OUTBREAKS	1990	42	16.6	0.3	50.7
3	PATH-BREAKING	41	1	0.3	0.0	43.8
4	BREAKPOINT	32	1	0.3	0.0	34.2
5	PATHBREAKING	38	0	0.3	0.0	31.7
6	STRIKEBREAKING	34	0	0.3	0.0	28.4
7	RULE-BREAKING	26	1	0.2	0.0	27.8
8	BREAKPOINTS	20	0	0.2	0.0	16.7
9	ICEBREAKERS	20	3	0.2	0.0	7.1
10	BREAKAGE	97	15	0.8	0.1	6.9
11	OUTBREAK	2357	421	19.7	3.3	6.0
12	STRIKEBREAKERS	42	9	0.4	0.1	5.0
13	BREAKDOWNS	171	44	1.4	0.3	4.2
14	BREAK-EVEN	31	9	0.3	0.1	3.7

Researchers can also see *all* words that are used much more in one genre (or sub-genre) than in another. For example, the words at the left are words that are used in COCA: Academic: Medicine than in COCA: Academic generally. Users could easily find words related to any domain, such as business, medicine, law, or engineering.



SEC 1 (ACAD:Medicine): 10,809,528 WORDS

	WORD/PHRASE	TOKENS 1	TOKENS 2	PM 1	PM 2	RATIO
1	MASTOID	226	1	20.9	0.0	2,278.5
2	PAROTID	388	2	35.9	0.0	1,955.9
3	TONSILLAR	184	0	17.0	0.0	1,702.2
4	MEDIASTINAL	142	0	13.1	0.0	1,313.7
5	TRANSCUTANEOUS	122	1	11.3	0.0	1,230.0
6	SCAPULAR	114	1	10.5	0.0	1,149.3
7	PLEOMORPHIC	110	1	10.2	0.0	1,109.0
8	OTOLOGIC	110	1	10.2	0.0	1,109.0
9	FASCIAL	118	0	10.9	0.0	1,091.6
10	ANTIHYPERTENSIVE	115	0	10.6	0.0	1,063.9
11	OTOTOXIC	114	0	10.5	0.0	1,054.6
12	ETHMOID	113	0	10.5	0.0	1,045.4
13	SPHENOID	112	0	10.4	0.0	1,036.1
14	COELIAC	110	0	10.2	0.0	1,017.6

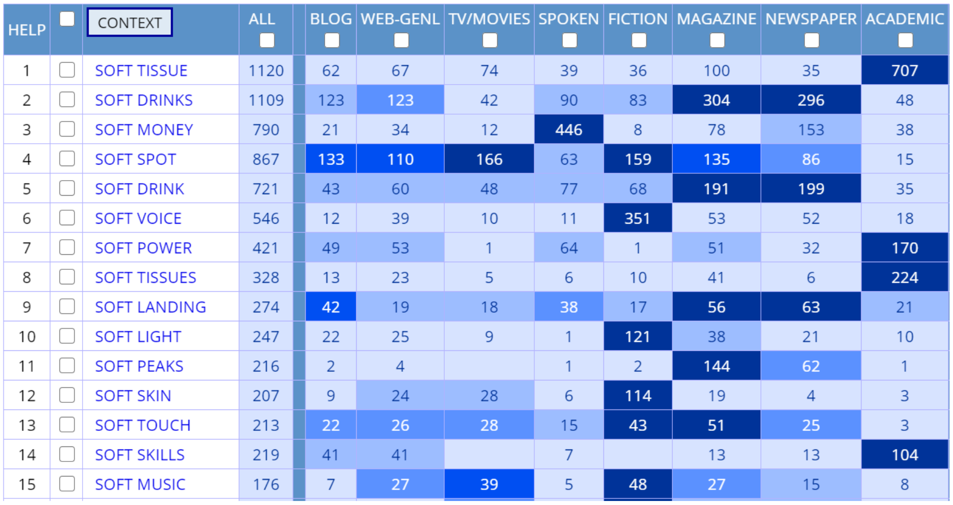


SEC 2 (ACADEMIC): 108,980,928 WORDS

	WORD/PHRASE	TOKENS 2	TOKENS 1	PM 2	PM 1	RATIO
1	GIFTED	8038	8	73.8	0.7	99.7
2	THEOLOGICAL	3626	4	33.3	0.4	89.9
3	FEMINIST	2605	3	23.9	0.3	86.1
4	ISLAMIC	7551	10	69.3	0.9	74.9
5	ARAB	8591	12	78.8	1.1	71.0
6	FICTIONAL	2853	4	26.2	0.4	70.7
7	PEDAGOGICAL	2079	3	19.1	0.3	68.7
8	LITERARY	9567	14	87.8	1.3	67.8
9	PROTESTANT	1880	3	17.3	0.3	62.2
10	RULING	1760	3	16.1	0.3	58.2
11	RITUAL	2342	4	21.5	0.4	58.1
12	IRAQI	3439	6	31.6	0.6	56.9
13	BIBLICAL	2060	4	18.9	0.4	51.1
14	NATIONALIST	2044	4	18.8	0.4	50.7

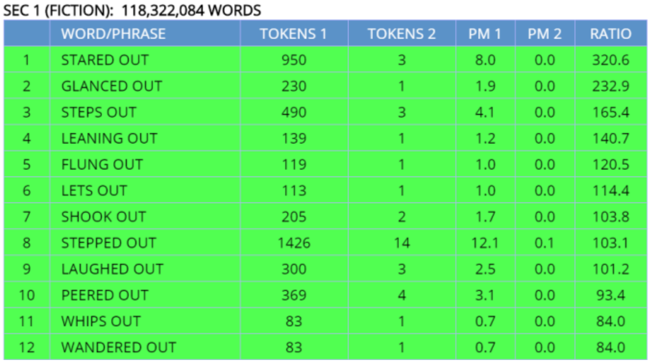
Phrases and collocations (strings of words) (go to beginning)

Of course, users can search for much more than individual words. The following table shows phrases with *soft + NOUN* in the different genres of COCA. Notice *soft tissue(s), power, skills* in academic, *soft spot* in TV/Movies, *soft voice, light, skin, touch, music* in fiction, and *soft drink(s)* or *landing* in newspapers and magazines. Again, a large "blob" of 15-20 billion words – with no indication of genre – would miss out on all of this.



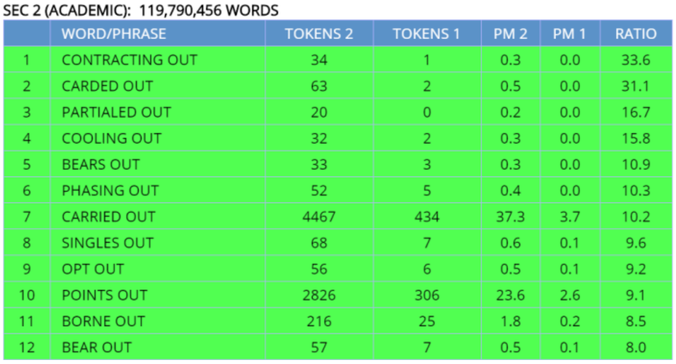
HELP	□	CONTEXT	ALL □	BLOG □	WEB-GENL □	TV/MOVIES □	SPOKEN □	FICTION □	MAGAZINE □	NEWSPAPER □	ACADEMIC □
1	□	SOFT TISSUE	1120	62	67	74	39	36	100	35	707
2	□	SOFT DRINKS	1109	123	123	42	90	83	304	296	48
3	□	SOFT MONEY	790	21	34	12	446	8	78	153	38
4	□	SOFT SPOT	867	133	110	166	63	159	135	86	15
5	□	SOFT DRINK	721	43	60	48	77	68	191	199	35
6	□	SOFT VOICE	546	12	39	10	11	351	53	52	18
7	□	SOFT POWER	421	49	53	1	64	1	51	32	170
8	□	SOFT TISSUES	328	13	23	5	6	10	41	6	224
9	□	SOFT LANDING	274	42	19	18	38	17	56	63	21
10	□	SOFT LIGHT	247	22	25	9	1	121	38	21	10
11	□	SOFT PEAKS	216	2	4		1	2	144	62	1
12	□	SOFT SKIN	207	9	24	28	6	114	19	4	3
13	□	SOFT TOUCH	213	22	26	28	15	43	51	25	3
14	□	SOFT SKILLS	219	41	41		7		13	13	104
15	□	SOFT MUSIC	176	7	27	39	5	48	27	15	8

Users can compare two sections of the corpora to find phrases that are much common in one section than the other. For example, these are phrasal verbs with *out* that are much more common in fiction (left) or academic (right).



SEC 1 (FICTION): 118,322,084 WORDS

	WORD/PHRASE	TOKENS 1	TOKENS 2	PM 1	PM 2	RATIO
1	STARED OUT	950	3	8.0	0.0	320.6
2	GLANCED OUT	230	1	1.9	0.0	232.9
3	STEPS OUT	490	3	4.1	0.0	165.4
4	LEANING OUT	139	1	1.2	0.0	140.7
5	FLUNG OUT	119	1	1.0	0.0	120.5
6	LETS OUT	113	1	1.0	0.0	114.4
7	SHOOK OUT	205	2	1.7	0.0	103.8
8	STEPPED OUT	1426	14	12.1	0.1	103.1
9	LAUGHED OUT	300	3	2.5	0.0	101.2
10	PEERED OUT	369	4	3.1	0.0	93.4
11	WHIPS OUT	83	1	0.7	0.0	84.0
12	WANDERED OUT	83	1	0.7	0.0	84.0

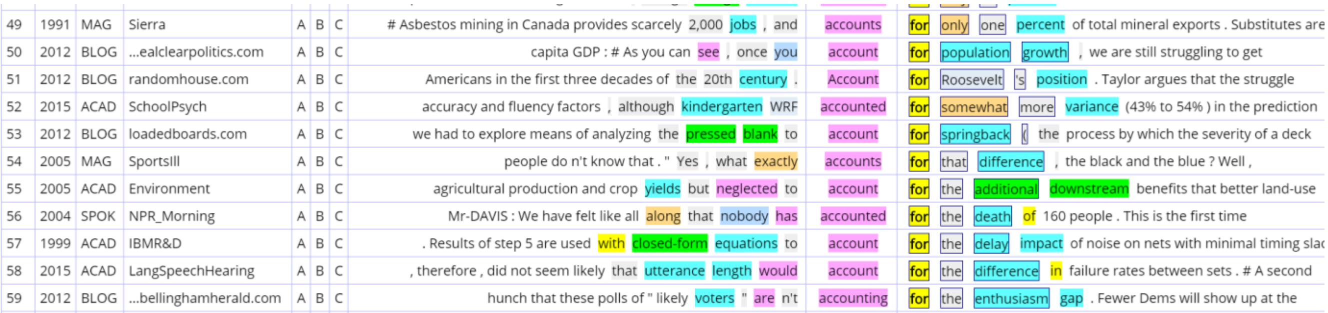


SEC 2 (ACADEMIC): 119,790,456 WORDS

	WORD/PHRASE	TOKENS 2	TOKENS 1	PM 2	PM 1	RATIO
1	CONTRACTING OUT	34	1	0.3	0.0	33.6
2	CARDED OUT	63	2	0.5	0.0	31.1
3	PARTIALED OUT	20	0	0.2	0.0	16.7
4	COOLING OUT	32	2	0.3	0.0	15.8
5	BEARS OUT	33	3	0.3	0.0	10.9
6	PHASING OUT	52	5	0.4	0.0	10.3
7	CARRIED OUT	4467	434	37.3	3.7	10.2
8	SINGLES OUT	68	7	0.6	0.1	9.6
9	OPT OUT	56	6	0.5	0.1	9.2
10	POINTS OUT	2826	306	23.6	2.6	9.1
11	BORNE OUT	216	25	1.8	0.2	8.5
12	BEAR OUT	57	7	0.5	0.1	8.0

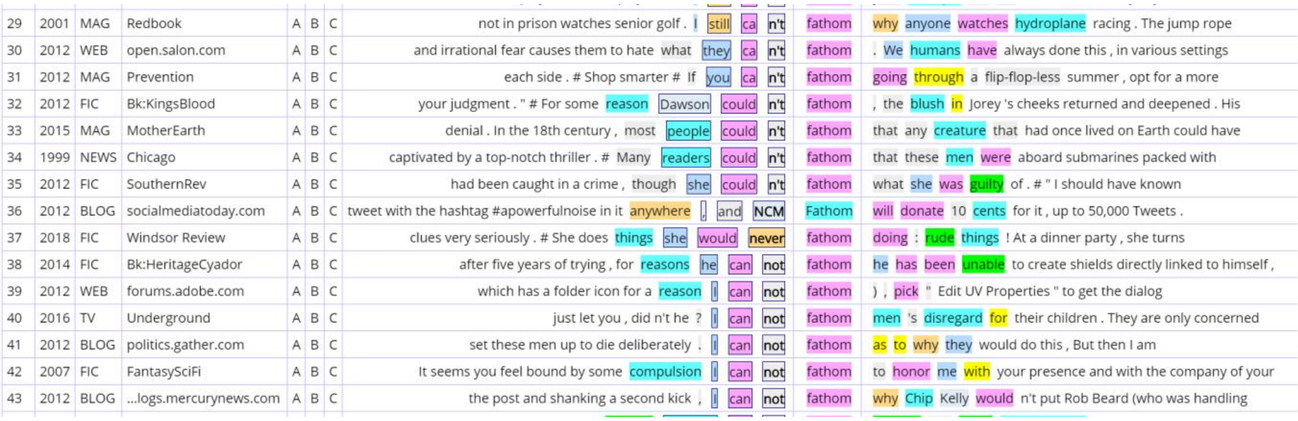
Patterns (go to beginning)

The corpora can also show the patterns in which words and phrases occur. Words do not occur in isolation, and learners need to understand the patterns that a given word takes. For example, *account* as a verb is nearly always followed by *for*:



49	1991	MAG	Sierra	A	B	C	# Asbestos mining in Canada provides scarcely 2,000 jobs , and	accounts	for only one percent of total mineral exports . Substitutes are
50	2012	BLOG	...ealclearpolitics.com	A	B	C	capita GDP : # As you can see , once you	account	for population growth , we are still struggling to get
51	2012	BLOG	randomhouse.com	A	B	C	Americans in the first three decades of the 20th century .	Account	for Roosevelt 's position . Taylor argues that the struggle
52	2015	ACAD	SchoolPsych	A	B	C	accuracy and fluency factors , although kindergarten WRF	accounted	for somewhat more variance (43% to 54%) in the prediction
53	2012	BLOG	loadedboards.com	A	B	C	we had to explore means of analyzing the pressed blank to	account	for springback (the process by which the severity of a deck
54	2005	MAG	SportsIll	A	B	C	people do n't know that . " Yes , what exactly	accounts	for that difference , the black and the blue ? Well ,
55	2005	ACAD	Environment	A	B	C	agricultural production and crop yields but neglected to	account	for the additional downstream benefits that better land-use
56	2004	SPOK	NPR_Morning	A	B	C	Mr-DAVIS : We have felt like all along that nobody has	accounted	for the death of 160 people . This is the first time
57	1999	ACAD	IBMR&D	A	B	C	. Results of step 5 are used with closed-form equations to	account	for the delay impact of noise on nets with minimal timing slac
58	2015	ACAD	LangSpeechHearing	A	B	C	, therefore , did not seem likely that utterance length would	account	for the difference in failure rates between sets . # A second
59	2012	BLOG	...bellinghamherald.com	A	B	C	hunch that these polls of " likely voters " are n't	accounting	for the enthusiasm gap . Fewer Dems will show up at the

And *fathom* is nearly always preceded by a negative word. This is why a sentence like *I totally fathom what you're saying* (without any negation before the verb) would sound strange to a native speaker.

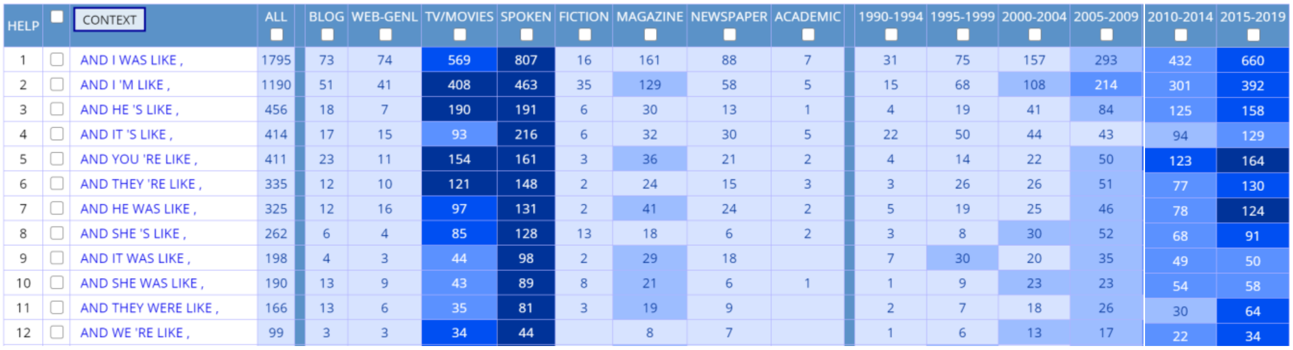


29	2001	MAG	Redbook	A	B	C	not in prison watches senior golf . I still ca n't	fathom	why anyone watches hydroplane racing . The jump rope
30	2012	WEB	open.salon.com	A	B	C	and irrational fear causes them to hate what they ca n't	fathom	. We humans have always done this , in various settings
31	2012	MAG	Prevention	A	B	C	each side . # Shop smarter # If you ca n't	fathom	going through a flip-flop-less summer , opt for a more
32	2012	FIC	Bk:KingsBlood	A	B	C	your judgment . " # For some reason Dawson could n't	fathom	, the blush in Jorey 's cheeks returned and deepened . His
33	2015	MAG	MotherEarth	A	B	C	denial . In the 18th century , most people could n't	fathom	that any creature that had once lived on Earth could have
34	1999	NEWS	Chicago	A	B	C	captivated by a top-notch thriller . # Many readers could n't	fathom	that these men were aboard submarines packed with
35	2012	FIC	SouthernRev	A	B	C	had been caught in a crime , though she could n't	fathom	what she was guilty of . # " I should have known
36	2012	BLOG	socialmediatoday.com	A	B	C	tweet with the hashtag #apowerfulnoise in it anywhere , and NCM	Fathom	will donate 10 cents for it , up to 50,000 Tweets .
37	2018	FIC	Windsor Review	A	B	C	clues very seriously . # She does things she would never	fathom	doing : rude things ! At a dinner party , she turns
38	2014	FIC	Bk:HeritageCyador	A	B	C	after five years of trying , for reasons he can not	fathom	he has been unable to create shields directly linked to himself ,
39	2012	WEB	forums.adobe.com	A	B	C	which has a folder icon for a reason I can not	fathom	) , pick " Edit UV Properties " to get the dialog
40	2016	TV	Underground	A	B	C	just let you , did n't he ? I can not	fathom	men 's disregard for their children . They are only concerned
41	2012	BLOG	politics.gather.com	A	B	C	set these men up to die deliberately . I can not	fathom	as to why they would do this , But then I am
42	2007	FIC	FantasySciFi	A	B	C	It seems you feel bound by some compulsion I can not	fathom	to honor me with your presence and with the company of your
43	2012	BLOG	...logs.mercurynews.com	A	B	C	the post and shanking a second kick , I can not	fathom	why Chip Kelly would n't put Rob Beard (who was handling

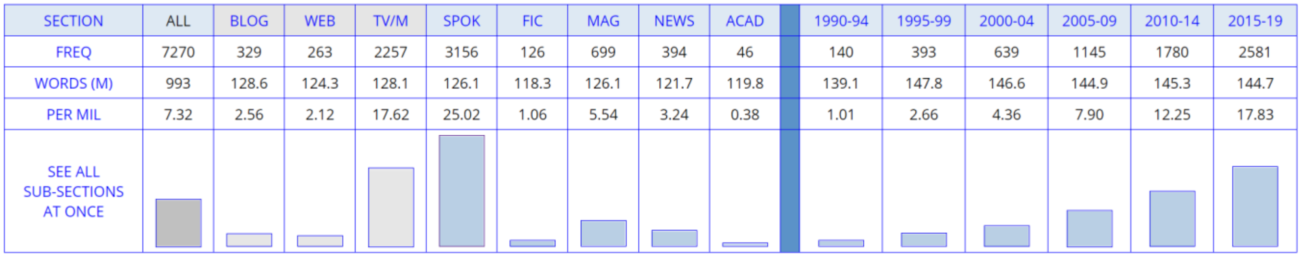
Corpora move far beyond a simple dictionary to show the patterns in which words occur.

Grammar / syntax (go to beginning)

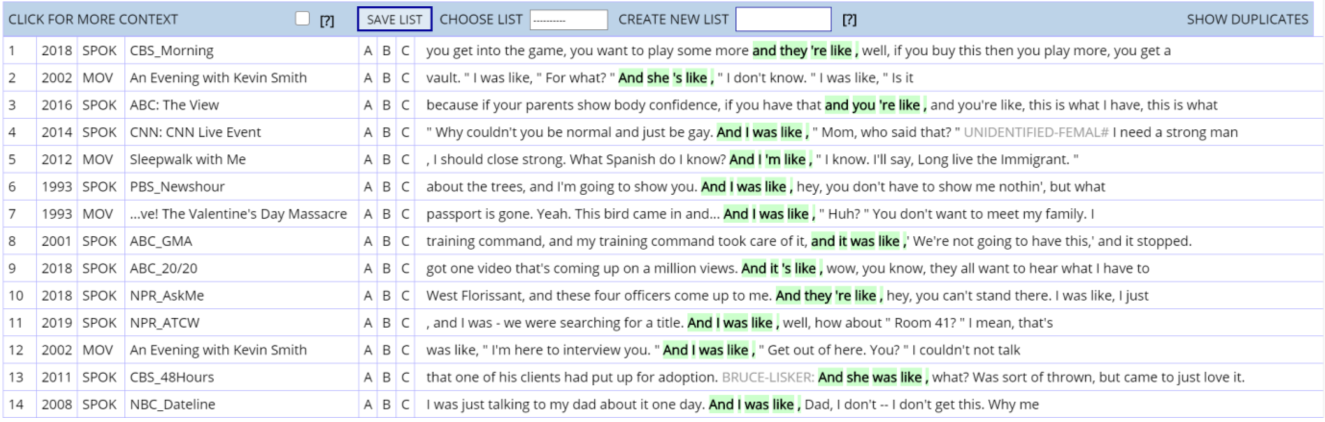
One of the best uses of the corpora is to look at the frequency and use of syntactic constructions. For example, consider the "like construction" (*and I'm like, he can't do it*, or *but she was like, let's just buy it*). The corpora can show the frequency of all matching phrases, as well as the frequency across sections of the corpus (in this case, genres and time periods 1990-2019 in COCA).



HELP	☐	CONTEXT	ALL ☐	BLOG ☐	WEB-GENL ☐	TV/MOVIES ☐	SPOKEN ☐	FICTION ☐	MAGAZINE ☐	NEWSPAPER ☐	ACADEMIC ☐	1990-1994 ☐	1995-1999 ☐	2000-2004 ☐	2005-2009 ☐	2010-2014 ☐	2015-2019 ☐
1	☐	AND I WAS LIKE ,	1795	73	74	569	807	16	161	88	7	31	75	157	293	432	660
2	☐	AND I 'M LIKE ,	1190	51	41	408	463	35	129	58	5	15	68	108	214	301	392
3	☐	AND HE 'S LIKE ,	456	18	7	190	191	6	30	13	1	4	19	41	84	125	158
4	☐	AND IT 'S LIKE ,	414	17	15	93	216	6	32	30	5	22	50	44	43	94	129
5	☐	AND YOU 'RE LIKE ,	411	23	11	154	161	3	36	21	2	4	14	22	50	123	164
6	☐	AND THEY 'RE LIKE ,	335	12	10	121	148	2	24	15	3	3	26	26	51	77	130
7	☐	AND HE WAS LIKE ,	325	12	16	97	131	2	41	24	2	5	19	25	46	78	124
8	☐	AND SHE 'S LIKE ,	262	6	4	85	128	13	18	6	2	3	8	30	52	68	91
9	☐	AND IT WAS LIKE ,	198	4	3	44	98	2	29	18		7	30	20	35	49	50
10	☐	AND SHE WAS LIKE ,	190	13	9	43	89	8	21	6	1	1	9	23	23	54	58
11	☐	AND THEY WERE LIKE ,	166	13	6	35	81	3	19	9		2	7	18	26	30	64
12	☐	AND WE 'RE LIKE ,	99	3	3	34	44		8	7		1	6	13	17	22	34



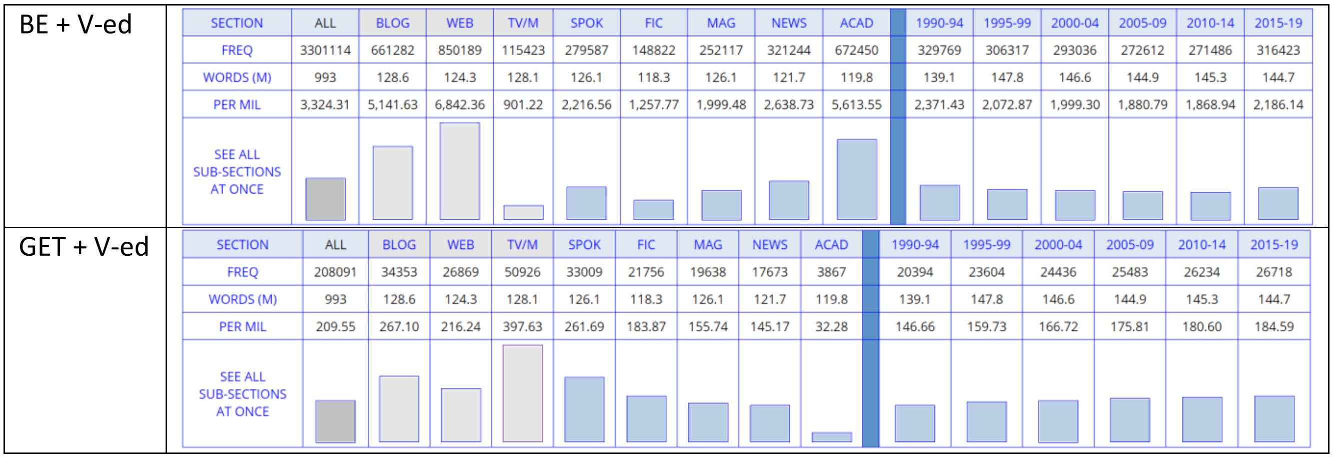
SECTION	ALL	BLOG	WEB	TV/M	SPOK	FIC	MAG	NEWS	ACAD	1990-94	1995-99	2000-04	2005-09	2010-14	2015-19
FREQ	7270	329	263	2257	3156	126	699	394	46	140	393	639	1145	1780	2581
WORDS (M)	993	128.6	124.3	128.1	126.1	118.3	126.1	121.7	119.8	139.1	147.8	146.6	144.9	145.3	144.7
PER MIL	7.32	2.56	2.12	17.62	25.02	1.06	5.54	3.24	0.38	1.01	2.66	4.36	7.90	12.25	17.83
SEE ALL SUB-SECTIONS AT ONCE															



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

1	2018	SPOK	CBS_Morning	A	B	C	you get into the game, you want to play some more **and they 're like ,** well, if you buy this then you play more, you get a
2	2002	MOV	An Evening with Kevin Smith	A	B	C	vault. " I was like, " For what? " **And she 's like ,** " I don't know. " I was like, " Is it
3	2016	SPOK	ABC: The View	A	B	C	because if your parents show body confidence, if you have that **and you 're like ,** and you're like, this is what I have, this is what
4	2014	SPOK	CNN: CNN Live Event	A	B	C	" Why couldn't you be normal and just be gay. **And I was like ,** " Mom, who said that? " UNIDENTIFIED-FEMAL# I need a strong man
5	2012	MOV	Sleepwalk with Me	A	B	C	, I should close strong. What Spanish do I know? **And I 'm like ,** " I know. I'll say, Long live the Immigrant. "
6	1993	SPOK	PBS_Newshour	A	B	C	about the trees, and I'm going to show you. **And I was like ,** hey, you don't have to show me nothin', but what
7	1993	MOV	...ve! The Valentine's Day Massacre	A	B	C	passport is gone. Yeah. This bird came in and... **And I was like ,** " Huh? " You don't want to meet my family. I
8	2001	SPOK	ABC_GMA	A	B	C	training command, and my training command took care of it, **and it was like ,**' We're not going to have this,' and it stopped.
9	2018	SPOK	ABC_20/20	A	B	C	got one video that's coming up on a million views. **And it 's like ,** wow, you know, they all want to hear what I have to
10	2018	SPOK	NPR_AskMe	A	B	C	West Florissant, and these four officers come up to me. **And they 're like ,** hey, you can't stand there. I was like, I just
11	2019	SPOK	NPR_ATCW	A	B	C	, and I was - we were searching for a title. **And I was like ,** well, how about " Room 41? " I mean, that's
12	2002	MOV	An Evening with Kevin Smith	A	B	C	was like, " I'm here to interview you. " **And I was like ,** " Get out of here. You? " I couldn't not talk
13	2011	SPOK	CBS_48Hours	A	B	C	that one of his clients had put up for adoption. BRUCE-LISKER: **And she was like ,** what? Was sort of thrown, but came to just love it.
14	2008	SPOK	NBC_Dateline	A	B	C	I was just talking to my dad about it one day. **And I was like ,** Dad, I don't -- I don't get this. Why me

Or consider the frequency of the "BE passive" (*he was hired; it was paid*) or the "GET passive" (*he got hired; it got paid*) in COCA. The BE passive is more frequent in formal genres (which disproves the idea that the passive occurs mainly in "sloppy" speech) and it is slightly decreasing over time, while the GET passive occurs more in informal genres and is increasing over time. So if someone is writing an academic paper in English, it would sound much better to use the BE passive than the GET passive, which is too informal.



BE + V-ed

SECTION	ALL	BLOG	WEB	TV/M	SPOK	FIC	MAG	NEWS	ACAD	1990-94	1995-99	2000-04	2005-09	2010-14	2015-19
FREQ	3301114	661282	850189	115423	279587	148822	252117	321244	672450	329769	306317	293036	272612	271486	316423
WORDS (M)	993	128.6	124.3	128.1	126.1	118.3	126.1	121.7	119.8	139.1	147.8	146.6	144.9	145.3	144.7
PER MIL	3,324.31	5,141.63	6,842.36	901.22	2,216.56	1,257.77	1,999.48	2,638.73	5,613.55	2,371.43	2,072.87	1,999.30	1,880.79	1,868.94	2,186.14
SEE ALL SUB-SECTIONS AT ONCE															

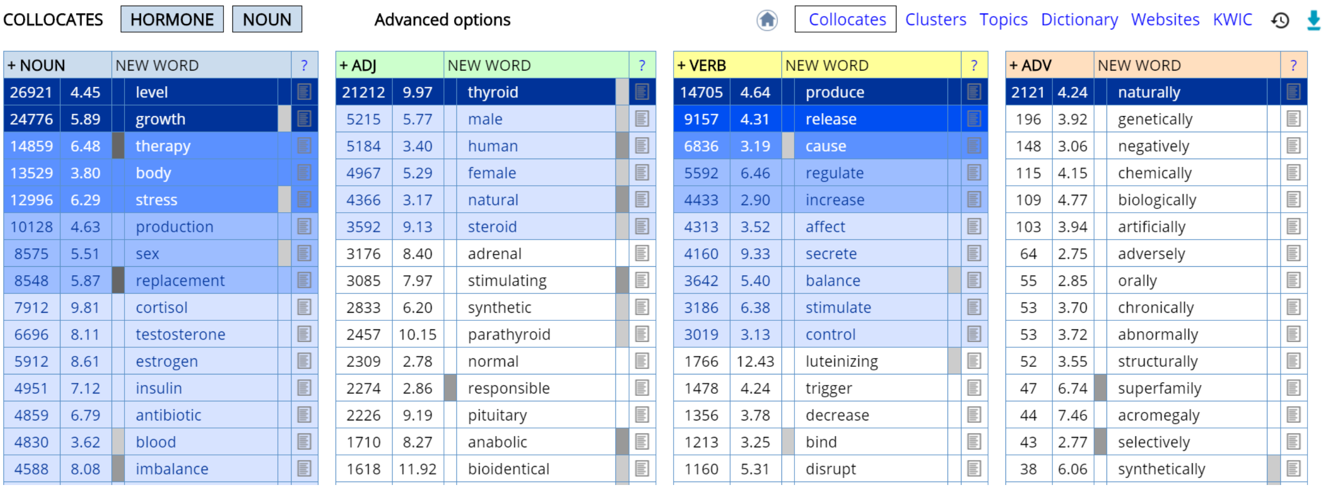
GET + V-ed

SECTION	ALL	BLOG	WEB	TV/M	SPOK	FIC	MAG	NEWS	ACAD	1990-94	1995-99	2000-04	2005-09	2010-14	2015-19
FREQ	208091	34353	26869	50926	33009	21756	19638	17673	3867	20394	23604	24436	25483	26234	26718
WORDS (M)	993	128.6	124.3	128.1	126.1	118.3	126.1	121.7	119.8	139.1	147.8	146.6	144.9	145.3	144.7
PER MIL	209.55	267.10	216.24	397.63	261.69	183.87	155.74	145.17	32.28	146.66	159.73	166.72	175.81	180.60	184.59
SEE ALL SUB-SECTIONS AT ONCE															

Because COCA is the only corpus of English that 1) has texts from a wide range of genres, 2) is large, and 3) is recent, it has been used as the basis for hundreds of in-depth studies of such syntactic variation in English.

Semantics (meaning and usage) (go to beginning)

Collocates (nearby words) can provide extremely useful insight into the **meaning and usage** of a word or phrase, following the idea that "you can tell a lot about a word by the words that it hangs out with". In iWeb (composed of 14 billion words from the Web) and COCA (one billion words, genre-balanced), users can see the frequency of collocates by part of speech (with indications about whether the collocates tend to occur before or after the word in question, and how "tightly bound" together the two words are). For example, these are the collocates of *hormone* in iWeb (via WORD search, and then COLLOCATES):



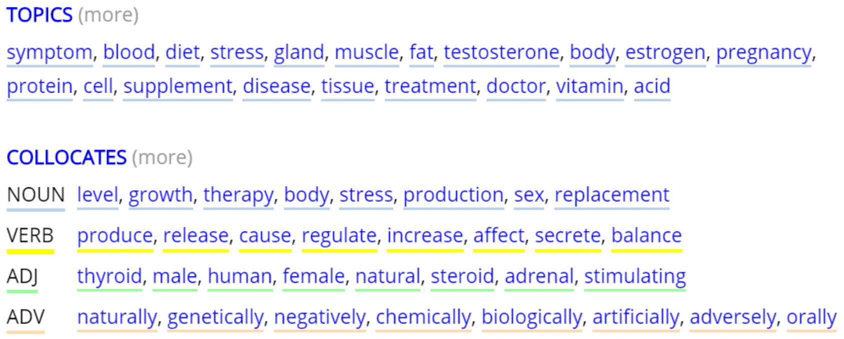
+ NOUN		NEW WORD
26921	4.45	level
24776	5.89	growth
14859	6.48	therapy
13529	3.80	body
12996	6.29	stress
10128	4.63	production
8575	5.51	sex
8548	5.87	replacement
7912	9.81	cortisol
6696	8.11	testosterone
5912	8.61	estrogen
4951	7.12	insulin
4859	6.79	antibiotic
4830	3.62	blood
4588	8.08	imbalance

+ ADJ		NEW WORD
21212	9.97	thyroid
5215	5.77	male
5184	3.40	human
4967	5.29	female
4366	3.17	natural
3592	9.13	steroid
3176	8.40	adrenal
3085	7.97	stimulating
2833	6.20	synthetic
2457	10.15	parathyroid
2309	2.78	normal
2274	2.86	responsible
2226	9.19	pituitary
1710	8.27	anabolic
1618	11.92	bioidentical

+ VERB		NEW WORD
14705	4.64	produce
9157	4.31	release
6836	3.19	cause
5592	6.46	regulate
4433	2.90	increase
4313	3.52	affect
4160	9.33	secrete
3642	5.40	balance
3186	6.38	stimulate
3019	3.13	control
1766	12.43	luteinizing
1478	4.24	trigger
1356	3.78	decrease
1213	3.25	bind
1160	5.31	disrupt

+ ADV		NEW WORD
2121	4.24	naturally
196	3.92	genetically
148	3.06	negatively
115	4.15	chemically
109	4.77	biologically
103	3.94	artificially
64	2.75	adversely
55	2.85	orally
53	3.70	chronically
53	3.72	abnormally
52	3.55	structurally
47	6.74	superfamily
44	7.46	acromegaly
43	2.77	selectively
38	6.06	synthetically

Collocates typically look at "nearby" words (e.g. 4 words left to 4 words right). Topics (which are unique to English-Corpora.org) look at words that co-occur *anywhere* in the text. In many cases, **topics provide even better insight** into the meaning and usage of a word (once again, *hormone* in iWeb):



TOPICS (more)

symptom, blood, diet, stress, gland, muscle, fat, testosterone, body, estrogen, pregnancy, protein, cell, supplement, disease, tissue, treatment, doctor, vitamin, acid

COLLOCATES (more)

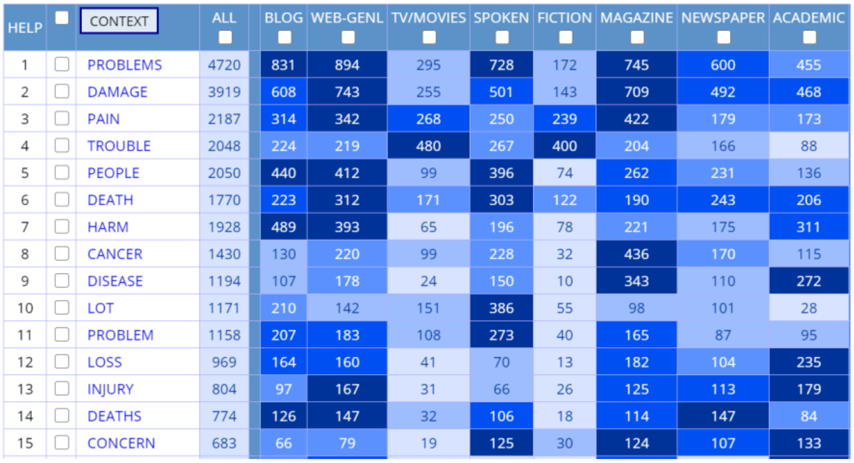
NOUN level, growth, therapy, body, stress, production, sex, replacement

VERB produce, release, cause, regulate, increase, affect, secrete, balance

ADJ thyroid, male, human, female, natural, steroid, adrenal, stimulating

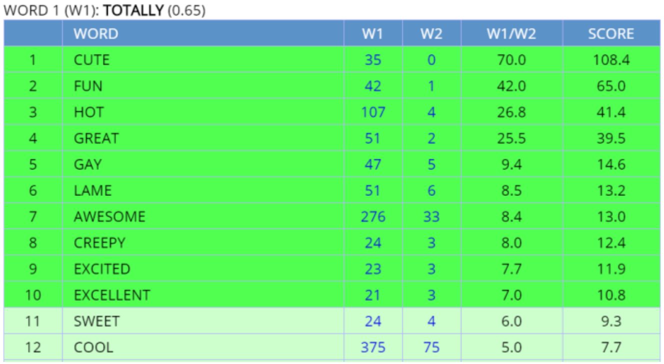
ADV naturally, genetically, negatively, chemically, biologically, artificially, adversely, orally

Collocates sometimes show that a word has different "semantic prosody" than what might first be expected, where "semantic prosody" refers to the preference of certain words for negative or positive collocates. For example, notice how negative the noun collocates of *cause* (as a verb) are in COCA:



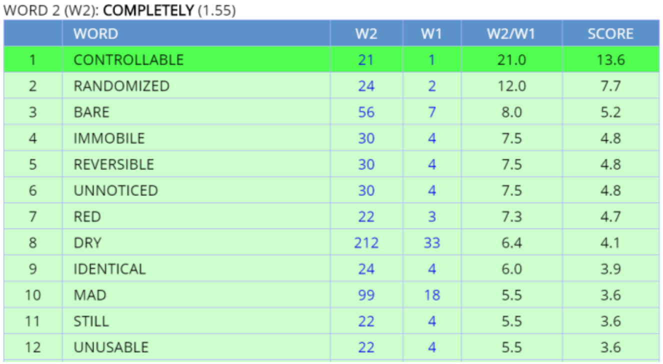
HELP	☐	CONTEXT	ALL	BLOG	WEB-GENL	TV/MOVIES	SPOKEN	FICTION	MAGAZINE	NEWSPAPER	ACADEMIC
			☐	☐	☐	☐	☐	☐	☐	☐	☐
1	☐	PROBLEMS	4720	831	894	295	728	172	745	600	455
2	☐	DAMAGE	3919	608	743	255	501	143	709	492	468
3	☐	PAIN	2187	314	342	268	250	239	422	179	173
4	☐	TROUBLE	2048	224	219	480	267	400	204	166	88
5	☐	PEOPLE	2050	440	412	99	396	74	262	231	136
6	☐	DEATH	1770	223	312	171	303	122	190	243	206
7	☐	HARM	1928	489	393	65	196	78	221	175	311
8	☐	CANCER	1430	130	220	99	228	32	436	170	115
9	☐	DISEASE	1194	107	178	24	150	10	343	110	272
10	☐	LOT	1171	210	142	151	386	55	98	101	28
11	☐	PROBLEM	1158	207	183	108	273	40	165	87	95
12	☐	LOSS	969	164	160	41	70	13	182	104	235
13	☐	INJURY	804	97	167	31	66	26	125	113	179
14	☐	DEATHS	774	126	147	32	106	18	114	147	84
15	☐	CONCERN	683	66	79	19	125	30	124	107	133

Collocates can also be used to investigate the difference between **words with similar meaning**, such as *totally vs completely* (+ADJ); note how much more informal the collocates of *totally* are (left).



WORD 1 (W1): TOTALLY (0.65)

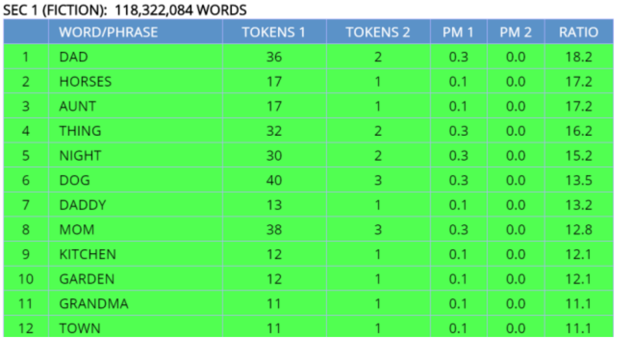
	WORD	W1	W2	W1/W2	SCORE
1	CUTE	35	0	70.0	108.4
2	FUN	42	1	42.0	65.0
3	HOT	107	4	26.8	41.4
4	GREAT	51	2	25.5	39.5
5	GAY	47	5	9.4	14.6
6	LAME	51	6	8.5	13.2
7	AWESOME	276	33	8.4	13.0
8	CREEPY	24	3	8.0	12.4
9	EXCITED	23	3	7.7	11.9
10	EXCELLENT	21	3	7.0	10.8
11	SWEET	24	4	6.0	9.3
12	COOL	375	75	5.0	7.7



WORD 2 (W2): COMPLETELY (1.55)

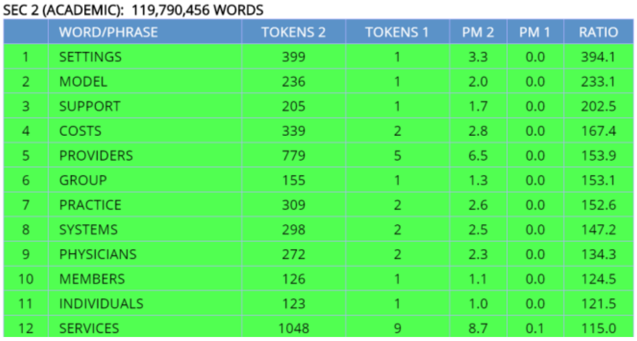
	WORD	W2	W1	W2/W1	SCORE
1	CONTROLLABLE	21	1	21.0	13.6
2	RANDOMIZED	24	2	12.0	7.7
3	BARE	56	7	8.0	5.2
4	IMMOBILE	30	4	7.5	4.8
5	REVERSIBLE	30	4	7.5	4.8
6	UNNOTICED	30	4	7.5	4.8
7	RED	22	3	7.3	4.7
8	DRY	212	33	6.4	4.1
9	IDENTICAL	24	4	6.0	3.9
10	MAD	99	18	5.5	3.6
11	STILL	22	4	5.5	3.6
12	UNUSABLE	22	4	5.5	3.6

Word meaning and usage can **vary by genre** as well. For example, consider the collocates of *care* in fiction (left; focus on what individuals *take care of*) and academic (right; more focus on institutions that provide *care*):



SEC 1 (FICTION): 118,322,084 WORDS

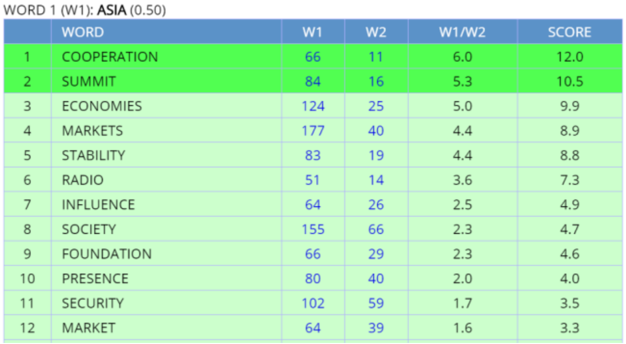
	WORD/PHRASE	TOKENS 1	TOKENS 2	PM 1	PM 2	RATIO
1	DAD	36	2	0.3	0.0	18.2
2	HORSES	17	1	0.1	0.0	17.2
3	AUNT	17	1	0.1	0.0	17.2
4	THING	32	2	0.3	0.0	16.2
5	NIGHT	30	2	0.3	0.0	15.2
6	DOG	40	3	0.3	0.0	13.5
7	DADDY	13	1	0.1	0.0	13.2
8	MOM	38	3	0.3	0.0	12.8
9	KITCHEN	12	1	0.1	0.0	12.1
10	GARDEN	12	1	0.1	0.0	12.1
11	GRANDMA	11	1	0.1	0.0	11.1
12	TOWN	11	1	0.1	0.0	11.1



SEC 2 (ACADEMIC): 119,790,456 WORDS

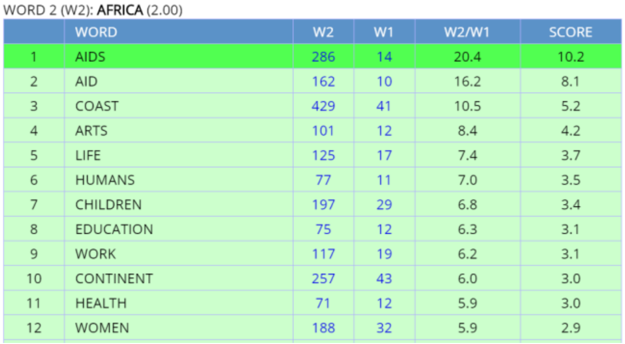
	WORD/PHRASE	TOKENS 2	TOKENS 1	PM 2	PM 1	RATIO
1	SETTINGS	399	1	3.3	0.0	394.1
2	MODEL	236	1	2.0	0.0	233.1
3	SUPPORT	205	1	1.7	0.0	202.5
4	COSTS	339	2	2.8	0.0	167.4
5	PROVIDERS	779	5	6.5	0.0	153.9
6	GROUP	155	1	1.3	0.0	153.1
7	PRACTICE	309	2	2.6	0.0	152.6
8	SYSTEMS	298	2	2.5	0.0	147.2
9	PHYSICIANS	272	2	2.3	0.0	134.3
10	MEMBERS	126	1	1.1	0.0	124.5
11	INDIVIDUALS	123	1	1.0	0.0	121.5
12	SERVICES	1048	9	8.7	0.1	115.0

Collocates can also move beyond strict "word meaning" to show **"what we are saying" about different topics**. For example, consider the collocates of *Asia* (left; perhaps more focus on countries and institutions) and *Africa* (right; perhaps more focus on individuals, health and well-being).



WORD 1 (W1): ASIA (0.50)

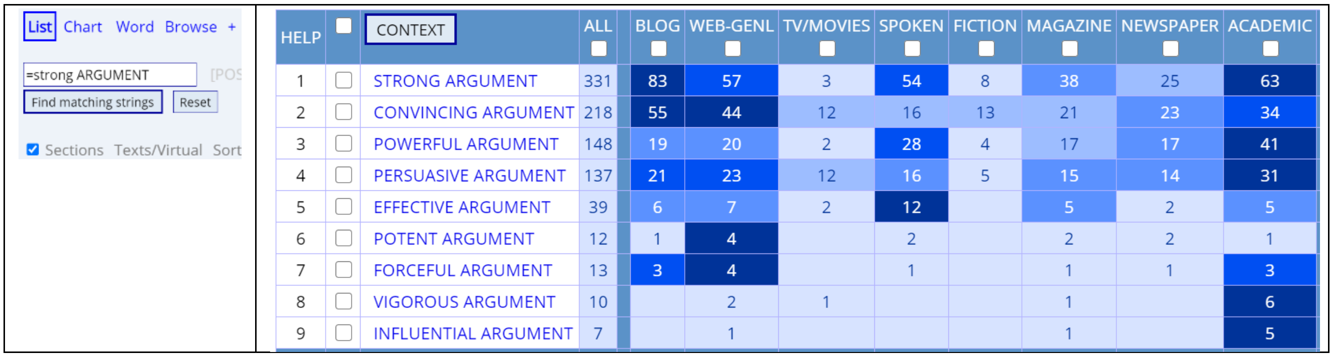
	WORD	W1	W2	W1/W2	SCORE
1	COOPERATION	66	11	6.0	12.0
2	SUMMIT	84	16	5.3	10.5
3	ECONOMIES	124	25	5.0	9.9
4	MARKETS	177	40	4.4	8.9
5	STABILITY	83	19	4.4	8.8
6	RADIO	51	14	3.6	7.3
7	INFLUENCE	64	26	2.5	4.9
8	SOCIETY	155	66	2.3	4.7
9	FOUNDATION	66	29	2.3	4.6
10	PRESENCE	80	40	2.0	4.0
11	SECURITY	102	59	1.7	3.5
12	MARKET	64	39	1.6	3.3



WORD 2 (W2): AFRICA (2.00)

	WORD	W2	W1	W2/W1	SCORE
1	AIDS	286	14	20.4	10.2
2	AID	162	10	16.2	8.1
3	COAST	429	41	10.5	5.2
4	ARTS	101	12	8.4	4.2
5	LIFE	125	17	7.4	3.7
6	HUMANS	77	11	7.0	3.5
7	CHILDREN	197	29	6.8	3.4
8	EDUCATION	75	12	6.3	3.1
9	WORK	117	19	6.2	3.1
10	CONTINENT	257	43	6.0	3.0
11	HEALTH	71	12	5.9	3.0
12	WOMEN	188	32	5.9	2.9

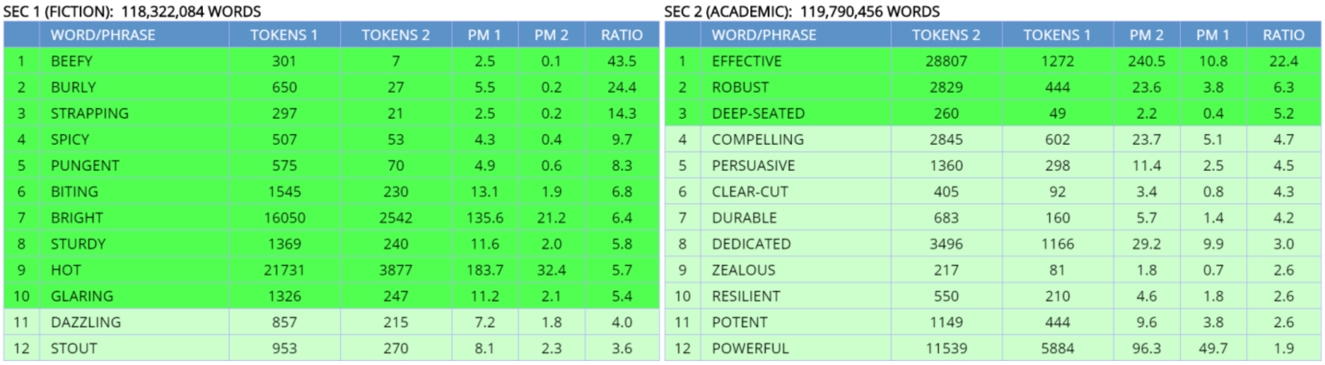
The corpora from English-Corpora.org are the only ones that can be searched by **synonym**, meaning that searches can focus on meaning as well as form (words). This can be extremely **useful for non-native speakers**, allowing them to see which of several "competing" words are actually used in a given context (such as *"strong" argument*) and thus have their writing or speech sound more "native-like".



=strong ARGUMENT

		ALL	BLOG	WEB-GENL	TV/MOVIES	SPOKEN	FICTION	MAGAZINE	NEWSPAPER	ACADEMIC
1	STRONG ARGUMENT	331	83	57	3	54	8	38	25	63
2	CONVINCING ARGUMENT	218	55	44	12	16	13	21	23	34
3	POWERFUL ARGUMENT	148	19	20	2	28	4	17	17	41
4	PERSUASIVE ARGUMENT	137	21	23	12	16	5	15	14	31
5	EFFECTIVE ARGUMENT	39	6	7	2	12		5	2	5
6	POTENT ARGUMENT	12	1	4		2		2	2	1
7	FORCEFUL ARGUMENT	13	3	4		1		1	1	3
8	VIGOROUS ARGUMENT	10		2	1			1		6
9	INFLUENTIAL ARGUMENT	7		1				1		5

Synonyms also vary by genre. For example, consider the synonyms of *strong* in fiction (left) and academic (right). All of these synonyms might appear together in a thesaurus, but only the corpus data shows, for example, that writers might refer to (= “strong”) *beefy, burly, strapping lumberjacks* in fiction, but (= “strong”) *effective, compelling, persuasive arguments* in academic writing.



SEC 1 (FICTION): 118,322,084 WORDS

	WORD/PHRASE	TOKENS 1	TOKENS 2	PM 1	PM 2	RATIO
1	BEEFY	301	7	2.5	0.1	43.5
2	BURLY	650	27	5.5	0.2	24.4
3	STRAPPING	297	21	2.5	0.2	14.3
4	SPICY	507	53	4.3	0.4	9.7
5	PUNGENT	575	70	4.9	0.6	8.3
6	BITING	1545	230	13.1	1.9	6.8
7	BRIGHT	16050	2542	135.6	21.2	6.4
8	STURDY	1369	240	11.6	2.0	5.8
9	HOT	21731	3877	183.7	32.4	5.7
10	GLARING	1326	247	11.2	2.1	5.4
11	DAZZLING	857	215	7.2	1.8	4.0
12	STOUT	953	270	8.1	2.3	3.6

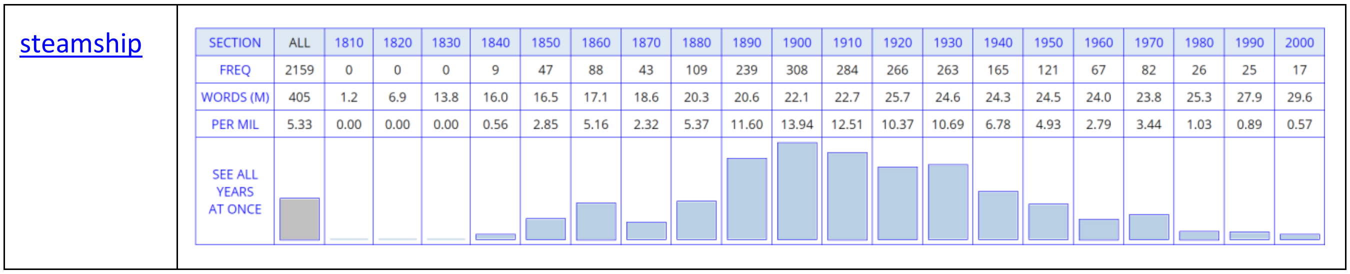
SEC 2 (ACADEMIC): 119,790,456 WORDS

	WORD/PHRASE	TOKENS 2	TOKENS 1	PM 2	PM 1	RATIO
1	EFFECTIVE	28807	1272	240.5	10.8	22.4
2	ROBUST	2829	444	23.6	3.8	6.3
3	DEEP-SEATED	260	49	2.2	0.4	5.2
4	COMPELLING	2845	602	23.7	5.1	4.7
5	PERSUASIVE	1360	298	11.4	2.5	4.5
6	CLEAR-CUT	405	92	3.4	0.8	4.3
7	DURABLE	683	160	5.7	1.4	4.2
8	DEDICATED	3496	1166	29.2	9.9	3.0
9	ZEALOUS	217	81	1.8	0.7	2.6
10	RESILIENT	550	210	4.6	1.8	2.6
11	POTENT	1149	444	9.6	3.8	2.6
12	POWERFUL	11539	5884	96.3	49.7	1.9

Historical change (go to beginning)

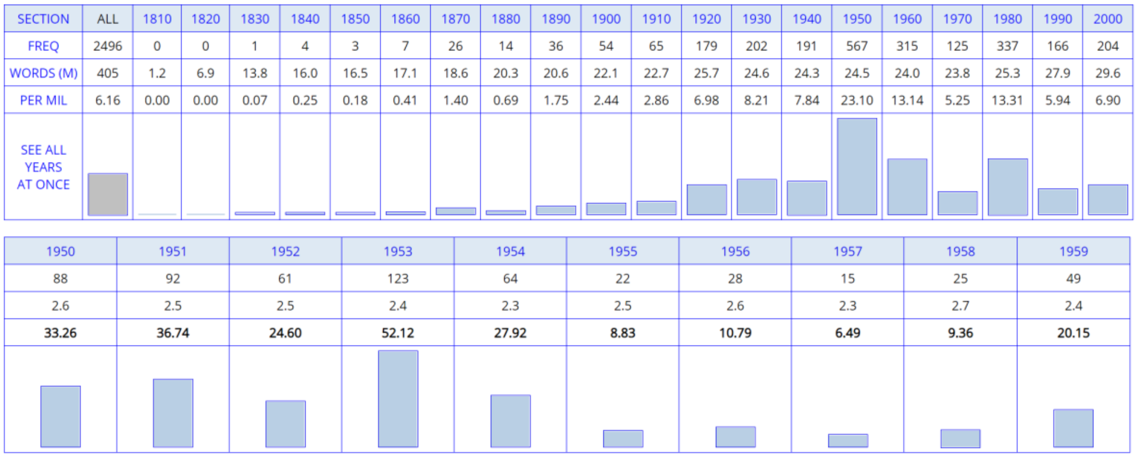
There are **many corpora from English-Corpora.org that provide very useful data on language change**, whether it is the 1400s-1600s (EEBO), 1810-2009 (COHA), 1800-2018 (US Supreme Court), 1803-2003 (Hansard; British Parliament), or 1926-2006 (TIME Magazine). The Movie Corpus (1930s-2010s) and the TV Corpus (1950s-2010s) are the only large corpora that provide a large amount of data on changes in very informal speech. And researchers can also focus on much more recent language change, as in COCA (1990-2019), the NOW Corpus (2010-2020) and the Coronavirus Corpus (2020). The last two corpora are updated *every night* with millions of words of data. Overall, there are billions of words of data, and most of these corpora are 50-100x as large as comparable historical corpora, which allows researchers to look at a **much wider range of phenomena**. In addition, these corpora allow a much wider range of searches than the simple searches for exact words and phrases in **Google Books n-grams**.

At the most basic level, researchers can see the **frequency of words and phrases by decade**. For example, the following charts from COHA (400 million words, 1810-2009) shows *steamship* by decade, and *Reds* by decade and even by year (note 1953, the year of the McCarthy hearings in the US Senate). As the search for *a most ADJ NOUN* shows, researchers can also look for phrases, including part of speech.



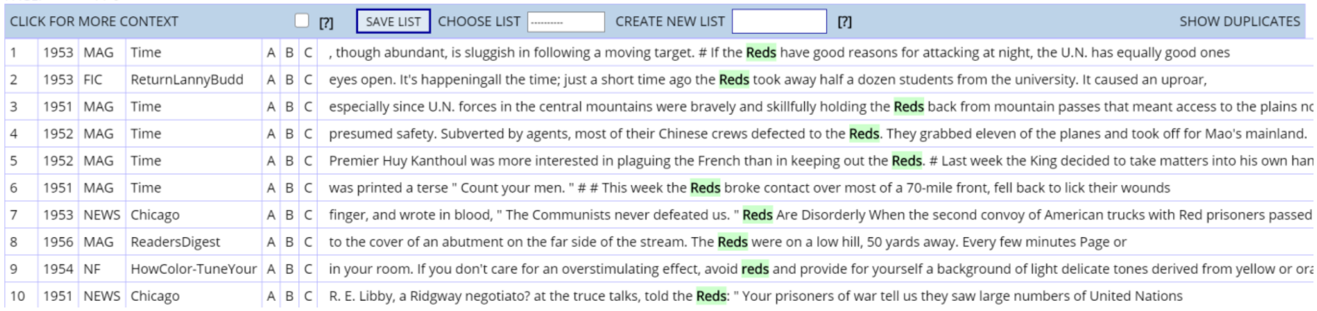
SECTION	ALL	1810	1820	1830	1840	1850	1860	1870	1880	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000
FREQ	2159	0	0	0	9	47	88	43	109	239	308	284	266	263	165	121	67	82	26	25	17
WORDS (M)	405	1.2	6.9	13.8	16.0	16.5	17.1	18.6	20.3	20.6	22.1	22.7	25.7	24.6	24.3	24.5	24.0	23.8	25.3	27.9	29.6
PER MIL	5.33	0.00	0.00	0.00	0.56	2.85	5.16	2.32	5.37	11.60	13.94	12.51	10.37	10.69	6.78	4.93	2.79	3.44	1.03	0.89	0.57

[Reds]



SECTION	ALL	1810	1820	1830	1840	1850	1860	1870	1880	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000
FREQ	2496	0	0	1	4	3	7	26	14	36	54	65	179	202	191	567	315	125	337	166	204
WORDS (M)	405	1.2	6.9	13.8	16.0	16.5	17.1	18.6	20.3	20.6	22.1	22.7	25.7	24.6	24.3	24.5	24.0	23.8	25.3	27.9	29.6
PER MIL	6.16	0.00	0.00	0.07	0.25	0.18	0.41	1.40	0.69	1.75	2.44	2.86	6.98	8.21	7.84	23.10	13.14	5.25	13.31	5.94	6.90
SEE ALL YEARS AT ONCE																					

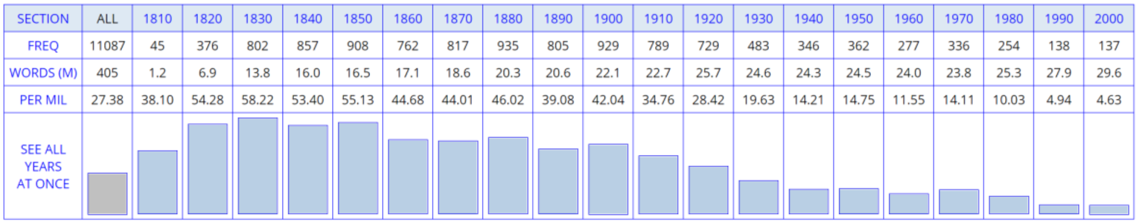
1950	1951	1952	1953	1954	1955	1956	1957	1958	1959
88	92	61	123	64	22	28	15	25	49
2.6	2.5	2.5	2.4	2.3	2.5	2.6	2.3	2.7	2.4
33.26	**36.74**	**24.60**	**52.12**	**27.92**	**8.83**	**10.79**	**6.49**	**9.36**	**20.15**



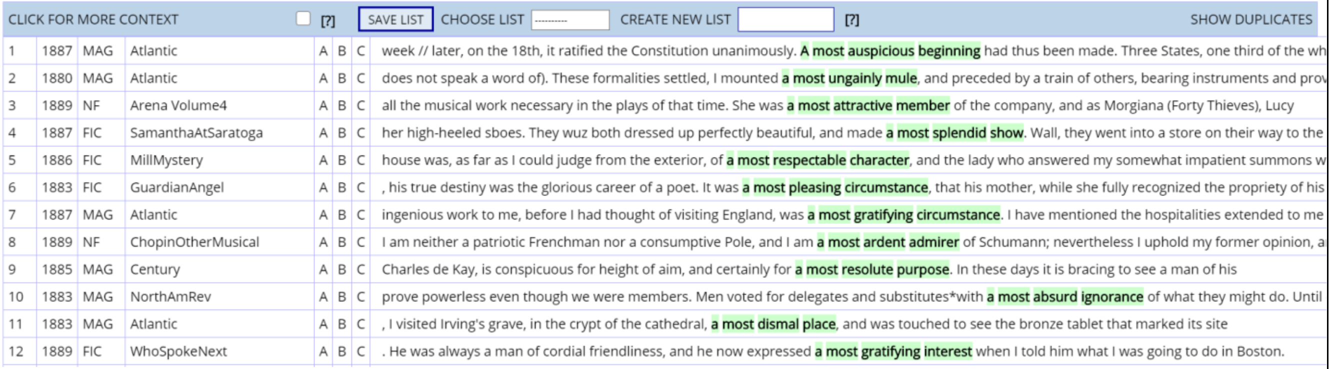
CLICK FOR MORE CONTEXT □ [?] SAVE LIST CHOOSE LIST --------- CREATE NEW LIST [?] SHOW DUPLICATES

1	1953	MAG	Time	A B C	, though abundant, is sluggish in following a moving target. # If the **Reds** have good reasons for attacking at night, the U.N. has equally good ones
2	1953	FIC	ReturnLannyBudd	A B C	eyes open. It's happeningall the time; just a short time ago the **Reds** took away half a dozen students from the university. It caused an uproar,
3	1951	MAG	Time	A B C	especially since U.N. forces in the central mountains were bravely and skillfully holding the **Reds** back from mountain passes that meant access to the plains nc
4	1952	MAG	Time	A B C	presumed safety. Subverted by agents, most of their Chinese crews defected to the **Reds**. They grabbed eleven of the planes and took off for Mao's mainland.
5	1952	MAG	Time	A B C	Premier Huy Kanthoul was more interested in plaguing the French than in keeping out the **Reds**. # Last week the King decided to take matters into his own han
6	1951	MAG	Time	A B C	was printed a terse " Count your men. " # # This week the **Reds** broke contact over most of a 70-mile front, fell back to lick their wounds
7	1953	NEWS	Chicago	A B C	finger, and wrote in blood, " The Communists never defeated us. " **Reds** Are Disorderly When the second convoy of American trucks with Red prisoners passed
8	1956	MAG	ReadersDigest	A B C	to the cover of an abutment on the far side of the stream. The **Reds** were on a low hill, 50 yards away. Every few minutes Page or
9	1954	NF	HowColor-TuneYour	A B C	in your room. If you don't care for an overstimulating effect, avoid **reds** and provide for yourself a background of light delicate tones derived from yellow or ora
10	1951	NEWS	Chicago	A B C	R. E. Libby, a Ridgway negotiato? at the truce talks, told the **Reds**: " Your prisoners of war tell us they saw large numbers of United Nations

[a most ADJ NOUN]



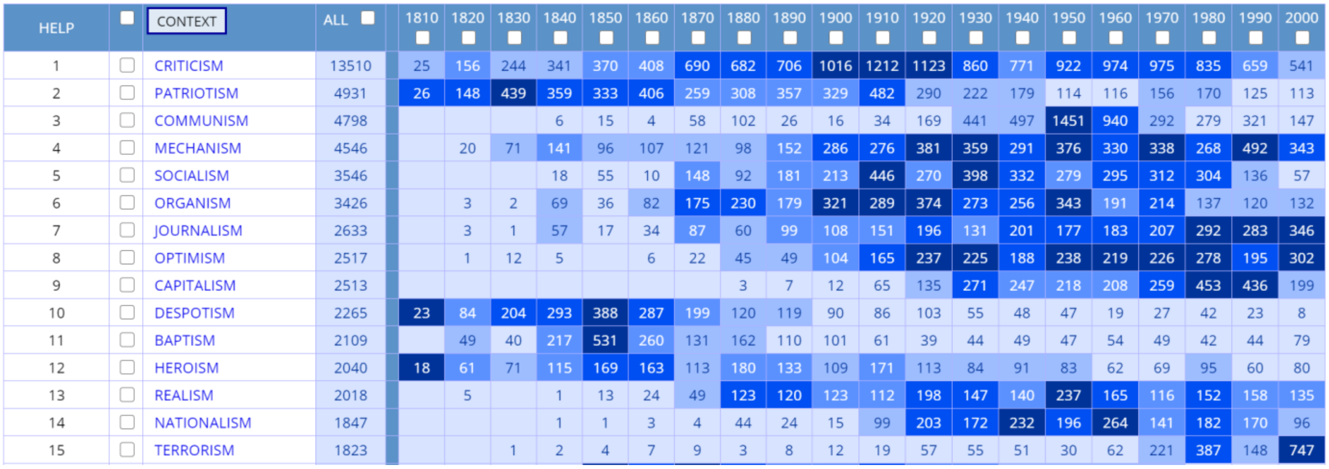
SECTION	ALL	1810	1820	1830	1840	1850	1860	1870	1880	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000
FREQ	11087	45	376	802	857	908	762	817	935	805	929	789	729	483	346	362	277	336	254	138	137
WORDS (M)	405	1.2	6.9	13.8	16.0	16.5	17.1	18.6	20.3	20.6	22.1	22.7	25.7	24.6	24.3	24.5	24.0	23.8	25.3	27.9	29.6
PER MIL	27.38	38.10	54.28	58.22	53.40	55.13	44.68	44.01	46.02	39.08	42.04	34.76	28.42	19.63	14.21	14.75	11.55	14.11	10.03	4.94	4.63
SEE ALL YEARS AT ONCE																					



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

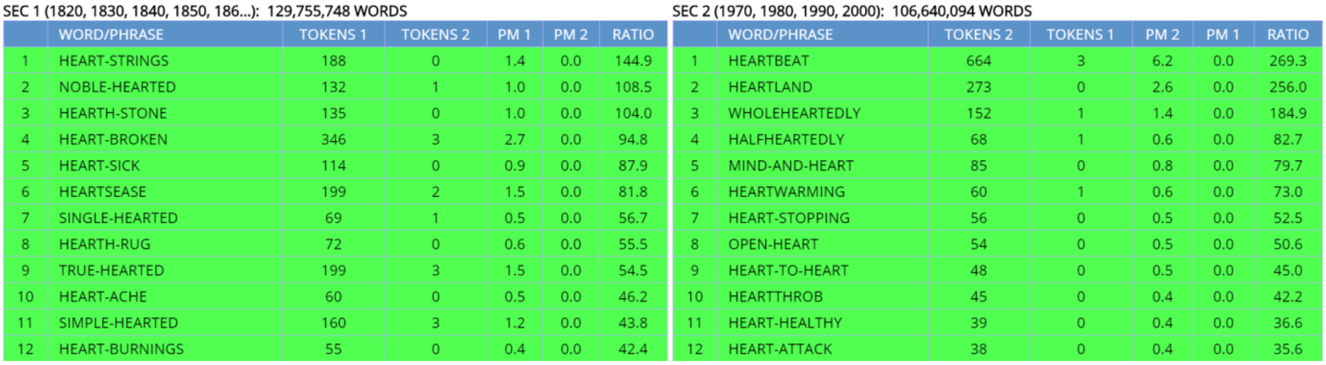
1	1887	MAG	Atlantic	A B C	week // later, on the 18th, it ratified the Constitution unanimously. **A most auspicious beginning** had thus been made. Three States, one third of the wh
2	1880	MAG	Atlantic	A B C	does not speak a word of). These formalities settled, I mounted **a most ungainly mule**, and preceded by a train of others, bearing instruments and prov
3	1889	NF	Arena Volume4	A B C	all the musical work necessary in the plays of that time. She was **a most attractive member** of the company, and as Morgiana (Forty Thieves), Lucy
4	1887	FIC	SamanthaAtSaratoga	A B C	her high-heeled sboes. They wuz both dressed up perfectly beautiful, and made **a most splendid show**. Wall, they went into a store on their way to the
5	1886	FIC	MillMystery	A B C	house was, as far as I could judge from the exterior, of **a most respectable character**, and the lady who answered my somewhat impatient summons w
6	1883	FIC	GuardianAngel	A B C	, his true destiny was the glorious career of a poet. It was **a most pleasing circumstance**, that his mother, while she fully recognized the propriety of his
7	1887	MAG	Atlantic	A B C	ingenious work to me, before I had thought of visiting England, was **a most gratifying circumstance**. I have mentioned the hospitalities extended to me
8	1889	NF	ChopinOtherMusical	A B C	I am neither a patriotic Frenchman nor a consumptive Pole, and I am **a most ardent admirer** of Schumann; nevertheless I uphold my former opinion, a
9	1885	MAG	Century	A B C	Charles de Kay, is conspicuous for height of aim, and certainly for **a most resolute purpose**. In these days it is bracing to see a man of his
10	1883	MAG	NorthAmRev	A B C	prove powerless even though we were members. Men voted for delegates and substitutes*with **a most absurd ignorance** of what they might do. Until
11	1883	MAG	Atlantic	A B C	, I visited Irving's grave, in the crypt of the cathedral, **a most dismal place**, and was touched to see the bronze tablet that marked its site
12	1889	FIC	WhoSpokeNext	A B C	. He was always a man of cordial friendliness, and he now expressed **a most gratifying interest** when I told him what I was going to do in Boston.

Researchers can find the **frequency of all matching string**s in all decades, such as *[*ism words](#)* in COHA. Note the higher frequency of *patriotism, despotism*, and *heroism* in the 1800s, *socialism*, *communism,* and *nationalism* in the mid-1900s, and *capitalism* and *terrorism* in the late 1900s and early 2000s.



HELP	□	CONTEXT	ALL □	1810 □	1820 □	1830 □	1840 □	1850 □	1860 □	1870 □	1880 □	1890 □	1900 □	1910 □	1920 □	1930 □	1940 □	1950 □	1960 □	1970 □	1980 □	1990 □	2000 □
1	□	CRITICISM	13510	25	156	244	341	370	408	690	682	706	1016	1212	1123	860	771	922	974	975	835	659	541
2	□	PATRIOTISM	4931	26	148	439	359	333	406	259	308	357	329	482	290	222	179	114	116	156	170	125	113
3	□	COMMUNISM	4798				6	15	4	58	102	26	16	34	169	441	497	1451	940	292	279	321	147
4	□	MECHANISM	4546		20	71	141	96	107	121	98	152	286	276	381	359	291	376	330	338	268	492	343
5	□	SOCIALISM	3546				18	55	10	148	92	181	213	446	270	398	332	279	295	312	304	136	57
6	□	ORGANISM	3426		3	2	69	36	82	175	230	179	321	289	374	273	256	343	191	214	137	120	132
7	□	JOURNALISM	2633		3	1	57	17	34	87	60	99	108	151	196	131	201	177	183	207	292	283	346
8	□	OPTIMISM	2517		1	12	5		6	22	45	49	104	165	237	225	188	238	219	226	278	195	302
9	□	CAPITALISM	2513								3	7	12	65	135	271	247	218	208	259	453	436	199
10	□	DESPOTISM	2265	23	84	204	293	388	287	199	120	119	90	86	103	55	48	47	19	27	42	23	8
11	□	BAPTISM	2109		49	40	217	531	260	131	162	110	101	61	39	44	49	47	54	49	42	44	79
12	□	HEROISM	2040	18	61	71	115	169	163	113	180	133	109	171	113	84	91	83	62	69	95	60	80
13	□	REALISM	2018		5		1	13	24	49	123	120	123	112	198	147	140	237	165	116	152	158	135
14	□	NATIONALISM	1847				1	1	3	4	44	24	15	99	203	172	232	196	264	141	182	170	96
15	□	TERRORISM	1823			1	2	4	7	9	3	8	12	19	57	55	51	30	62	221	387	148	747

It is also possible to find **all words that are more common in one time period** than in another. For example, words with [*heart*](#) in COHA in the 1800s (left) vs the late 1900s (right), or [*ess words](#) in TIME in the 1920s-1930s (left) vs the 1980s-2000s (right); note older feminine forms like *negress, authoress, sculptress, adventuress*, and *poetess*.

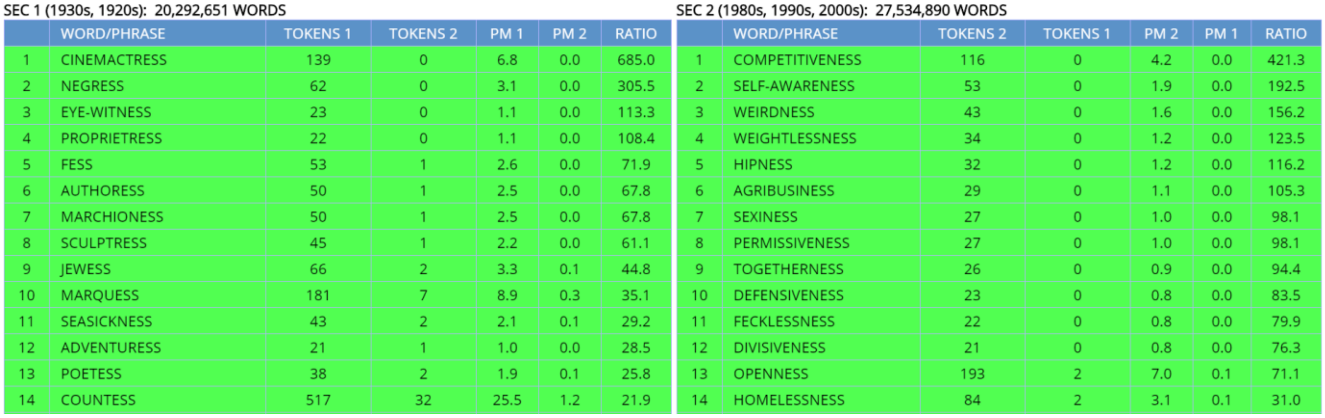


SEC 1 (1820, 1830, 1840, 1850, 186...): 129,755,748 WORDS

	WORD/PHRASE	TOKENS 1	TOKENS 2	PM 1	PM 2	RATIO
1	HEART-STRINGS	188	0	1.4	0.0	144.9
2	NOBLE-HEARTED	132	1	1.0	0.0	108.5
3	HEARTH-STONE	135	0	1.0	0.0	104.0
4	HEART-BROKEN	346	3	2.7	0.0	94.8
5	HEART-SICK	114	0	0.9	0.0	87.9
6	HEARTSEASE	199	2	1.5	0.0	81.8
7	SINGLE-HEARTED	69	1	0.5	0.0	56.7
8	HEARTH-RUG	72	0	0.6	0.0	55.5
9	TRUE-HEARTED	199	3	1.5	0.0	54.5
10	HEART-ACHE	60	0	0.5	0.0	46.2
11	SIMPLE-HEARTED	160	3	1.2	0.0	43.8
12	HEART-BURNINGS	55	0	0.4	0.0	42.4

SEC 2 (1970, 1980, 1990, 2000): 106,640,094 WORDS

	WORD/PHRASE	TOKENS 2	TOKENS 1	PM 2	PM 1	RATIO
1	HEARTBEAT	664	3	6.2	0.0	269.3
2	HEARTLAND	273	0	2.6	0.0	256.0
3	WHOLEHEARTEDLY	152	1	1.4	0.0	184.9
4	HALFHEARTEDLY	68	1	0.6	0.0	82.7
5	MIND-AND-HEART	85	0	0.8	0.0	79.7
6	HEARTWARMING	60	1	0.6	0.0	73.0
7	HEART-STOPPING	56	0	0.5	0.0	52.5
8	OPEN-HEART	54	0	0.5	0.0	50.6
9	HEART-TO-HEART	48	0	0.5	0.0	45.0
10	HEARTTHROB	45	0	0.4	0.0	42.2
11	HEART-HEALTHY	39	0	0.4	0.0	36.6
12	HEART-ATTACK	38	0	0.4	0.0	35.6



SEC 1 (1930s, 1920s): 20,292,651 WORDS

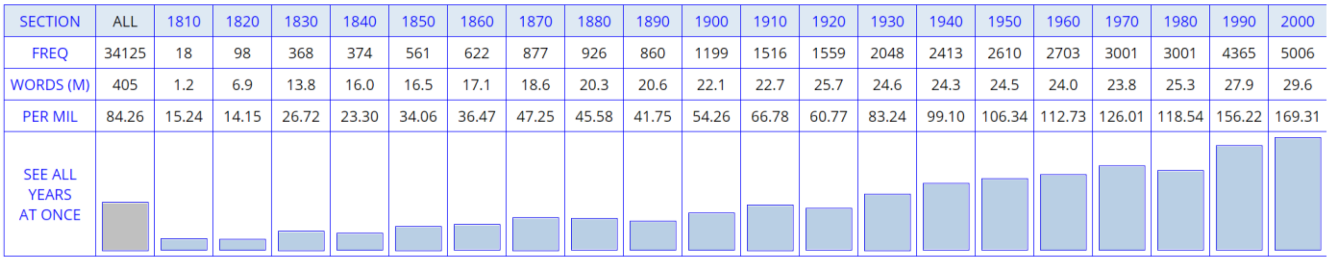
	WORD/PHRASE	TOKENS 1	TOKENS 2	PM 1	PM 2	RATIO
1	CINEMACTRESS	139	0	6.8	0.0	685.0
2	NEGRESS	62	0	3.1	0.0	305.5
3	EYE-WITNESS	23	0	1.1	0.0	113.3
4	PROPRIETRESS	22	0	1.1	0.0	108.4
5	FESS	53	1	2.6	0.0	71.9
6	AUTHORESS	50	1	2.5	0.0	67.8
7	MARCHIONESS	50	1	2.5	0.0	67.8
8	SCULPTRESS	45	1	2.2	0.0	61.1
9	JEWESS	66	2	3.3	0.1	44.8
10	MARQUESS	181	7	8.9	0.3	35.1
11	SEASICKNESS	43	2	2.1	0.1	29.2
12	ADVENTURESS	21	1	1.0	0.0	28.5
13	POETESS	38	2	1.9	0.1	25.8
14	COUNTESS	517	32	25.5	1.2	21.9

SEC 2 (1980s, 1990s, 2000s): 27,534,890 WORDS

	WORD/PHRASE	TOKENS 2	TOKENS 1	PM 2	PM 1	RATIO
1	COMPETITIVENESS	116	0	4.2	0.0	421.3
2	SELF-AWARENESS	53	0	1.9	0.0	192.5
3	WEIRDNESS	43	0	1.6	0.0	156.2
4	WEIGHTLESSNESS	34	0	1.2	0.0	123.5
5	HIPNESS	32	0	1.2	0.0	116.2
6	AGRIBUSINESS	29	0	1.1	0.0	105.3
7	SEXINESS	27	0	1.0	0.0	98.1
8	PERMISSIVENESS	27	0	1.0	0.0	98.1
9	TOGETHERNESS	26	0	0.9	0.0	94.4
10	DEFENSIVENESS	23	0	0.8	0.0	83.5
11	FECKLESSNESS	22	0	0.8	0.0	79.9
12	DIVISIVENESS	21	0	0.8	0.0	76.3
13	OPENNESS	193	2	7.0	0.1	71.1
14	HOMELESSNESS	84	2	3.1	0.1	31.0

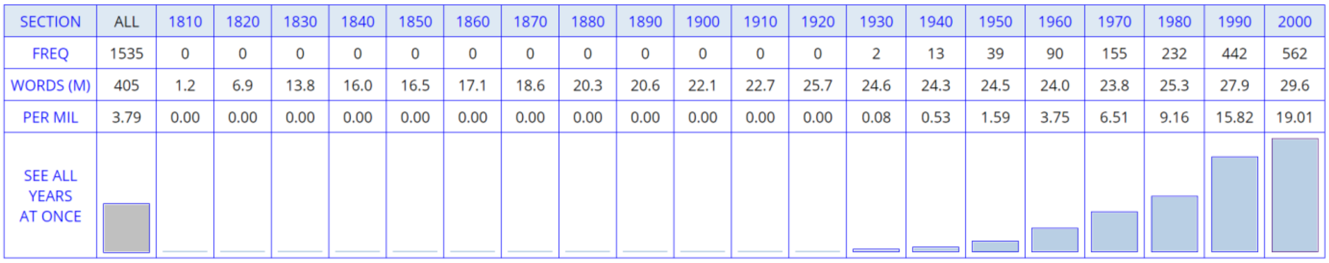
The corpora can also be used to investigate **grammatical change over time**, and they have been used for a wide range of studies during the last ten years (since COHA was released in 2010). For example, see the frequency of GET + V-ed (e.g. *get married, got painted*) in COHA during the last 200 years, or the frequency of END up V-ing (e.g. *ended up paying too much*); note how the construction only really began to be used about 100 years ago.

GET V-ed



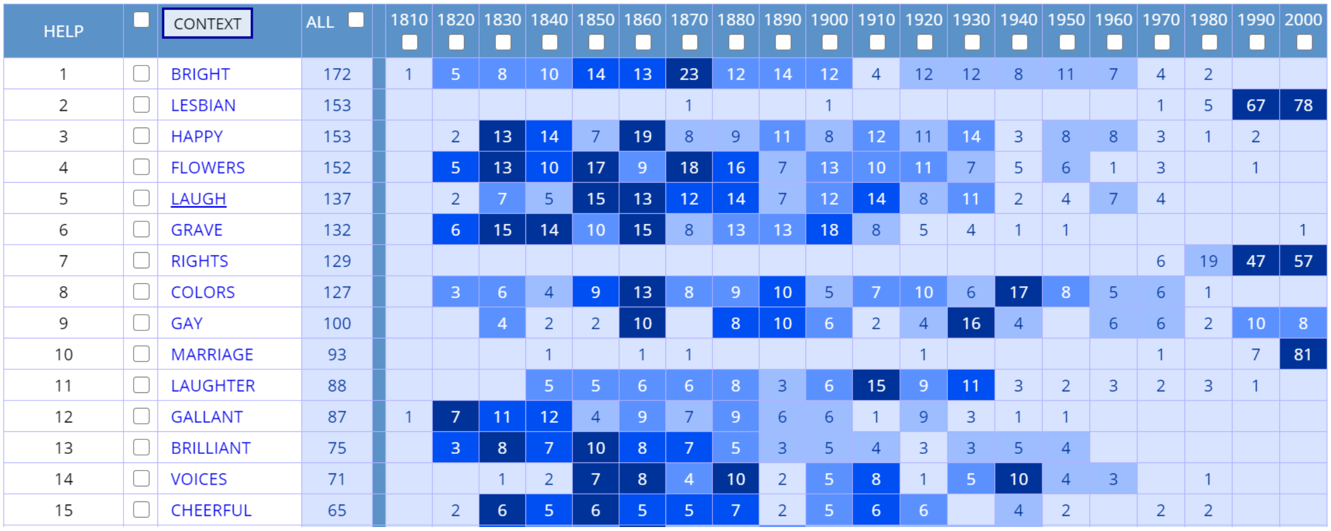
SECTION	ALL	1810	1820	1830	1840	1850	1860	1870	1880	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000
FREQ	34125	18	98	368	374	561	622	877	926	860	1199	1516	1559	2048	2413	2610	2703	3001	3001	4365	5006
WORDS (M)	405	1.2	6.9	13.8	16.0	16.5	17.1	18.6	20.3	20.6	22.1	22.7	25.7	24.6	24.3	24.5	24.0	23.8	25.3	27.9	29.6
PER MIL	84.26	15.24	14.15	26.72	23.30	34.06	36.47	47.25	45.58	41.75	54.26	66.78	60.77	83.24	99.10	106.34	112.73	126.01	118.54	156.22	169.31
SEE ALL YEARS AT ONCE																					

END up V-ing



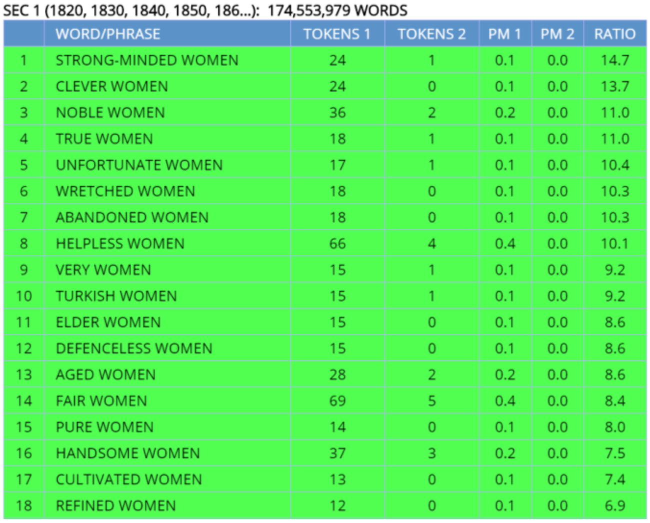
SECTION	ALL	1810	1820	1830	1840	1850	1860	1870	1880	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000
FREQ	1535	0	0	0	0	0	0	0	0	0	0	0	0	2	13	39	90	155	232	442	562
WORDS (M)	405	1.2	6.9	13.8	16.0	16.5	17.1	18.6	20.3	20.6	22.1	22.7	25.7	24.6	24.3	24.5	24.0	23.8	25.3	27.9	29.6
PER MIL	3.79	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.08	0.53	1.59	3.75	6.51	9.16	15.82	19.01
SEE ALL YEARS AT ONCE																					

Researchers can also investigate **changes in meaning using collocates**, with the idea that changes in nearby words can signal changes in meaning or usage. These are the collocates of *gay* decade by decade during the last 200 years. Notice the change from "happy, joyful" in the 1800s to "sexual orientation" in the second half of the 1900s.



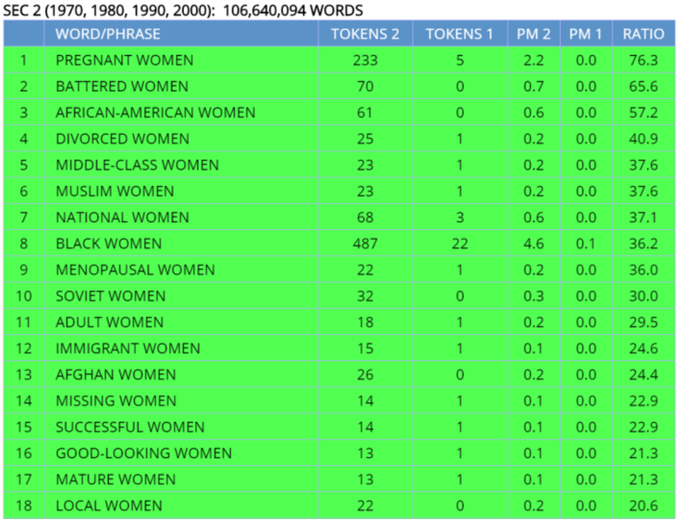
HELP	☐	CONTEXT	ALL ☐	1810 ☐	1820 ☐	1830 ☐	1840 ☐	1850 ☐	1860 ☐	1870 ☐	1880 ☐	1890 ☐	1900 ☐	1910 ☐	1920 ☐	1930 ☐	1940 ☐	1950 ☐	1960 ☐	1970 ☐	1980 ☐	1990 ☐	2000 ☐
1	☐	BRIGHT	172	1	5	8	10	14	13	23	12	14	12	4	12	12	8	11	7	4	2		
2	☐	LESBIAN	153							1			1							1	5	67	78
3	☐	HAPPY	153		2	13	14	7	19	8	9	11	8	12	11	14	3	8	8	3	1	2	
4	☐	FLOWERS	152		5	13	10	17	9	18	16	7	13	10	11	7	5	6	1	3		1	
5	☐	LAUGH	137		2	7	5	15	13	12	14	7	12	14	8	11	2	4	7	4			
6	☐	GRAVE	132		6	15	14	10	15	8	13	13	18	8	5	4	1	1					1
7	☐	RIGHTS	129																	6	19	47	57
8	☐	COLORS	127		3	6	4	9	13	8	9	10	5	7	10	6	17	8	5	6	1		
9	☐	GAY	100			4	2	2	10		8	10	6	2	4	16	4		6	6	2	10	8
10	☐	MARRIAGE	93				1		1	1					1					1		7	81
11	☐	LAUGHTER	88				5	5	6	6	8	3	6	15	9	11	3	2	3	2	3	1	
12	☐	GALLANT	87	1	7	11	12	4	9	7	9	6	6	1	9	3	1	1					
13	☐	BRILLIANT	75		3	8	7	10	8	7	5	3	5	4	3	3	5	4					
14	☐	VOICES	71			1	2	7	8	4	10	2	5	8	1	5	10	4	3		1		
15	☐	CHEERFUL	65		2	6	5	6	5	5	7	2	5	6	6		4	2		2	2		

Collocates can also signal **changes in "what we are saying" about certain topics**. For example, the collocates of *women* from texts in the 1800s (left) show a very sexist worldview, in which women were evaluated according to their moral characteristics (*noble, true, pure, cultivated, refined, wretched*); they were often seen as being weak (*unfortunate, abandoned, helpless*); and women that were intelligent or independent were marked as being unusual (*strong-minded, clever*).



SEC 1 (1820, 1830, 1840, 1850, 186...): 174,553,979 WORDS

	WORD/PHRASE	TOKENS 1	TOKENS 2	PM 1	PM 2	RATIO
1	STRONG-MINDED WOMEN	24	1	0.1	0.0	14.7
2	CLEVER WOMEN	24	0	0.1	0.0	13.7
3	NOBLE WOMEN	36	2	0.2	0.0	11.0
4	TRUE WOMEN	18	1	0.1	0.0	11.0
5	UNFORTUNATE WOMEN	17	1	0.1	0.0	10.4
6	WRETCHED WOMEN	18	0	0.1	0.0	10.3
7	ABANDONED WOMEN	18	0	0.1	0.0	10.3
8	HELPLESS WOMEN	66	4	0.4	0.0	10.1
9	VERY WOMEN	15	1	0.1	0.0	9.2
10	TURKISH WOMEN	15	1	0.1	0.0	9.2
11	ELDER WOMEN	15	0	0.1	0.0	8.6
12	DEFENCELESS WOMEN	15	0	0.1	0.0	8.6
13	AGED WOMEN	28	2	0.2	0.0	8.6
14	FAIR WOMEN	69	5	0.4	0.0	8.4
15	PURE WOMEN	14	0	0.1	0.0	8.0
16	HANDSOME WOMEN	37	3	0.2	0.0	7.5
17	CULTIVATED WOMEN	13	0	0.1	0.0	7.4
18	REFINED WOMEN	12	0	0.1	0.0	6.9



SEC 2 (1970, 1980, 1990, 2000): 106,640,094 WORDS

	WORD/PHRASE	TOKENS 2	TOKENS 1	PM 2	PM 1	RATIO
1	PREGNANT WOMEN	233	5	2.2	0.0	76.3
2	BATTERED WOMEN	70	0	0.7	0.0	65.6
3	AFRICAN-AMERICAN WOMEN	61	0	0.6	0.0	57.2
4	DIVORCED WOMEN	25	1	0.2	0.0	40.9
5	MIDDLE-CLASS WOMEN	23	1	0.2	0.0	37.6
6	MUSLIM WOMEN	23	1	0.2	0.0	37.6
7	NATIONAL WOMEN	68	3	0.6	0.0	37.1
8	BLACK WOMEN	487	22	4.6	0.1	36.2
9	MENOPAUSAL WOMEN	22	1	0.2	0.0	36.0
10	SOVIET WOMEN	32	0	0.3	0.0	30.0
11	ADULT WOMEN	18	1	0.2	0.0	29.5
12	IMMIGRANT WOMEN	15	1	0.1	0.0	24.6
13	AFGHAN WOMEN	26	0	0.2	0.0	24.4
14	MISSING WOMEN	14	1	0.1	0.0	22.9
15	SUCCESSFUL WOMEN	14	1	0.1	0.0	22.9
16	GOOD-LOOKING WOMEN	13	1	0.1	0.0	21.3
17	MATURE WOMEN	13	1	0.1	0.0	21.3
18	LOCAL WOMEN	22	0	0.2	0.0	20.6

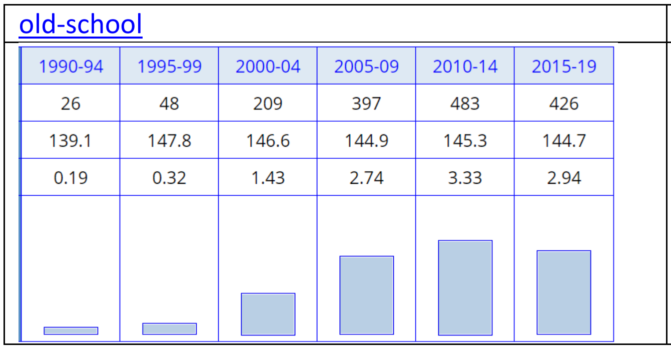
Other than the corpora from English-Corpora.org, *no other historical corpora* are 1) large enough and 2) have a robust enough architecture, to allow studies like these two collocates-based searches. And note that complex searches like those shown above – which provide a wealth of useful data – take just 1-2 seconds in the 400 million word COHA corpus or in any of the other historical corpora.

More recent changes (go to beginning)

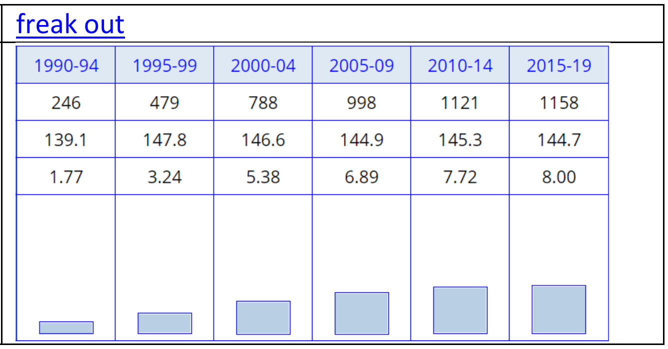
EEBO, COHA, US Supreme Court, and Hansard (British Parliament) focus on changes hundreds of years ago, or during the last 200 years or so. But the corpora from English-Corpora.org are also unique in the way that they allow researchers to look at more recent changes in the language. The Movie Corpus (1930s-2010s) and the TV Corpus (1950s-2010s) are the **only corpora anywhere that focus on recent changes in very informal language**, using large corpora. For example, they show words that were much more common from the 1930s-1960s (left) compared to the 1990s-2010s (right) (including lots more profanity in movies in recent decades).

	More common 1930-1969 (movies)	More common 1990-2018 (movies)
ADJ	swell, splendid, sore, fond, delighted, dreadful, darn, phony, blasted, satisfactory, snappy, darned, apt, no-good, cockeyed, screwy, disgraceful, crummy, beastly, frightful, double-crossing, phoney, bashful, confounded, shrewd, soapy, daffy	f--king, okay, cool, weird, damn, g--d---, huge, awesome, pregnant, super, sexy, scary, unbelievable, sexual, boring, pathetic, gross, massive, nuclear, creepy, global, creative, magical, intense, ultimate, sh-tty, homeless, random, corporate, pissed
NOUN	darling, fellow, pardon, dough, wagon, headquarters, chap, cigar, railroad, brandy, telegram, corporal, crook, hunch, regiment, squadron, handkerchief, shilling, cinch, butler, skipper, chauffeur, plenty, tailor, sonny, mink, nuisance, mammy, waltz, newspaperman	sh-t, hell, mom, f--k, a-s, b-tch, dude, sex, drug, a--h---, tv, bullsh-t, m-f-r, b-st-rd, girlfriend, relationship, d-ck, computer, video, tape, crap, bro, p-ssy, n-g--, grunt, role, bike, chick, cancer, butt
VERB	shall, suppose, pardon, phone, spoil, frighten, telephone, permit, object, congratulate, oblige, dine, notify, faint, quarrel, acquaint, delight, amuse, intrude, dislike, slug, scram, furnish, sock, darn, consent, tangle, fuss, peddle, double-cross	f--k, suck, screw, p-ss, focus, freak, date, r-pe, pee, film, score, b-tch, sh-t, chill, define, stress, evolve, f-rt, activate, surf, tape, participate, process, monitor, target, manipulate, trigger, puke, initiate, generate

We saw above how COCA can be used to look at genre-based variation in English. But because it has almost exactly the same genre-balance each year from 1990-2019, this billion word corpus can also look at **language change during the last 30 years** (and it is the only corpus in the world that allows such searches). For example, users can look at the frequency of words and phrases in five year periods (and if desired, even single years), such as the increase with *old-school* or *freak out* (which is more than four times as frequent than 25-30 years ago).

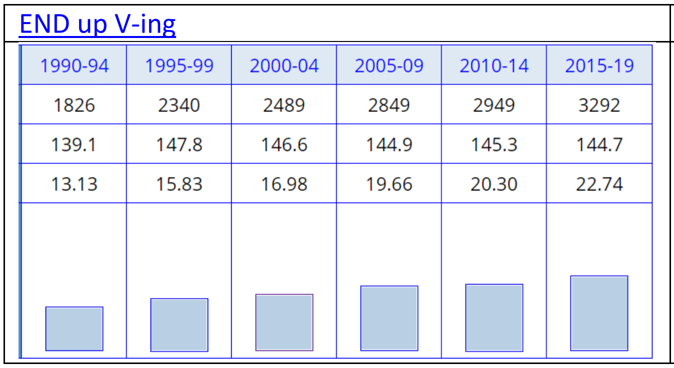

old-school

1990-94	1995-99	2000-04	2005-09	2010-14	2015-19
26	48	209	397	483	426
139.1	147.8	146.6	144.9	145.3	144.7
0.19	0.32	1.43	2.74	3.33	2.94

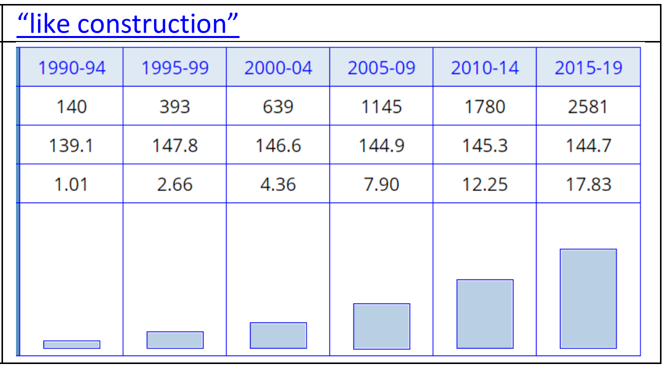

freak out

1990-94	1995-99	2000-04	2005-09	2010-14	2015-19
246	479	788	998	1121	1158
139.1	147.8	146.6	144.9	145.3	144.7
1.77	3.24	5.38	6.89	7.72	8.00

Researchers can also investigate **recent syntactic shifts** in English, such as the increase in END up V-ing (e.g. *we ended up leaving at 9 AM instead*) or the "like construction" (e.g. *and I was like, I guess they can come*).

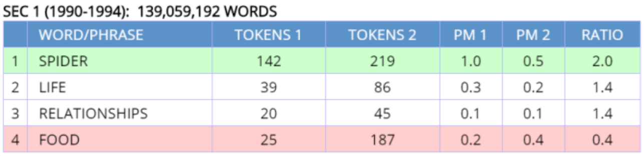

END up V-ing

1990-94	1995-99	2000-04	2005-09	2010-14	2015-19
1826	2340	2489	2849	2949	3292
139.1	147.8	146.6	144.9	145.3	144.7
13.13	15.83	16.98	19.66	20.30	22.74

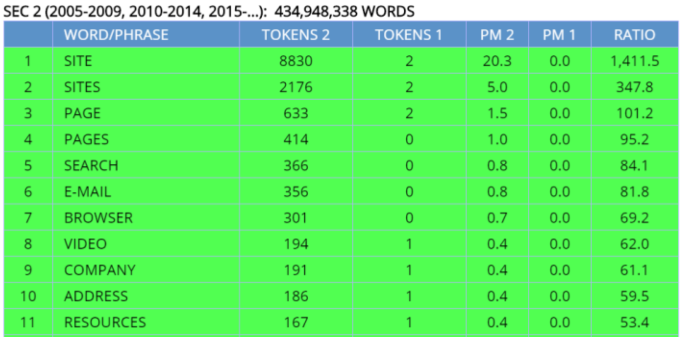

"like construction"

1990-94	1995-99	2000-04	2005-09	2010-14	2015-19
140	393	639	1145	1780	2581
139.1	147.8	146.6	144.9	145.3	144.7
1.01	2.66	4.36	7.90	12.25	17.83

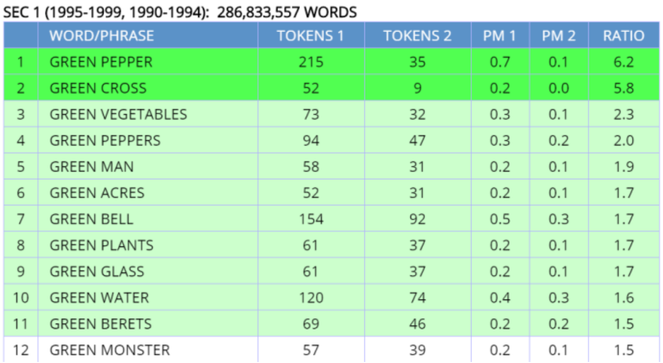
We saw above how collocates could be used with *gay* to show changes in meaning in COHA. We can do the same with words in COCA to show **changes in meaning and usage during the last 30 years**. See the collocates of *web* (note the increase in words(right) referring to the World Wide Web after the early 1990s), and the noun collocates of *green* in the 2010s (below, right), which show the newer meaning of "environmentally friendly".


SEC 1 (1990-1994): 139,059,192 WORDS

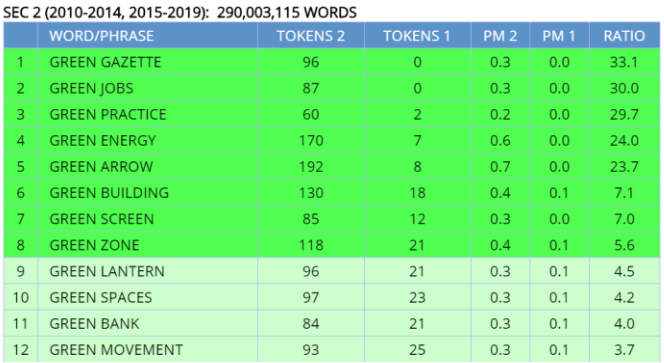
	WORD/PHRASE	TOKENS 1	TOKENS 2	PM 1	PM 2	RATIO
1	SPIDER	142	219	1.0	0.5	2.0
2	LIFE	39	86	0.3	0.2	1.4
3	RELATIONSHIPS	20	45	0.1	0.1	1.4
4	FOOD	25	187	0.2	0.4	0.4


SEC 2 (2005-2009, 2010-2014, 2015-...): 434,948,338 WORDS

	WORD/PHRASE	TOKENS 2	TOKENS 1	PM 2	PM 1	RATIO
1	SITE	8830	2	20.3	0.0	1,411.5
2	SITES	2176	2	5.0	0.0	347.8
3	PAGE	633	2	1.5	0.0	101.2
4	PAGES	414	0	1.0	0.0	95.2
5	SEARCH	366	0	0.8	0.0	84.1
6	E-MAIL	356	0	0.8	0.0	81.8
7	BROWSER	301	0	0.7	0.0	69.2
8	VIDEO	194	1	0.4	0.0	62.0
9	COMPANY	191	1	0.4	0.0	61.1
10	ADDRESS	186	1	0.4	0.0	59.5
11	RESOURCES	167	1	0.4	0.0	53.4


SEC 1 (1995-1999, 1990-1994): 286,833,557 WORDS

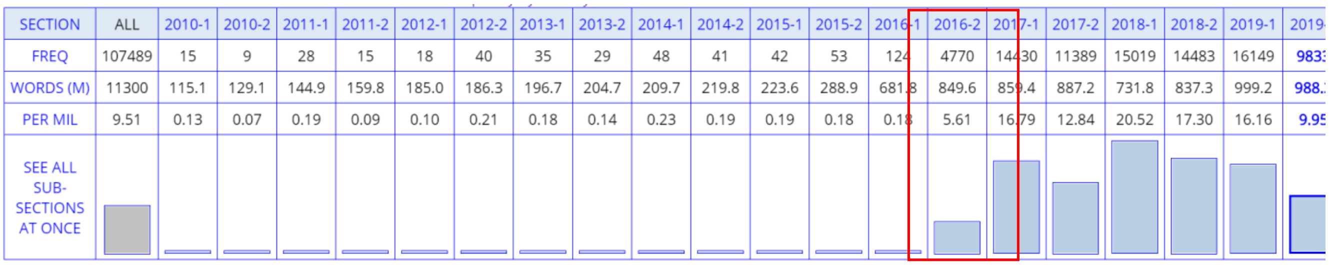
	WORD/PHRASE	TOKENS 1	TOKENS 2	PM 1	PM 2	RATIO
1	GREEN PEPPER	215	35	0.7	0.1	6.2
2	GREEN CROSS	52	9	0.2	0.0	5.8
3	GREEN VEGETABLES	73	32	0.3	0.1	2.3
4	GREEN PEPPERS	94	47	0.3	0.2	2.0
5	GREEN MAN	58	31	0.2	0.1	1.9
6	GREEN ACRES	52	31	0.2	0.1	1.7
7	GREEN BELL	154	92	0.5	0.3	1.7
8	GREEN PLANTS	61	37	0.2	0.1	1.7
9	GREEN GLASS	61	37	0.2	0.1	1.7
10	GREEN WATER	120	74	0.4	0.3	1.6
11	GREEN BERETS	69	46	0.2	0.2	1.5
12	GREEN MONSTER	57	39	0.2	0.1	1.5


SEC 2 (2010-2014, 2015-2019): 290,003,115 WORDS

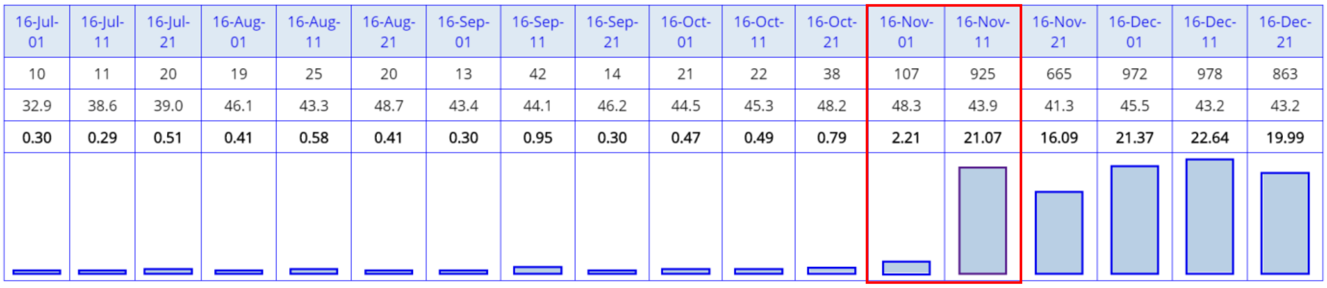
	WORD/PHRASE	TOKENS 2	TOKENS 1	PM 2	PM 1	RATIO
1	GREEN GAZETTE	96	0	0.3	0.0	33.1
2	GREEN JOBS	87	0	0.3	0.0	30.0
3	GREEN PRACTICE	60	2	0.2	0.0	29.7
4	GREEN ENERGY	170	7	0.6	0.0	24.0
5	GREEN ARROW	192	8	0.7	0.0	23.7
6	GREEN BUILDING	130	18	0.4	0.1	7.1
7	GREEN SCREEN	85	12	0.3	0.0	7.0
8	GREEN ZONE	118	21	0.4	0.1	5.6
9	GREEN LANTERN	96	21	0.3	0.1	4.5
10	GREEN SPACES	97	23	0.3	0.1	4.2
11	GREEN BANK	84	21	0.3	0.1	4.0
12	GREEN MOVEMENT	93	25	0.3	0.1	3.7

The **NOW Corpu**s is virtually unique in its ability to look at very recent changes. As of late 2020, it contains about 11.5 billion words from 2010 to the current time (literally, yesterday). *Every day*, 6-10 million words of data are added to the corpus, or about 200-250 million words each month. Users can see the **frequency of words and**

phrases in six-month increments (and even 10-day increments, if desired). For example, the following figures show the spike in *fake news* in the second half of 2016 (2016-2, in the chart), and they can zero in even more to see that it spiked between November 1-10 and November 11-20, which is immediately after the US presidential elections on 8 November 2016.

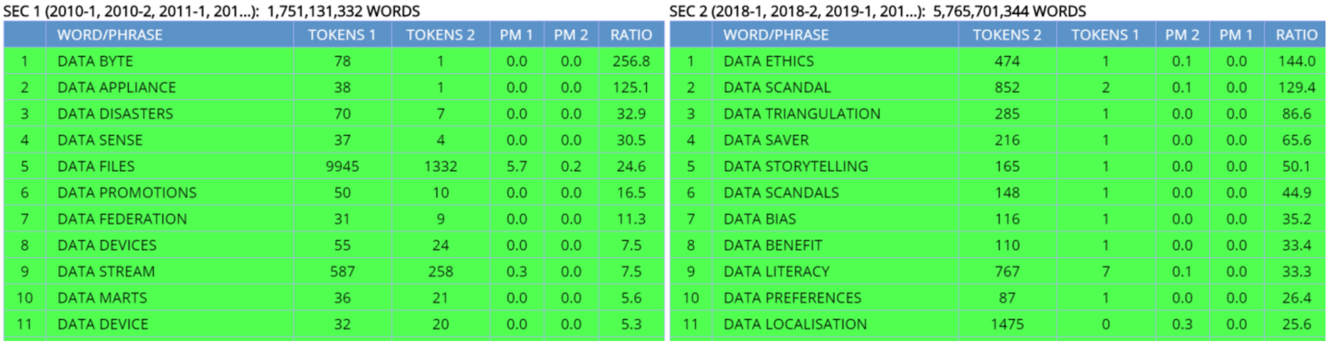


SECTION	ALL	2010-1	2010-2	2011-1	2011-2	2012-1	2012-2	2013-1	2013-2	2014-1	2014-2	2015-1	2015-2	2016-1	2016-2	2017-1	2017-2	2018-1	2018-2	2019-1	2019-
FREQ	107489	15	9	28	15	18	40	35	29	48	41	42	53	124	4770	14430	11389	15019	14483	16149	983
WORDS (M)	11300	115.1	129.1	144.9	159.8	185.0	186.3	196.7	204.7	209.7	219.8	223.6	288.9	681.8	849.6	859.4	887.2	731.8	837.3	999.2	988.
PER MIL	9.51	0.13	0.07	0.19	0.09	0.10	0.21	0.18	0.14	0.23	0.19	0.19	0.18	0.18	5.61	16.79	12.84	20.52	17.30	16.16	9.95
SEE ALL SUB-SECTIONS AT ONCE																					



16-Jul-01	16-Jul-11	16-Jul-21	16-Aug-01	16-Aug-11	16-Aug-21	16-Sep-01	16-Sep-11	16-Sep-21	16-Oct-01	16-Oct-11	16-Oct-21	16-Nov-01	16-Nov-11	16-Nov-21	16-Dec-01	16-Dec-11	16-Dec-21
10	11	20	19	25	20	13	42	14	21	22	38	107	925	665	972	978	863
32.9	38.6	39.0	46.1	43.3	48.7	43.4	44.1	46.2	44.5	45.3	48.2	48.3	43.9	41.3	45.5	43.2	43.2
0.30	**0.29**	**0.51**	**0.41**	**0.58**	**0.41**	**0.30**	**0.95**	**0.30**	**0.47**	**0.49**	**0.79**	**2.21**	**21.07**	**16.09**	**21.37**	**22.64**	**19.99**

The corpus also shows **changes in phrases during the last ten years**, such as phrases with *data* + NOUN that are more frequent from 2018-2020 (right; e.g. *data ethics, data scandal*) than in 2010-2013 (left).



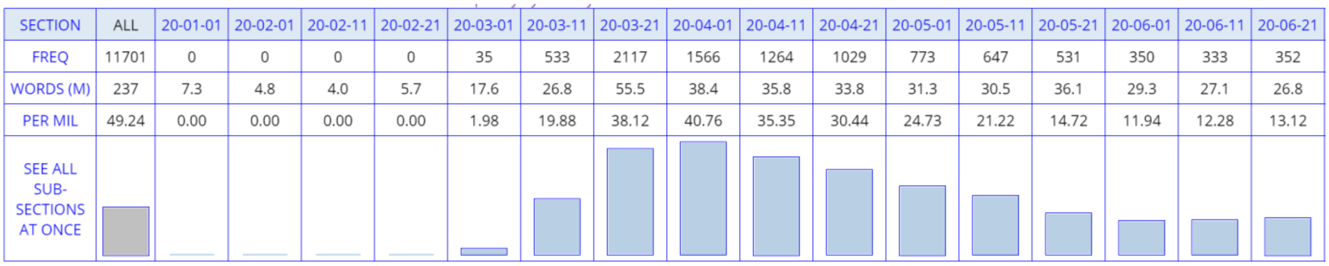
SEC 1 (2010-1, 2010-2, 2011-1, 201...): 1,751,131,332 WORDS

	WORD/PHRASE	TOKENS 1	TOKENS 2	PM 1	PM 2	RATIO
1	DATA BYTE	78	1	0.0	0.0	256.8
2	DATA APPLIANCE	38	1	0.0	0.0	125.1
3	DATA DISASTERS	70	7	0.0	0.0	32.9
4	DATA SENSE	37	4	0.0	0.0	30.5
5	DATA FILES	9945	1332	5.7	0.2	24.6
6	DATA PROMOTIONS	50	10	0.0	0.0	16.5
7	DATA FEDERATION	31	9	0.0	0.0	11.3
8	DATA DEVICES	55	24	0.0	0.0	7.5
9	DATA STREAM	587	258	0.3	0.0	7.5
10	DATA MARTS	36	21	0.0	0.0	5.6
11	DATA DEVICE	32	20	0.0	0.0	5.3

SEC 2 (2018-1, 2018-2, 2019-1, 201...): 5,765,701,344 WORDS

	WORD/PHRASE	TOKENS 2	TOKENS 1	PM 2	PM 1	RATIO
1	DATA ETHICS	474	1	0.1	0.0	144.0
2	DATA SCANDAL	852	2	0.1	0.0	129.4
3	DATA TRIANGULATION	285	1	0.0	0.0	86.6
4	DATA SAVER	216	1	0.0	0.0	65.6
5	DATA STORYTELLING	165	1	0.0	0.0	50.1
6	DATA SCANDALS	148	1	0.0	0.0	44.9
7	DATA BIAS	116	1	0.0	0.0	35.2
8	DATA BENEFIT	110	1	0.0	0.0	33.4
9	DATA LITERACY	767	7	0.1	0.0	33.3
10	DATA PREFERENCES	87	1	0.0	0.0	26.4
11	DATA LOCALISATION	1475	0	0.3	0.0	25.6

The **Coronavirus Corpus** is a subset of the NOW Corpus, and it contains articles from 2020 and beyond, which deal with COVID-19. As of late 2020 it is about 700 million words in size, and it is growing by about 60-70 million words each month. It shows the **frequency of words and phrases in ten-day increments** since January 2020, such as *flatten the curve*, which peaks in mid-March 2020, and has then "flattened out" since June 2020.



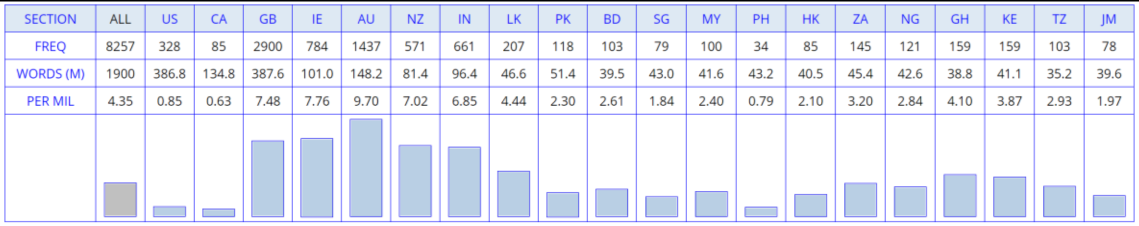
SECTION	ALL	20-01-01	20-02-01	20-02-11	20-02-21	20-03-01	20-03-11	20-03-21	20-04-01	20-04-11	20-04-21	20-05-01	20-05-11	20-05-21	20-06-01	20-06-11	20-06-21
FREQ	11701	0	0	0	0	35	533	2117	1566	1264	1029	773	647	531	350	333	352
WORDS (M)	237	7.3	4.8	4.0	5.7	17.6	26.8	55.5	38.4	35.8	33.8	31.3	30.5	36.1	29.3	27.1	26.8
PER MIL	49.24	0.00	0.00	0.00	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
SEE ALL SUB-SECTIONS AT ONCE																	

Dialectal variation (go to beginning)

The **GloWbE** Corpus contains about two billion words from 20 different English-speaking countries, and it allows researchers **to look at changes between dialects in ways that are not possible with any other corpus**. Since it was released in 2013, a large number of articles have been published that are based on this corpus.

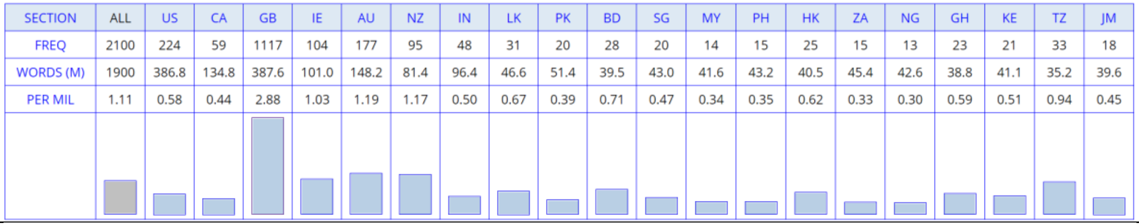
At the most basic level, researchers can see the frequency of a word or phrase in all 20 countries, such as *fortnight* (notice its virtual absence in American and Canadian English, as well as Philippine English, which is based on American English), *rather more ADJ* (definitely the most frequent in GB: Great Britain), *Eve teas** (which means "sexual harassment", and a word that is found almost exclusively in South Asia), and *equipments* (note the plural form), which occurs in most of the countries other than the six "Inner Circle" countries (US, Canada, Great Britain, Ireland, Australia, and New Zealand).

fortnight



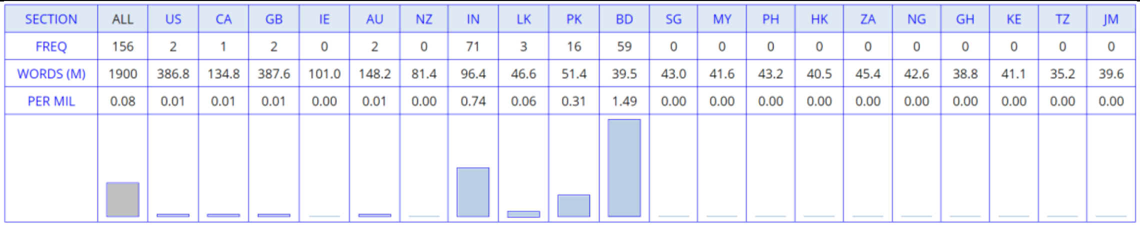
SECTION	ALL	US	CA	GB	IE	AU	NZ	IN	LK	PK	BD	SG	MY	PH	HK	ZA	NG	GH	KE	TZ	JM
FREQ	8257	328	85	2900	784	1437	571	661	207	118	103	79	100	34	85	145	121	159	159	103	78
WORDS (M)	1900	386.8	134.8	387.6	101.0	148.2	81.4	96.4	46.6	51.4	39.5	43.0	41.6	43.2	40.5	45.4	42.6	38.8	41.1	35.2	39.6
PER MIL	4.35	0.85	0.63	7.48	7.76	9.70	7.02	6.85	4.44	2.30	2.61	1.84	2.40	0.79	2.10	3.20	2.84	4.10	3.87	2.93	1.97

rather more ADJ



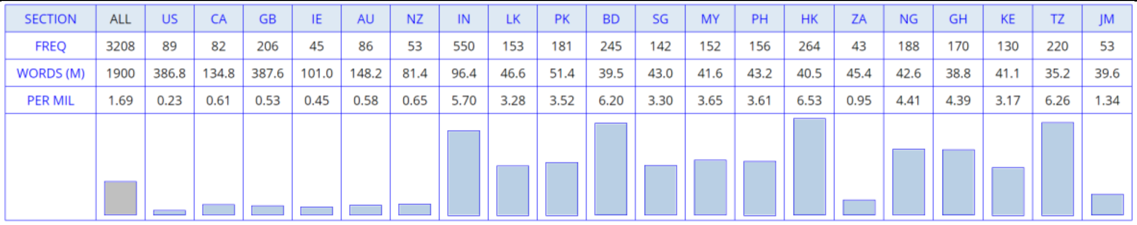
SECTION	ALL	US	CA	GB	IE	AU	NZ	IN	LK	PK	BD	SG	MY	PH	HK	ZA	NG	GH	KE	TZ	JM
FREQ	2100	224	59	1117	104	177	95	48	31	20	28	20	14	15	25	15	13	23	21	33	18
WORDS (M)	1900	386.8	134.8	387.6	101.0	148.2	81.4	96.4	46.6	51.4	39.5	43.0	41.6	43.2	40.5	45.4	42.6	38.8	41.1	35.2	39.6
PER MIL	1.11	0.58	0.44	2.88	1.03	1.19	1.17	0.50	0.67	0.39	0.71	0.47	0.34	0.35	0.62	0.33	0.30	0.59	0.51	0.94	0.45

Eve teas*



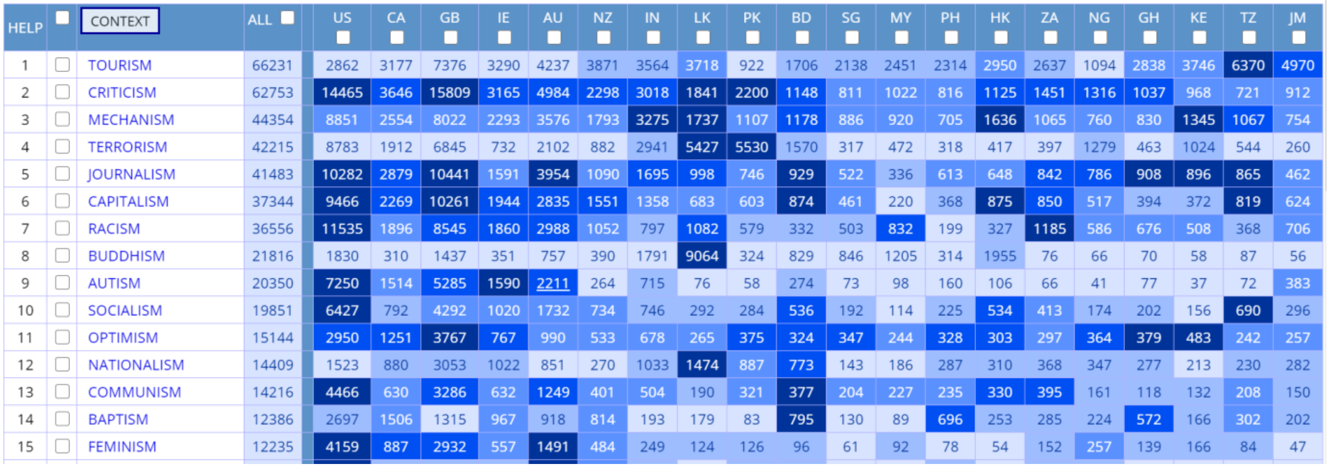
SECTION	ALL	US	CA	GB	IE	AU	NZ	IN	LK	PK	BD	SG	MY	PH	HK	ZA	NG	GH	KE	TZ	JM
FREQ	156	2	1	2	0	2	0	71	3	16	59	0	0	0	0	0	0	0	0	0	0
WORDS (M)	1900	386.8	134.8	387.6	101.0	148.2	81.4	96.4	46.6	51.4	39.5	43.0	41.6	43.2	40.5	45.4	42.6	38.8	41.1	35.2	39.6
PER MIL	0.08	0.01	0.01	0.01	0.00	0.01	0.00	0.74	0.06	0.31	1.49	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

equipments



SECTION	ALL	US	CA	GB	IE	AU	NZ	IN	LK	PK	BD	SG	MY	PH	HK	ZA	NG	GH	KE	TZ	JM
FREQ	3208	89	82	206	45	86	53	550	153	181	245	142	152	156	264	43	188	170	130	220	53
WORDS (M)	1900	386.8	134.8	387.6	101.0	148.2	81.4	96.4	46.6	51.4	39.5	43.0	41.6	43.2	40.5	45.4	42.6	38.8	41.1	35.2	39.6
PER MIL	1.69	0.23	0.61	0.53	0.45	0.58	0.65	5.70	3.28	3.52	6.20	3.30	3.65	3.61	6.53	0.95	4.41	4.39	3.17	6.26	1.34

It is also possible to see the frequency of a number of words matching a particular string, in all 20 countries. For example, the following chart shows the most frequent **ism* words.

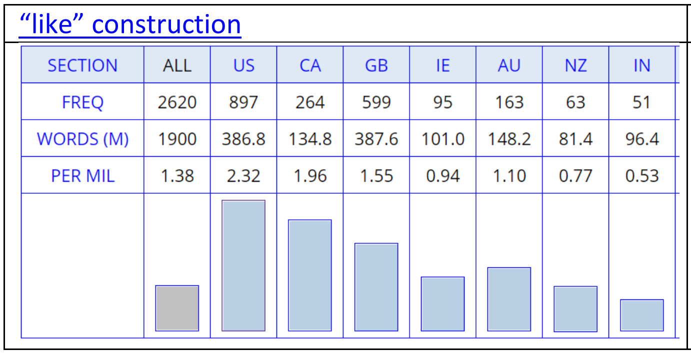


HELP	☐	CONTEXT	ALL ☐	US ☐	CA ☐	GB ☐	IE ☐	AU ☐	NZ ☐	IN ☐	LK ☐	PK ☐	BD ☐	SG ☐	MY ☐	PH ☐	HK ☐	ZA ☐	NG ☐	GH ☐	KE ☐	TZ ☐	JM ☐
1	☐	TOURISM	66231	2862	3177	7376	3290	4237	3871	3564	3718	922	1706	2138	2451	2314	2950	2637	1094	2838	3746	6370	4970
2	☐	CRITICISM	62753	14465	3646	15809	3165	4984	2298	3018	1841	2200	1148	811	1022	816	1125	1451	1316	1037	968	721	912
3	☐	MECHANISM	44354	8851	2554	8022	2293	3576	1793	3275	1737	1107	1178	886	920	705	1636	1065	760	830	1345	1067	754
4	☐	TERRORISM	42215	8783	1912	6845	732	2102	882	2941	5427	5530	1570	317	472	318	417	397	1279	463	1024	544	260
5	☐	JOURNALISM	41483	10282	2879	10441	1591	3954	1090	1695	998	746	929	522	336	613	648	842	786	908	896	865	462
6	☐	CAPITALISM	37344	9466	2269	10261	1944	2835	1551	1358	683	603	874	461	220	368	875	850	517	394	372	819	624
7	☐	RACISM	36556	11535	1896	8545	1860	2988	1052	797	1082	579	332	503	832	199	327	1185	586	676	508	368	706
8	☐	BUDDHISM	21816	1830	310	1437	351	757	390	1791	9064	324	829	846	1205	314	1955	76	66	70	58	87	56
9	☐	AUTISM	20350	7250	1514	5285	1590	2211	264	715	76	58	274	73	98	160	106	66	41	77	37	72	383
10	☐	SOCIALISM	19851	6427	792	4292	1020	1732	734	746	292	284	536	192	114	225	534	413	174	202	156	690	296
11	☐	OPTIMISM	15144	2950	1251	3767	767	990	533	678	265	375	324	347	244	328	303	297	364	379	483	242	257
12	☐	NATIONALISM	14409	1523	880	3053	1022	851	270	1033	1474	887	773	143	186	287	310	368	347	277	213	230	282
13	☐	COMMUNISM	14216	4466	630	3286	632	1249	401	504	190	321	377	204	227	235	330	395	161	118	132	208	150
14	☐	BAPTISM	12386	2697	1506	1315	967	918	814	193	179	83	795	130	89	696	253	285	224	572	166	302	202
15	☐	FEMINISM	12235	4159	887	2932	557	1491	484	249	124	126	96	61	92	78	54	152	257	139	166	84	47

The **TV Corpus and Movies corpora** can also provide useful information on **differences between dialects**, since they contain 575 million words of data of **extremely informal English** from the six "Inner Circle" countries. For example, the following table shows words that are much more common in American or in British English. Of course these two corpora could also compare anything else between these six dialects, including word formation, syntax, or word meaning and usage (via collocates).

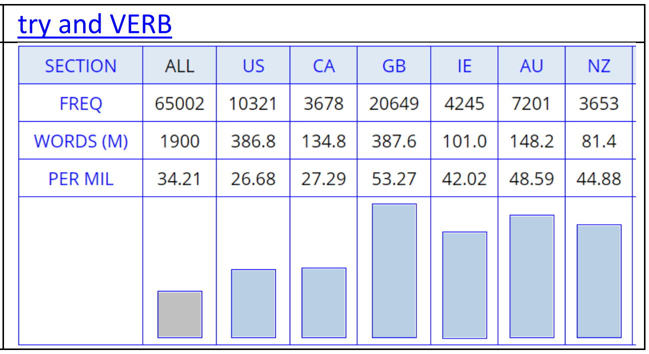
	American	British
ADJ	okay, crazy, damn, awesome, cute, dumb, federal, goddamn, gross, lame, adorable, lousy, crappy, sloppy, phony, downtown, cozy, busted, darn, cranky, high-end, one-time, high-school, canned, cellular, big-time, African-American, goofy, off-limits, old-school, sassy, condescending, puffy, big-a--, sketchy, wordy, charmed, disoriented, kick-a--, bitchy, narcissistic, crummy, self-centered, curt, trashy, whimsical, dorky, scrappy	daft, posh, dodgy, knackered, ruddy, barmy, sodding, poxy, dozy, soppy, mucky, disused, chuffed, tinned, whirly, manky, disorientated, pish, fiddly
NOUN	guy, mom, honey, dude, cop, agent, a--, movie, buddy, apartment, truck, chef, buck, dollar, sweetie, mommy, attorney, mayor, butt, cookie, grandma, a--h---, candy, grade, parking, senator, couch, vacation, closet, homicide, garbage, jerk, baseball, grandpa, elevator, trash, math, thanksgiving, shooter, roommate, bud, assignment, prom, tech, mall, dessert, heck, bout, zombie, soda, motel, halloween, therapist, basketball, counselor, lawsuit, diaper, congressman, chili,	mum, bloke, a-se, quid, rubbish, b-ll-ck, solicitor, railway, vicar, telly, guv, grandad, petrol, ladyship, mammy, shilling, maths, lorry, a---h---, advert, motorway, tosser, tenner, pence, nutter, punter, gearbox, footballer, windscreen, pensioner, barman, pram, tuppence, prat, flatmate, lodger, roundabout, vicarage, workhouse, pillock, sixpence
VERB	guess, figure, kid, damn, date, quit, hire, freak, yell, bust, file, hook, testify, pee, coach, assign, schedule, graduate, violate, practice, dial, jerk, sniffle, participate, brag, party, merge, poop, hustle, reschedule	reckon, fancy, shag, sod, flog, w-nk, queue, burgle, snigger, snog, plod, splutter, clamber

A number of studies have also used GloWbE to examine **syntactic differences between the different dialects**. To provide two simple examples here, the "like construction" (*and I'm like, no way can they do it*) is the most frequent in American English, but it also occurs in other related "Inner Circle" countries, like Canada, Great Britain, Ireland, Australia, and New Zealand (although less in each successive country). The second chart looks at the construction *try and VERB* (*I'm gonna try and talk to her,* vs *try to talk*), which is stigmatized as being "incorrect" in American and Canadian English (due to certain prescriptive grammars in these two countries 50-100 years ago). But in the other countries (where the prescriptive rule was never as important), the construction is much more common.



“like” construction

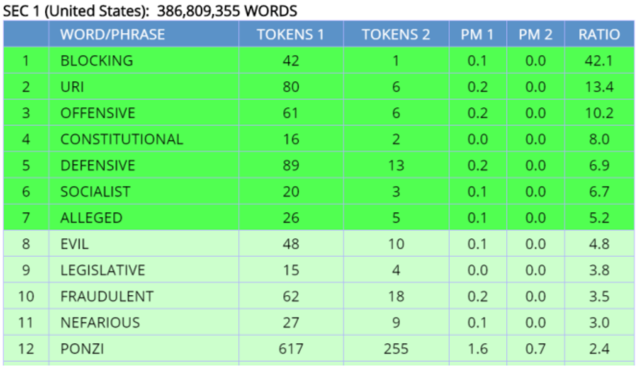
SECTION	ALL	US	CA	GB	IE	AU	NZ	IN
FREQ	2620	897	264	599	95	163	63	51
WORDS (M)	1900	386.8	134.8	387.6	101.0	148.2	81.4	96.4
PER MIL	1.38	2.32	1.96	1.55	0.94	1.10	0.77	0.53



try and VERB

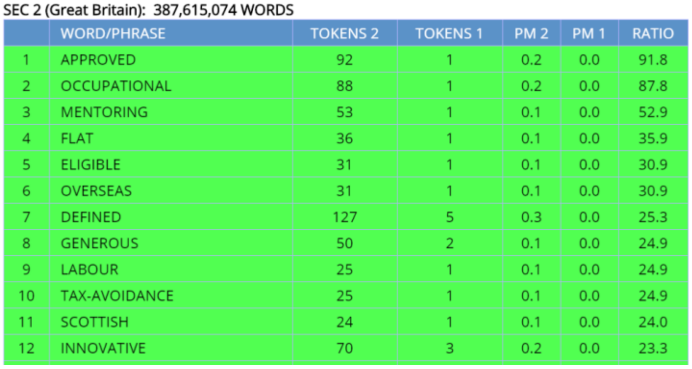
SECTION	ALL	US	CA	GB	IE	AU	NZ
FREQ	65002	10321	3678	20649	4245	7201	3653
WORDS (M)	1900	386.8	134.8	387.6	101.0	148.2	81.4
PER MIL	34.21	26.68	27.29	53.27	42.02	48.59	44.88

Due to the size of GloWbE (nearly two billion words) it is also possible to use collocates to look at **differences in meaning and usage between dialects**. For example, this chart shows the collocates of *scheme*, and shows that the word is much more negative in American English than in British English, as evidenced by the collocates (*alleged, evil, fraudulent, nefarious*).



SEC 1 (United States): 386,809,355 WORDS

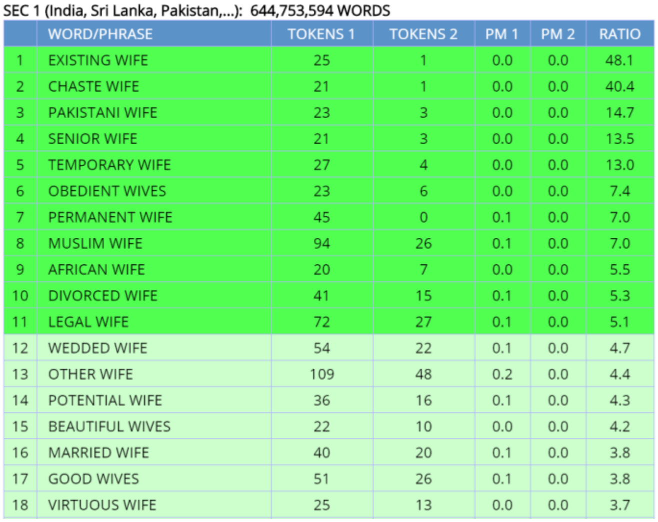
	WORD/PHRASE	TOKENS 1	TOKENS 2	PM 1	PM 2	RATIO
1	BLOCKING	42	1	0.1	0.0	42.1
2	URI	80	6	0.2	0.0	13.4
3	OFFENSIVE	61	6	0.2	0.0	10.2
4	CONSTITUTIONAL	16	2	0.0	0.0	8.0
5	DEFENSIVE	89	13	0.2	0.0	6.9
6	SOCIALIST	20	3	0.1	0.0	6.7
7	ALLEGED	26	5	0.1	0.0	5.2
8	EVIL	48	10	0.1	0.0	4.8
9	LEGISLATIVE	15	4	0.0	0.0	3.8
10	FRAUDULENT	62	18	0.2	0.0	3.5
11	NEFARIOUS	27	9	0.1	0.0	3.0
12	PONZI	617	255	1.6	0.7	2.4



SEC 2 (Great Britain): 387,615,074 WORDS

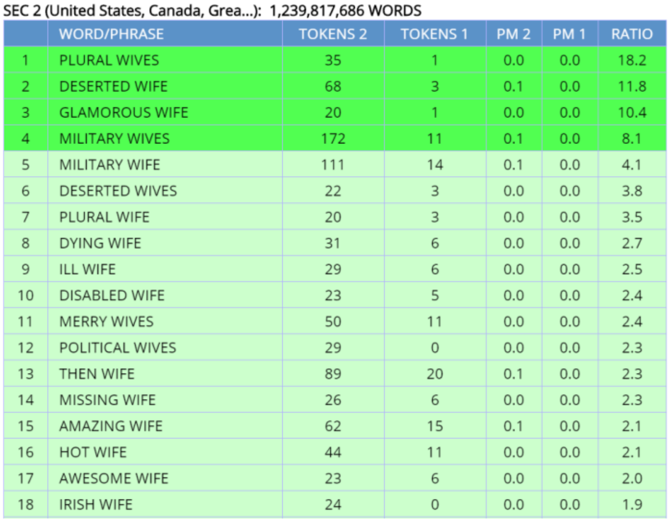
	WORD/PHRASE	TOKENS 2	TOKENS 1	PM 2	PM 1	RATIO
1	APPROVED	92	1	0.2	0.0	91.8
2	OCCUPATIONAL	88	1	0.2	0.0	87.8
3	MENTORING	53	1	0.1	0.0	52.9
4	FLAT	36	1	0.1	0.0	35.9
5	ELIGIBLE	31	1	0.1	0.0	30.9
6	OVERSEAS	31	1	0.1	0.0	30.9
7	DEFINED	127	5	0.3	0.0	25.3
8	GENEROUS	50	2	0.1	0.0	24.9
9	LABOUR	25	1	0.1	0.0	24.9
10	TAX-AVOIDANCE	25	1	0.1	0.0	24.9
11	SCOTTISH	24	1	0.1	0.0	24.0
12	INNOVATIVE	70	3	0.2	0.0	23.3

We can also use collocates to compare what is being said about different topics in different dialects, which may indicate interesting differences in culture and society. For example, the collocates of *wife* in the dialects of Asia and Africa (left) include words like *existing, temporary*, and *permanent*, which relate to cultural practices in these countries. Other collocates such as *chaste, obedient, good*, and *virtuous* also signal important cultural practices and norms in these countries. As we can see, a simple 2-3 second search can – with the right corpus – show interesting differences between the cultures of the different countries, which may be of interest to social scientists (in addition to linguists).



SEC 1 (India, Sri Lanka, Pakistan,...): 644,753,594 WORDS

	WORD/PHRASE	TOKENS 1	TOKENS 2	PM 1	PM 2	RATIO
1	EXISTING WIFE	25	1	0.0	0.0	48.1
2	CHASTE WIFE	21	1	0.0	0.0	40.4
3	PAKISTANI WIFE	23	3	0.0	0.0	14.7
4	SENIOR WIFE	21	3	0.0	0.0	13.5
5	TEMPORARY WIFE	27	4	0.0	0.0	13.0
6	OBEDIENT WIVES	23	6	0.0	0.0	7.4
7	PERMANENT WIFE	45	0	0.1	0.0	7.0
8	MUSLIM WIFE	94	26	0.1	0.0	7.0
9	AFRICAN WIFE	20	7	0.0	0.0	5.5
10	DIVORCED WIFE	41	15	0.1	0.0	5.3
11	LEGAL WIFE	72	27	0.1	0.0	5.1
12	WEDDED WIFE	54	22	0.1	0.0	4.7
13	OTHER WIFE	109	48	0.2	0.0	4.4
14	POTENTIAL WIFE	36	16	0.1	0.0	4.3
15	BEAUTIFUL WIVES	22	10	0.0	0.0	4.2
16	MARRIED WIFE	40	20	0.1	0.0	3.8
17	GOOD WIVES	51	26	0.1	0.0	3.8
18	VIRTUOUS WIFE	25	13	0.0	0.0	3.7



SEC 2 (United States, Canada, Grea...): 1,239,817,686 WORDS

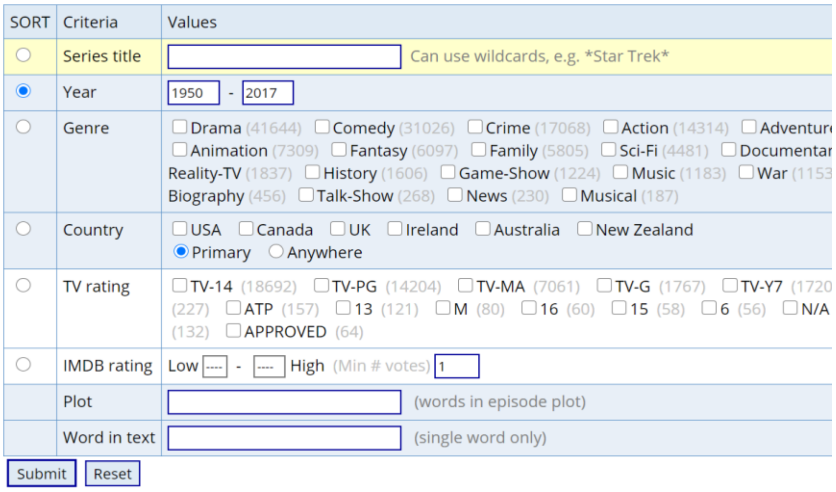
	WORD/PHRASE	TOKENS 2	TOKENS 1	PM 2	PM 1	RATIO
1	PLURAL WIVES	35	1	0.0	0.0	18.2
2	DESERTED WIFE	68	3	0.1	0.0	11.8
3	GLAMOROUS WIFE	20	1	0.0	0.0	10.4
4	MILITARY WIVES	172	11	0.1	0.0	8.1
5	MILITARY WIFE	111	14	0.1	0.0	4.1
6	DESERTED WIVES	22	3	0.0	0.0	3.8
7	PLURAL WIFE	20	3	0.0	0.0	3.5
8	DYING WIFE	31	6	0.0	0.0	2.7
9	ILL WIFE	29	6	0.0	0.0	2.5
10	DISABLED WIFE	23	5	0.0	0.0	2.4
11	MERRY WIVES	50	11	0.0	0.0	2.4
12	POLITICAL WIVES	29	0	0.0	0.0	2.3
13	THEN WIFE	89	20	0.1	0.0	2.3
14	MISSING WIFE	26	6	0.0	0.0	2.3
15	AMAZING WIFE	62	15	0.1	0.0	2.1
16	HOT WIFE	44	11	0.0	0.0	2.1
17	AWESOME WIFE	23	6	0.0	0.0	2.0
18	IRISH WIFE	24	0	0.0	0.0	1.9

Virtual Corpora (go to beginning)

In the sections above, the corpora have been divided into sections that the researcher can use for their searches – such as genres, decades, or countries. But users can quickly and easily **create their own collections of texts in the corpora, and then search that "Virtual Corpus"** just as if it were its own corpus. For example, they could focus on texts dealing with any topic (e.g. biology, investments, nuclear energy, basketball, or Harry Potter), a specific author or source (e.g. the *New York Times*, or *Astronomy* magazine), a specific sub-genre (e.g. reality shows in the TV Corpus, or finance articles in COCA or the BNC), a particular date range, or any combination of these.

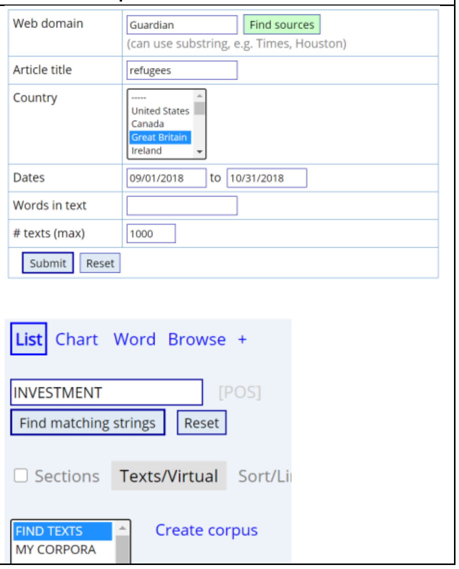
For example, the following is the page that researchers can use in the TV Corpus (left) and in the NOW Corpus (right) to create a Virtual Corpus, and similar pages are available in each of the 17 corpora from English-Corpora.org. They can also quickly and easily create a Virtual Corpus based just on words or phrases (lower, right).

TV Corpus



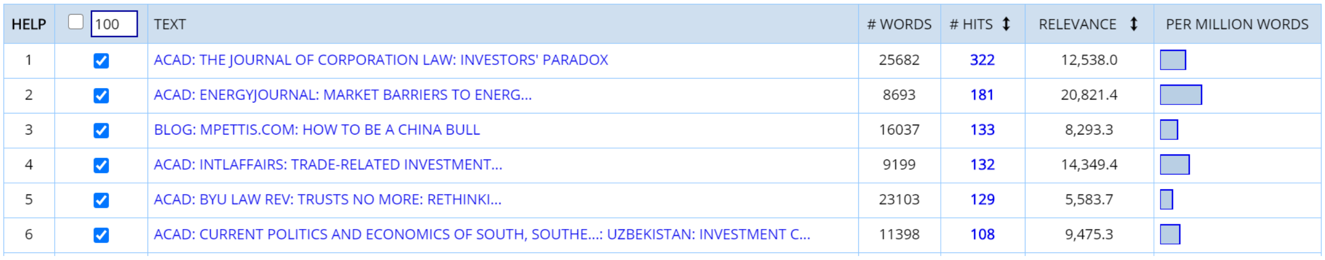
SORT	Criteria	Values
○	Series title	Can use wildcards, e.g. *Star Trek*
●	Year	1950 - 2017
○	Genre	□Drama (41644) □Comedy (31026) □Crime (17068) □Action (14314) □Adventur □Animation (7309) □Fantasy (6097) □Family (5805) □Sci-Fi (4481) □Documentar Reality-TV (1837) □History (1606) □Game-Show (1224) □Music (1183) □War (1153 Biography (456) □Talk-Show (268) □News (230) □Musical (187)
○	Country	□USA □Canada □UK □Ireland □Australia □New Zealand ●Primary ○Anywhere
○	TV rating	□TV-14 (18692) □TV-PG (14204) □TV-MA (7061) □TV-G (1767) □TV-Y7 (1720 (227) □ATP (157) □13 (121) □M (80) □16 (60) □15 (58) □6 (56) □N/A (132) □APPROVED (64)
○	IMDB rating	Low ---- - ---- High (Min # votes) 1
	Plot	(words in episode plot)
	Word in text	(single word only)

NOW Corpus



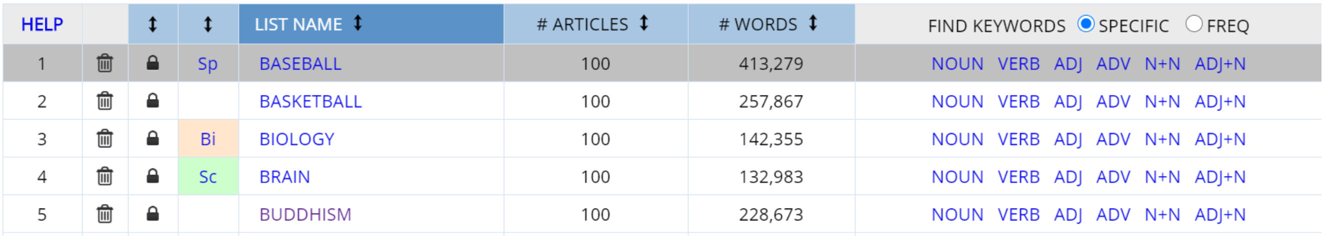
Web domain	Guardian (can use substring, e.g. Times, Houston)
Article title	refugees
Country	----- United States Canada Great Britain Ireland
Dates	09/01/2018 to 10/31/2018
Words in text	
# texts (max)	1000

The corpus then finds what it thinks are the best texts for the search, and users can select among these texts. They can also add and delete texts, or copy or move texts between other Virtual Corpora.



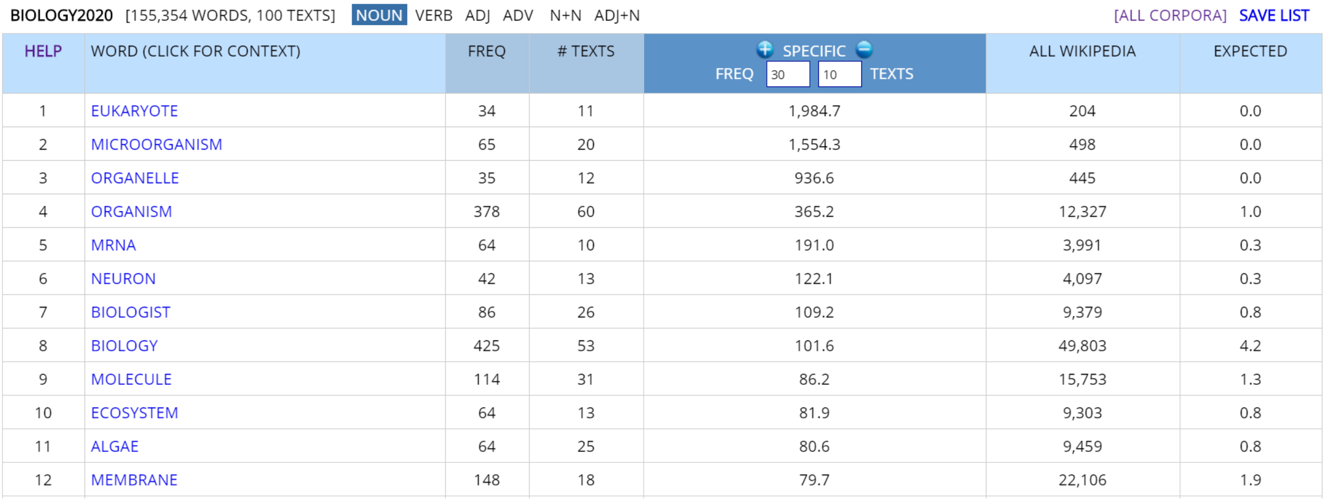
HELP	□ 100	TEXT	# WORDS	# HITS ↕	RELEVANCE ↕	PER MILLION WORDS
1	☑	ACAD: THE JOURNAL OF CORPORATION LAW: INVESTORS' PARADOX	25682	322	12,538.0	
2	☑	ACAD: ENERGYJOURNAL: MARKET BARRIERS TO ENERG...	8693	181	20,821.4	
3	☑	BLOG: MPETTIS.COM: HOW TO BE A CHINA BULL	16037	133	8,293.3	
4	☑	ACAD: INTLAFFAIRS: TRADE-RELATED INVESTMENT...	9199	132	14,349.4	
5	☑	ACAD: BYU LAW REV: TRUSTS NO MORE: RETHINKI...	23103	129	5,583.7	
6	☑	ACAD: CURRENT POLITICS AND ECONOMICS OF SOUTH, SOUTHE...: UZBEKISTAN: INVESTMENT C...	11398	108	9,475.3	

They can see all of their Virtual Corpora, and can organize them into user-defined category (e.g. science, finance, or sports).



HELP		↕	↕	LIST NAME ↕	# ARTICLES ↕	# WORDS ↕	FIND KEYWORDS ● SPECIFIC ○ FREQ
1	🗑	🔒	Sp	BASEBALL	100	413,279	NOUN VERB ADJ ADV N+N ADJ+N
2	🗑	🔒		BASKETBALL	100	257,867	NOUN VERB ADJ ADV N+N ADJ+N
3	🗑	🔒	Bi	BIOLOGY	100	142,355	NOUN VERB ADJ ADV N+N ADJ+N
4	🗑	🔒	Sc	BRAIN	100	132,983	NOUN VERB ADJ ADV N+N ADJ+N
5	🗑	🔒		BUDDHISM	100	228,673	NOUN VERB ADJ ADV N+N ADJ+N

Perhaps most importantly, they can see **keyword lists** from their Virtual Corpora, and can adjust how specific the words are to the Virtual Corpus. The following words are from the [biology] Virtual Corpus was created in the Wikipedia Corpus.

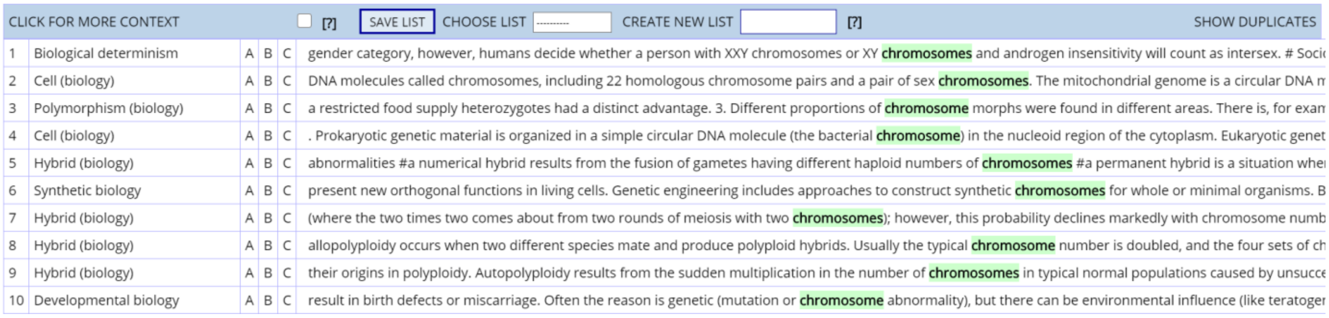


BIOLOGY2020 [155,354 WORDS, 100 TEXTS] NOUN VERB ADJ ADV N+N ADJ+N [ALL CORPORA] SAVE LIST

SPECIFIC: FREQ 30 10 TEXTS

HELP	WORD (CLICK FOR CONTEXT)	FREQ	# TEXTS	SPECIFIC	ALL WIKIPEDIA	EXPECTED
1	EUKARYOTE	34	11	1,984.7	204	0.0
2	MICROORGANISM	65	20	1,554.3	498	0.0
3	ORGANELLE	35	12	936.6	445	0.0
4	ORGANISM	378	60	365.2	12,327	1.0
5	MRNA	64	10	191.0	3,991	0.3
6	NEURON	42	13	122.1	4,097	0.3
7	BIOLOGIST	86	26	109.2	9,379	0.8
8	BIOLOGY	425	53	101.6	49,803	4.2
9	MOLECULE	114	31	86.2	15,753	1.3
10	ECOSYSTEM	64	13	81.9	9,303	0.8
11	ALGAE	64	25	80.6	9,459	0.8
12	MEMBRANE	148	18	79.7	22,106	1.9

When users click on a keyword, they see the concordance lines from this particular Virtual Corpus:



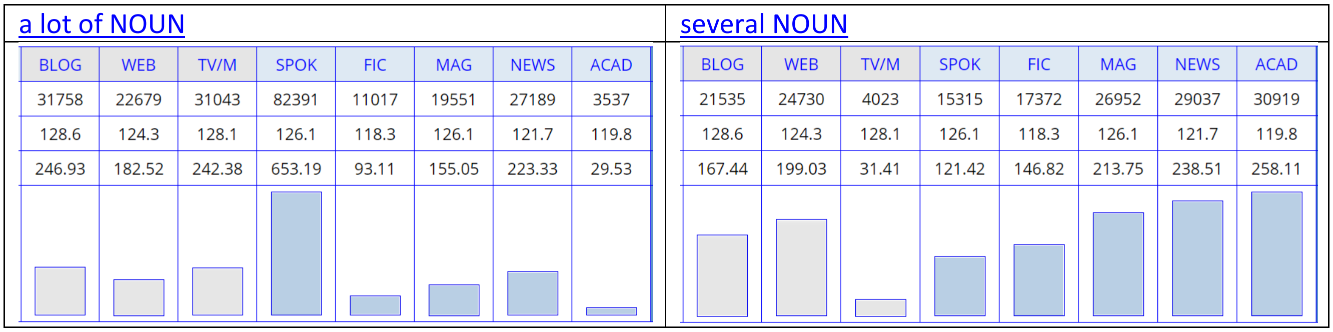
CLICK FOR MORE CONTEXT □ [?] SAVE LIST CHOOSE LIST --------- CREATE NEW LIST [?] SHOW DUPLICATES

1	Biological determinism	A	B	C	gender category, however, humans decide whether a person with XXY chromosomes or XY **chromosomes** and androgen insensitivity will count as intersex. # Soci
2	Cell (biology)	A	B	C	DNA molecules called chromosomes, including 22 homologous chromosome pairs and a pair of sex **chromosomes**. The mitochondrial genome is a circular DNA m
3	Polymorphism (biology)	A	B	C	a restricted food supply heterozygotes had a distinct advantage. 3. Different proportions of **chromosome** morphs were found in different areas. There is, for exan
4	Cell (biology)	A	B	C	. Prokaryotic genetic material is organized in a simple circular DNA molecule (the bacterial **chromosome**) in the nucleoid region of the cytoplasm. Eukaryotic genet
5	Hybrid (biology)	A	B	C	abnormalities #a numerical hybrid results from the fusion of gametes having different haploid numbers of **chromosomes** #a permanent hybrid is a situation whe
6	Synthetic biology	A	B	C	present new orthogonal functions in living cells. Genetic engineering includes approaches to construct synthetic **chromosomes** for whole or minimal organisms. B
7	Hybrid (biology)	A	B	C	(where the two times two comes about from two rounds of meiosis with two **chromosomes**); however, this probability declines markedly with chromosome numb
8	Hybrid (biology)	A	B	C	allopolyploidy occurs when two different species mate and produce polyploid hybrids. Usually the typical **chromosome** number is doubled, and the four sets of ch
9	Hybrid (biology)	A	B	C	their origins in polyploidy. Autopolyploidy results from the sudden multiplication in the number of **chromosomes** in typical normal populations caused by unsucce
10	Developmental biology	A	B	C	result in birth defects or miscarriage. Often the reason is genetic (mutation or **chromosome** abnormality), but there can be environmental influence (like teratoger

And of course, they can do any other corpus search – word, phrase, substring, synonyms, collocates, etc – and then limit the search just to a particular Virtual Corpus. In this way, a Virtual Corpus is like a "**corpus within a corpus**", and it may be much more useful to researchers who are interested in a specific topic. And unlike other corpus sites, it takes just a few clicks and a few seconds to create Virtual Corpora at English-Corpora.org.

Tools for language learners and teachers (go to beginning)

Many of the searches shown above provide useful information for learners and teachers of English. Simple **frequency charts** can be useful to have students "calibrate" their usage for particular genres. For example, learners might not know intuitively that the phrase *a lot of* sounds very informal and that it is very uncommon in academic writing, whereas *several NOUN* sounds much better in formal writing:



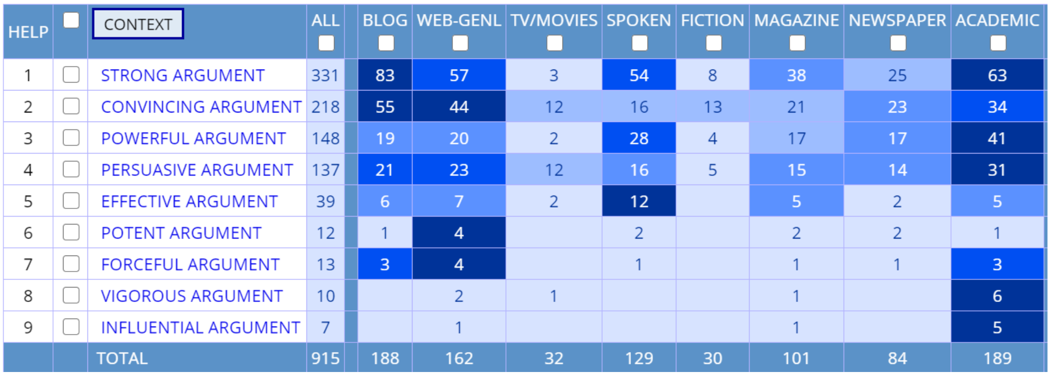
a lot of NOUN

BLOG	WEB	TV/M	SPOK	FIC	MAG	NEWS	ACAD
31758	22679	31043	82391	11017	19551	27189	3537
128.6	124.3	128.1	126.1	118.3	126.1	121.7	119.8
246.93	182.52	242.38	653.19	93.11	155.05	223.33	29.53

several NOUN

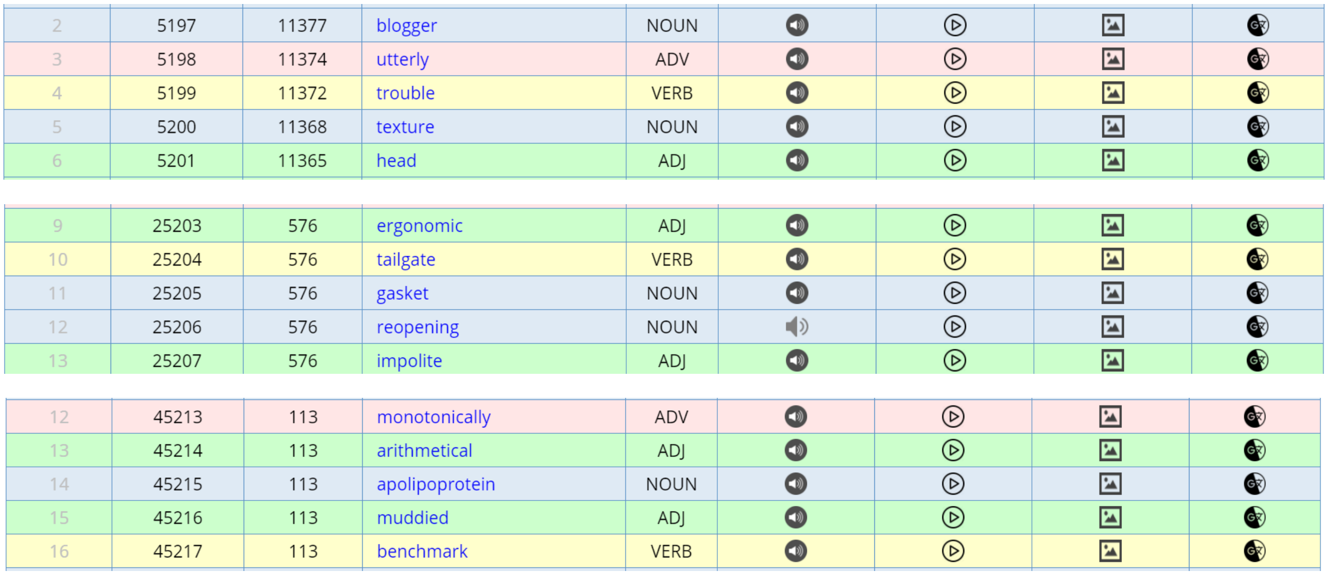
BLOG	WEB	TV/M	SPOK	FIC	MAG	NEWS	ACAD
21535	24730	4023	15315	17372	26952	29037	30919
128.6	124.3	128.1	126.1	118.3	126.1	121.7	119.8
167.44	199.03	31.41	121.42	146.82	213.75	238.51	258.11

As mentioned above, it is also very useful to see which of several **"competing" words** are the most common in a given context, such as the collocates of *powerful* before *argument*. Again, this is the type of knowledge that either comes with a thousands of hours of exposure to the second language or (alternatively) just a few seconds of searching in a corpus. And data like this can be invaluable to those writing in a second language, including researchers from a wide range of academic fields.



HELP	□	CONTEXT	ALL □	BLOG □	WEB-GENL □	TV/MOVIES □	SPOKEN □	FICTION □	MAGAZINE □	NEWSPAPER □	ACADEMIC □
1	□	STRONG ARGUMENT	331	83	57	3	54	8	38	25	63
2	□	CONVINCING ARGUMENT	218	55	44	12	16	13	21	23	34
3	□	POWERFUL ARGUMENT	148	19	20	2	28	4	17	17	41
4	□	PERSUASIVE ARGUMENT	137	21	23	12	16	5	15	14	31
5	□	EFFECTIVE ARGUMENT	39	6	7	2	12		5	2	5
6	□	POTENT ARGUMENT	12	1	4		2		2	2	1
7	□	FORCEFUL ARGUMENT	13	3	4		1		1	1	3
8	□	VIGOROUS ARGUMENT	10		2	1			1		6
9	□	INFLUENTIAL ARGUMENT	7		1				1		5
		TOTAL	915	188	162	32	129	30	101	84	189

In addition to the many types of searches shown above, there are other features of the corpora that are designed specifically for language learners, and which are definitely not available from any other large corpora. For example, in COCA and iWeb, users can **browse through** a list of the **top 60,000 words** in the corpus (these are the only large, carefully corrected frequency lists of English). The small extracts below show samples of words at three different frequency bands: near 5,000 (i.e. the 5,000th most frequent word in the corpus), 25,000, and 45,000. For each word, there is a link to a "home page" for that word (see below), audio, video, images, and translations.

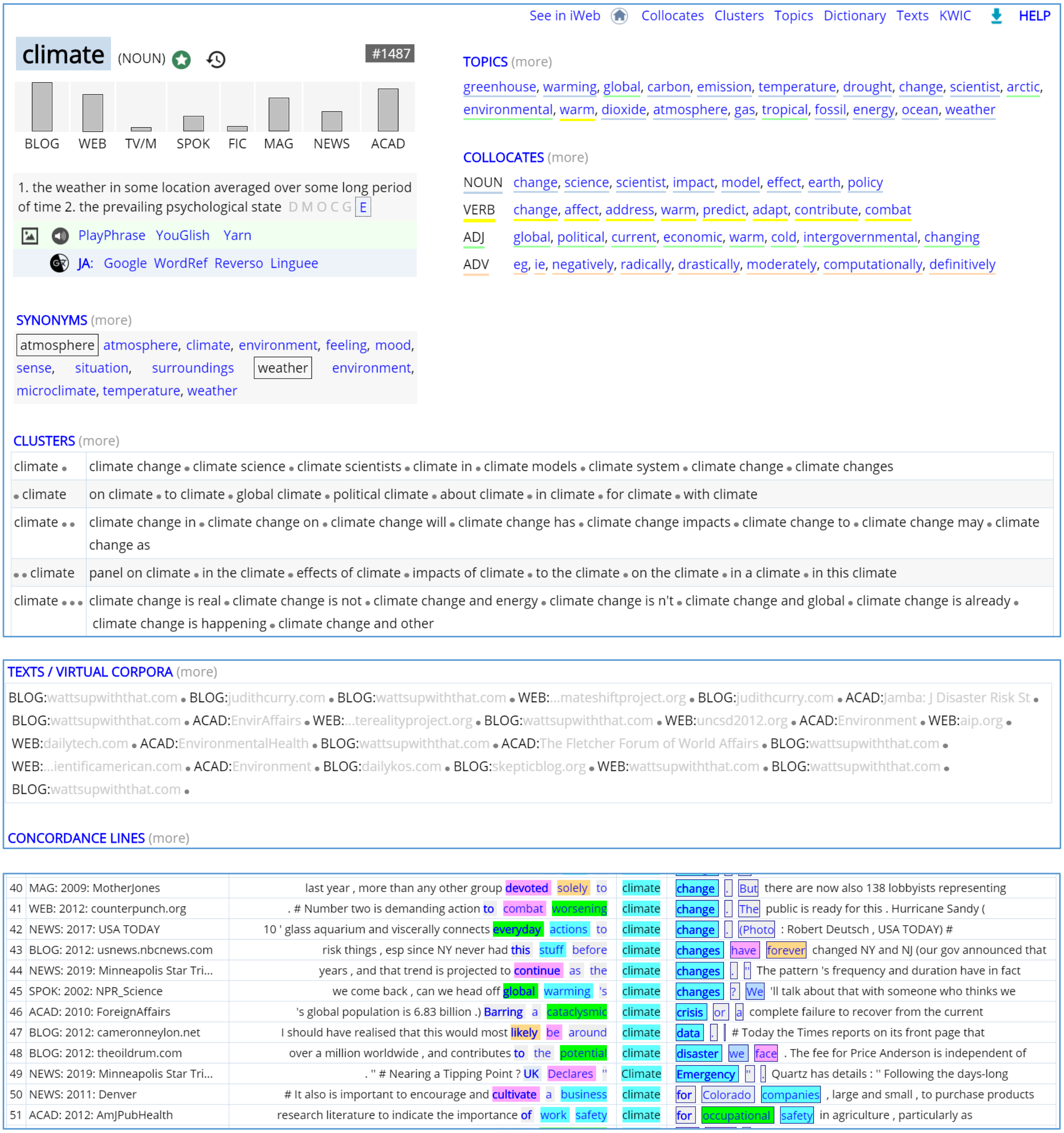


2	5197	11377	blogger	NOUN	🔊	▷	🖼	G
3	5198	11374	utterly	ADV	🔊	▷	🖼	G
4	5199	11372	trouble	VERB	🔊	▷	🖼	G
5	5200	11368	texture	NOUN	🔊	▷	🖼	G
6	5201	11365	head	ADJ	🔊	▷	🖼	G

9	25203	576	ergonomic	ADJ	🔊	▷	🖼	G
10	25204	576	tailgate	VERB	🔊	▷	🖼	G
11	25205	576	gasket	NOUN	🔊	▷	🖼	G
12	25206	576	reopening	NOUN	🔊	▷	🖼	G
13	25207	576	impolite	ADJ	🔊	▷	🖼	G

12	45213	113	monotonically	ADV	🔊	▷	🖼	G
13	45214	113	arithmetical	ADJ	🔊	▷	🖼	G
14	45215	113	apolipoprotein	NOUN	🔊	▷	🖼	G
15	45216	113	muddied	ADJ	🔊	▷	🖼	G
16	45217	113	benchmark	VERB	🔊	▷	🖼	G

For **each of the top 60,000 words** (lemmas) in the corpus, there is a "home page", which provides an **incredible wealth of information**, including: frequency, word rank (e.g. #1-60,000), frequency by genre, definitions, links to additional definitions and etymologies online, images, videos, translations (to more than 100 languages), related topics, collocates, synonyms, clusters (2, 3, and 4 word strings), texts that use the word the most, and sample concordance lines.



See in iWeb Collocates Clusters Topics Dictionary Texts KWIC HELP

climate (NOUN) #1487

BLOG WEB TV/M SPOK FIC MAG NEWS ACAD

1. the weather in some location averaged over some long period of time 2. the prevailing psychological state D M O C G E

PlayPhrase YouGlish Yarn

JA: Google WordRef Reverso Linguee

SYNONYMS (more)

atmosphere atmosphere, climate, environment, feeling, mood, sense, situation, surroundings weather environment, microclimate, temperature, weather

TOPICS (more)

greenhouse, warming, global, carbon, emission, temperature, drought, change, scientist, arctic, environmental, warm, dioxide, atmosphere, gas, tropical, fossil, energy, ocean, weather

COLLOCATES (more)

NOUN	change, science, scientist, impact, model, effect, earth, policy
VERB	change, affect, address, warm, predict, adapt, contribute, combat
ADJ	global, political, current, economic, warm, cold, intergovernmental, changing
ADV	eg, ie, negatively, radically, drastically, moderately, computationally, definitively

CLUSTERS (more)

climate •	climate change • climate science • climate scientists • climate in • climate models • climate system • climate change • climate changes
• climate	on climate • to climate • global climate • political climate • about climate • in climate • for climate • with climate
climate • •	climate change in • climate change on • climate change will • climate change has • climate change impacts • climate change to • climate change may • climate change as
• • climate	panel on climate • in the climate • effects of climate • impacts of climate • to the climate • on the climate • in a climate • in this climate
climate • • •	climate change is real • climate change is not • climate change and energy • climate change is n't • climate change and global • climate change is already • climate change is happening • climate change and other

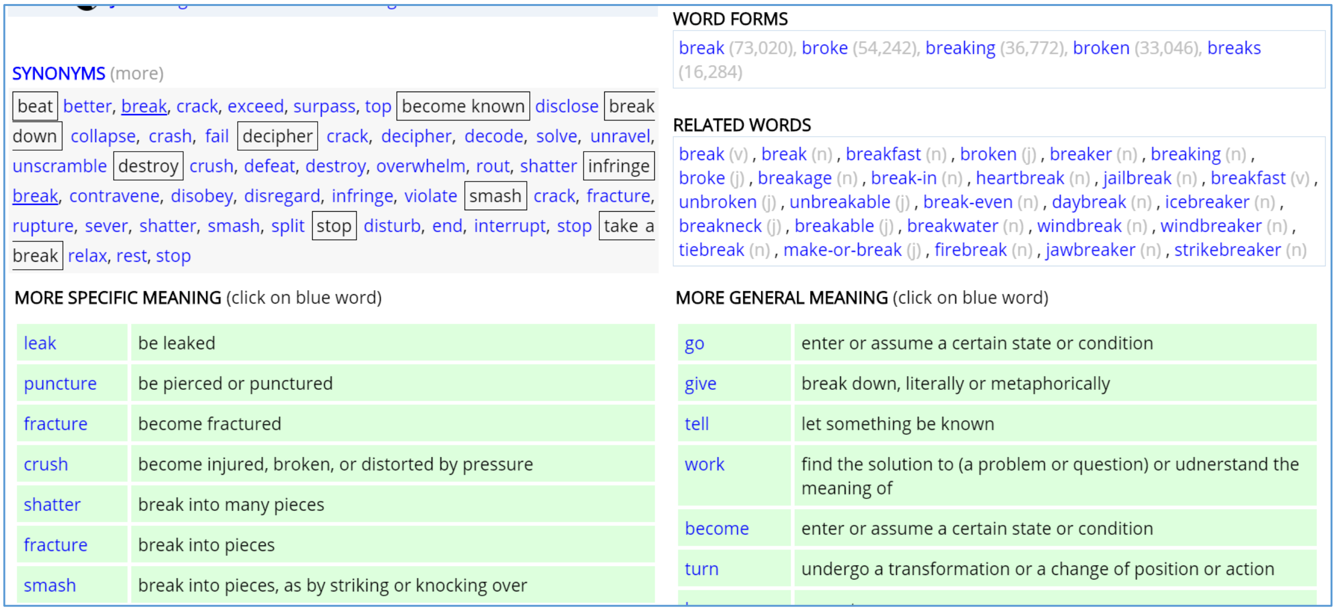
TEXTS / VIRTUAL CORPORA (more)

BLOG:wattsupwiththat.com • BLOG:judithcurry.com • BLOG:wattsupwiththat.com • WEB:...mateshiftproject.org • BLOG:judithcurry.com • ACAD:Jamba: J Disaster Risk St • BLOG:wattsupwiththat.com • ACAD:EnvirAffairs • WEB:...terealityproject.org • BLOG:wattsupwiththat.com • WEB:uncsd2012.org • ACAD:Environment • WEB:aip.org • WEB:dailytech.com • ACAD:EnvironmentalHealth • BLOG:wattsupwiththat.com • ACAD:The Fletcher Forum of World Affairs • BLOG:wattsupwiththat.com • WEB:...ientificamerican.com • ACAD:Environment • BLOG:dailykos.com • BLOG:skepticblog.org • WEB:wattsupwiththat.com • BLOG:wattsupwiththat.com • BLOG:wattsupwiththat.com •

CONCORDANCE LINES (more)

40	MAG: 2009: MotherJones	last year , more than any other group devoted solely to	climate	change . But there are now also 138 lobbyists representing
41	WEB: 2012: counterpunch.org	. # Number two is demanding action to combat worsening	climate	change . The public is ready for this . Hurricane Sandy (
42	NEWS: 2017: USA TODAY	10 ' glass aquarium and viscerally connects everyday actions to	climate	change . (Photo : Robert Deutsch , USA TODAY) #
43	BLOG: 2012: usnews.nbcnews.com	risk things , esp since NY never had this stuff before	climate	changes have forever changed NY and NJ (our gov announced that
44	NEWS: 2019: Minneapolis Star Tri...	years , and that trend is projected to continue as the	climate	changes . '' The pattern 's frequency and duration have in fact
45	SPOK: 2002: NPR_Science	we come back , can we head off global warming 's	climate	changes ? We 'll talk about that with someone who thinks we
46	ACAD: 2010: ForeignAffairs	's global population is 6.83 billion .) Barring a cataclysmic	climate	crisis or a complete failure to recover from the current
47	BLOG: 2012: cameronneylon.net	I should have realised that this would most likely be around	climate	data . # Today the Times reports on its front page that
48	BLOG: 2012: theoildrum.com	over a million worldwide , and contributes to the potential	climate	disaster we face . The fee for Price Anderson is independent of
49	NEWS: 2019: Minneapolis Star Tri...	. '' # Nearing a Tipping Point ? UK Declares ''	Climate	Emergency '' . Quartz has details : '' Following the days-long
50	NEWS: 2011: Denver	# It also is important to encourage and cultivate a business	climate	for Colorado companies , large and small , to purchase products
51	ACAD: 2012: AmJPubHealth	research literature to indicate the importance of work safety	climate	for occupational safety in agriculture , particularly as

All of the sections on the "home page" are just overviews, and users can click on almost any section for **even more information**. For example, the "dictionary" page for *break* as a verb (one of seven pages for this word that are available in COCA or iWeb) shows synonyms, frequency of word forms, related words, and more specific and more general words. Users can click on any word on the page to go to the "home page" for that word. In other words, all of the words are connected, which allows users to follow a "semantic trail" through related words.



SYNONYMS (more)

beat better, break, crack, exceed, surpass, top become known disclose break down collapse, crash, fail decipher crack, decipher, decode, solve, unravel, unscramble destroy crush, defeat, destroy, overwhelm, rout, shatter infringe break, contravene, disobey, disregard, infringe, violate smash crack, fracture, rupture, sever, shatter, smash, split stop disturb, end, interrupt, stop take a break relax, rest, stop

WORD FORMS

break (73,020), broke (54,242), breaking (36,772), broken (33,046), breaks (16,284)

RELATED WORDS

break (v) , break (n) , breakfast (n) , broken (j) , breaker (n) , breaking (n) , broke (j) , breakage (n) , break-in (n) , heartbreak (n) , jailbreak (n) , breakfast (v) , unbroken (j) , unbreakable (j) , break-even (n) , daybreak (n) , icebreaker (n) , breakneck (j) , breakable (j) , breakwater (n) , windbreak (n) , windbreaker (n) , tiebreak (n) , make-or-break (j) , firebreak (n) , jawbreaker (n) , strikebreaker (n)

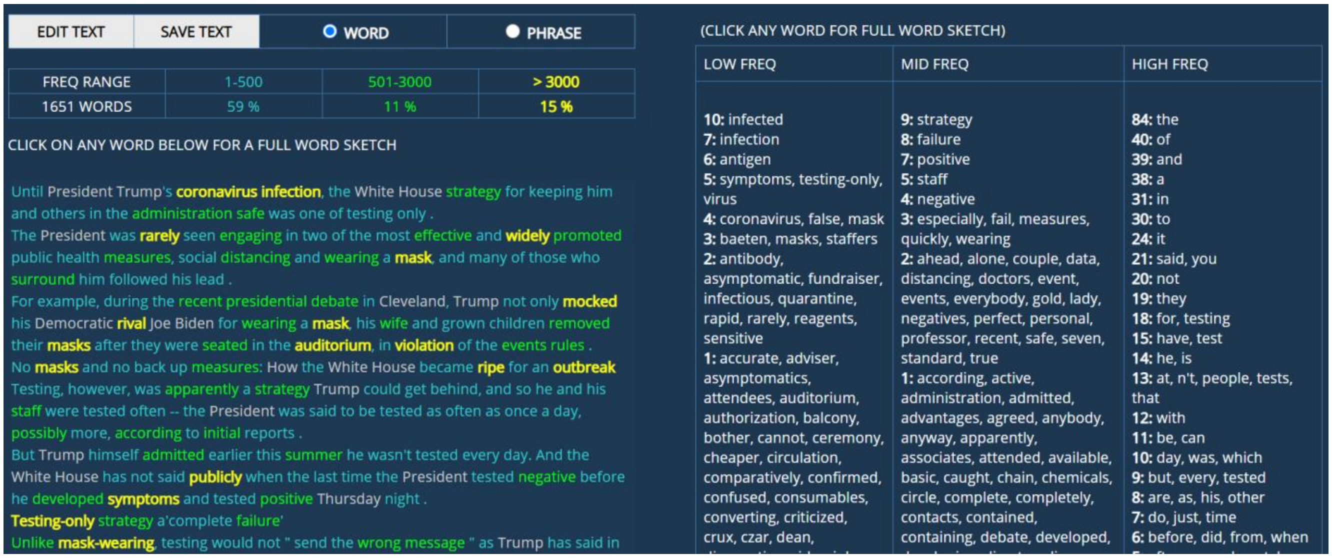
MORE SPECIFIC MEANING (click on blue word)

leak	be leaked
puncture	be pierced or punctured
fracture	become fractured
crush	become injured, broken, or distorted by pressure
shatter	break into many pieces
fracture	break into pieces
smash	break into pieces, as by striking or knocking over

MORE GENERAL MEANING (click on blue word)

go	enter or assume a certain state or condition
give	break down, literally or metaphorically
tell	let something be known
work	find the solution to (a problem or question) or udnerstand the meaning of
become	enter or assume a certain state or condition
turn	undergo a transformation or a change of position or action

Finally, the "**analyze text**" functionality in COCA provides many features that are very useful to language learners and teachers. Users can enter entire texts (e.g. compositions that they have written, or articles from online newspapers or magazines). The corpus then **highlights words** in the text that are less frequent generally in English (and which are words that the learner might not know), and it shows the percentage of words in different frequency bands of English. It also shows the specific words in each of these frequency bands, ordered by frequency, which provide good information on the **keywords in the text**. So for example, in the following article from CNN (dealing with identifying carriers of COVID-19), some of the top keywords are *infected, infection, antigen, symptoms*, and *virus*.



FREQ RANGE	1-500	501-3000	> 3000
1651 WORDS	59 %	11 %	15 %

CLICK ON ANY WORD BELOW FOR A FULL WORD SKETCH

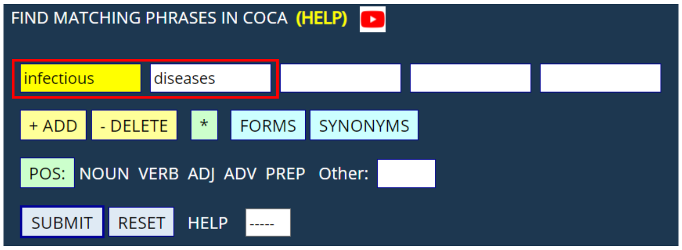
Until President Trump's coronavirus infection, the White House strategy for keeping him and others in the administration safe was one of testing only .
The President was rarely seen engaging in two of the most effective and widely promoted public health measures, social distancing and wearing a mask, and many of those who surround him followed his lead .
For example, during the recent presidential debate in Cleveland, Trump not only mocked his Democratic rival Joe Biden for wearing a mask, his wife and grown children removed their masks after they were seated in the auditorium, in violation of the events rules .
No masks and no back up measures: How the White House became ripe for an outbreak
Testing, however, was apparently a strategy Trump could get behind, and so he and his staff were tested often -- the President was said to be tested as often as once a day, possibly more, according to initial reports .
But Trump himself admitted earlier this summer he wasn't tested every day. And the White House has not said publicly when the last time the President tested negative before he developed symptoms and tested positive Thursday night .
Testing-only strategy a'complete failure'
Unlike mask-wearing, testing would not " send the wrong message " as Trump has said in

(CLICK ANY WORD FOR FULL WORD SKETCH)

LOW FREQ	MID FREQ	HIGH FREQ
10: infected **7:** infection **6:** antigen **5:** symptoms, testing-only, virus **4:** coronavirus, false, mask **3:** baeten, masks, staffers **2:** antibody, asymptomatic, fundraiser, infectious, quarantine, rapid, rarely, reagents, sensitive **1:** accurate, adviser, asymptomatics, attendees, auditorium, authorization, balcony, bother, cannot, ceremony, cheaper, circulation, comparatively, confirmed, confused, consumables, converting, criticized, crux, czar, dean,	**9:** strategy **8:** failure **7:** positive **5:** staff **4:** negative **3:** especially, fail, measures, quickly, wearing **2:** ahead, alone, couple, data, distancing, doctors, event, events, everybody, gold, lady, negatives, perfect, personal, professor, recent, safe, seven, standard, true **1:** according, active, administration, admitted, advantages, agreed, anybody, anyway, apparently, associates, attended, available, basic, caught, chain, chemicals, circle, complete, completely, contacts, contained, containing, debate, developed,	**84:** the **40:** of **39:** and **38:** a **31:** in **30:** to **24:** it **21:** said, you **20:** not **19:** they **18:** for, testing **15:** have, test **14:** he, is **13:** at, n't, people, tests, that **12:** with **11:** be, can **10:** day, was, which **9:** but, every, tested **8:** are, as, his, other **7:** do, just, time **6:** before, did, from, when

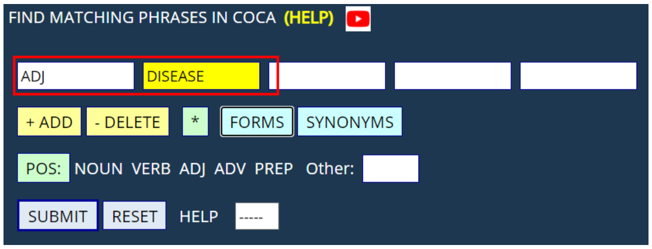
Users can then click on any word in the text, or any of the words in the frequency lists from the text, to see the full entry on that word, as was discussed above. This ability to easily **browse through unfamiliar words** and then to see detailed information on any of the words is completely unique to COCA.

Finally, users can click on any words in the text to form phrases, and then **quickly and easily find related phrases in COCA**. For example, the phrase *infectious diseases* occurs in this text. Users can click on these two words (below, left) and then click on POS (Part of Speech) to show that they want any adjective instead of *infectious*, and then FORMS to find any form of *diseases* (right).



FIND MATCHING PHRASES IN COCA (HELP)

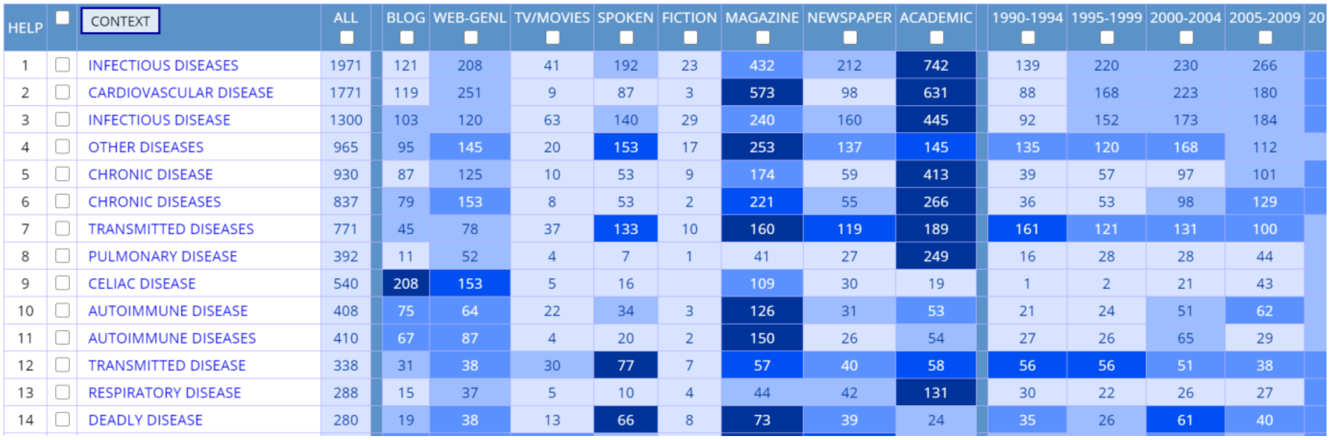
infectious | diseases



FIND MATCHING PHRASES IN COCA (HELP)

ADJ | DISEASE

After clicking on SUBMIT, they can see the matching phrases in COCA, ordered by frequency in the different genres.

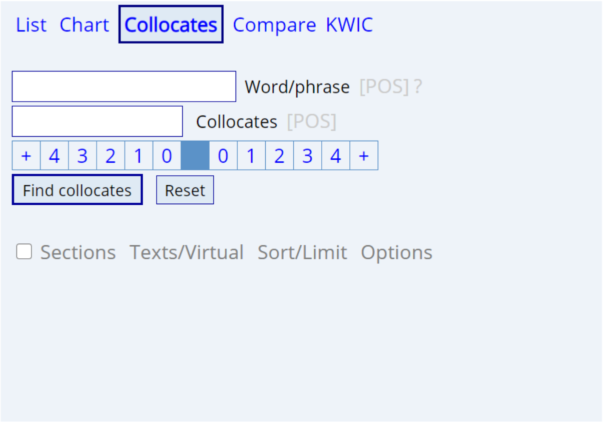


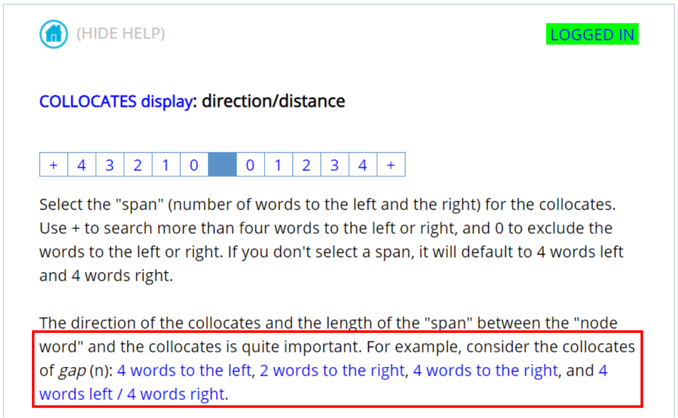
HELP	☐	CONTEXT	ALL	BLOG	WEB-GENL	TV/MOVIES	SPOKEN	FICTION	MAGAZINE	NEWSPAPER	ACADEMIC	1990-1994	1995-1999	2000-2004	2005-2009	20
			☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	
1	☐	INFECTIOUS DISEASES	1971	121	208	41	192	23	432	212	742	139	220	230	266	
2	☐	CARDIOVASCULAR DISEASE	1771	119	251	9	87	3	573	98	631	88	168	223	180	
3	☐	INFECTIOUS DISEASE	1300	103	120	63	140	29	240	160	445	92	152	173	184	
4	☐	OTHER DISEASES	965	95	145	20	153	17	253	137	145	135	120	168	112	
5	☐	CHRONIC DISEASE	930	87	125	10	53	9	174	59	413	39	57	97	101	
6	☐	CHRONIC DISEASES	837	79	153	8	53	2	221	55	266	36	53	98	129	
7	☐	TRANSMITTED DISEASES	771	45	78	37	133	10	160	119	189	161	121	131	100	
8	☐	PULMONARY DISEASE	392	11	52	4	7	1	41	27	249	16	28	28	44	
9	☐	CELIAC DISEASE	540	208	153	5	16		109	30	19	1	2	21	43	
10	☐	AUTOIMMUNE DISEASE	408	75	64	22	34	3	126	31	53	21	24	51	62	
11	☐	AUTOIMMUNE DISEASES	410	67	87	4	20	2	150	26	54	27	26	65	29	
12	☐	TRANSMITTED DISEASE	338	31	38	30	77	7	57	40	58	56	56	51	38	
13	☐	RESPIRATORY DISEASE	288	15	37	5	10	4	44	42	131	30	22	26	27	
14	☐	DEADLY DISEASE	280	19	38	13	66	8	73	39	24	35	26	61	40	

The ability to “click and see” many related phrases might be particularly useful for teaching writing, or for non-native researchers writing in English. They can **click on any of the phrases in their composition**, for example, and see the frequency across genres (e.g. is it a formal or informal phrase), and quickly and easily find related phrases that might be even better (such as with phrases related to *powerful argument*, shown above).

Other tools and features (go to beginning)

As is shown above, users can do a wide range of queries. Especially at the beginning, however, this can sometimes be overwhelming. Fortunately, every page has a wide range of **“context sensitive” help files** that guide users through the options (e.g. of [Collocates] below). Most of these context-sensitive help files also have sample searches that users can click on, and thus interact with the corpus even more.



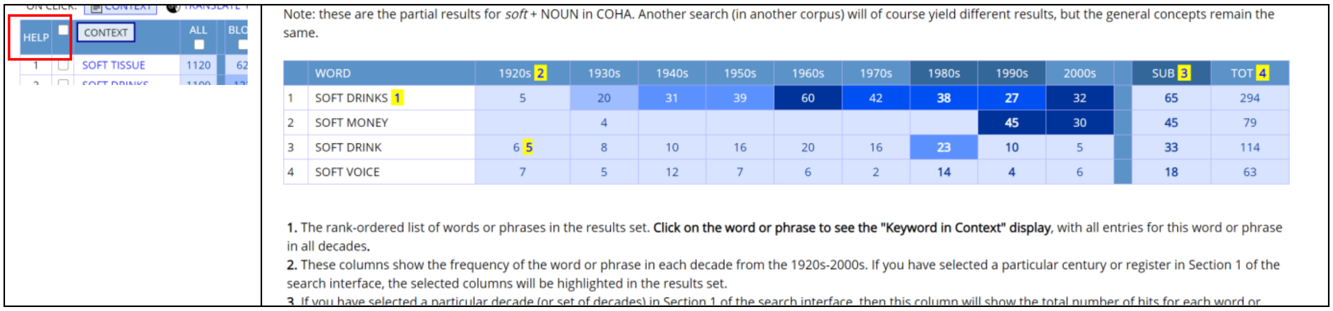


COLLOCATES display: direction/distance

Select the "span" (number of words to the left and the right) for the collocates. Use + to search more than four words to the left or right, and 0 to exclude the words to the left or right. If you don't select a span, it will default to 4 words left and 4 words right.

The direction of the collocates and the length of the "span" between the "node word" and the collocates is quite important. For example, consider the collocates of *gap* (n): 4 words to the left, 2 words to the right, 4 words to the right, and 4 words left / 4 words right.

In addition, each of the “results” pages has a [HELP] link, which helps users to understand what the data means:



HELP	☐	CONTEXT	ALL	BLO
1	☐	SOFT TISSUE	1120	62

Note: these are the partial results for *soft* + NOUN in COHA. Another search (in another corpus) will of course yield different results, but the general concepts remain the same.

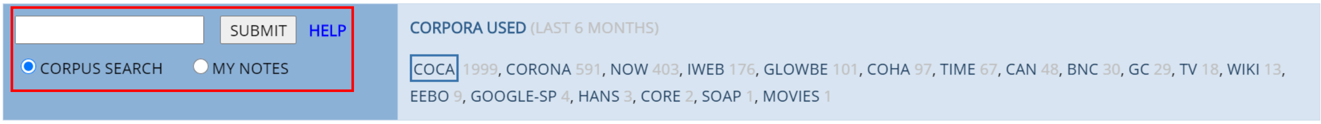
	WORD	1920s 2	1930s	1940s	1950s	1960s	1970s	1980s	1990s	2000s	SUB 3	TOT 4
1	SOFT DRINKS 1	5	20	31	39	60	42	38	27	32	65	294
2	SOFT MONEY		4						45	30	45	79
3	SOFT DRINK	6 5	8	10	16	20	16	23	10	5	33	114
4	SOFT VOICE	7	5	12	7	6	2	14	4	6	18	63

1. The rank-ordered list of words or phrases in the results set. **Click on the word or phrase to see the "Keyword in Context" display**, with all entries for this word or phrase in all decades.

2. These columns show the frequency of the word or phrase in each decade from the 1920s-2000s. If you have selected a particular century or register in Section 1 of the search interface, the selected columns will be highlighted in the results set.

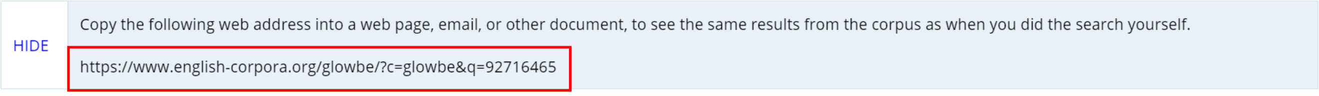
3. If you have selected a particular decade (or set of decades) in Section 1 of the search interface, then this column will show the total number of hits for each word or

Users can see a **"history" of their searches**, and can even find past searches that contain specific words or phrases. They can then **copy links** to their searches and embed them in research papers or web pages, so that other people will see exactly what the user saw when s/he originally did the search (and thus help make the findings from the corpora "replicable").



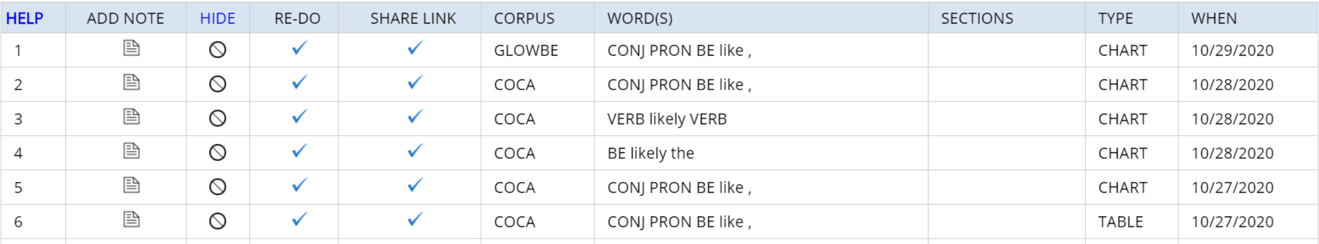
CORPORA USED (LAST 6 MONTHS)

COCA 1999, CORONA 591, NOW 403, IWEB 176, GLOWBE 101, COHA 97, TIME 67, CAN 48, BNC 30, GC 29, TV 18, WIKI 13, EEBO 9, GOOGLE-SP 4, HANS 3, CORE 2, SOAP 1, MOVIES 1



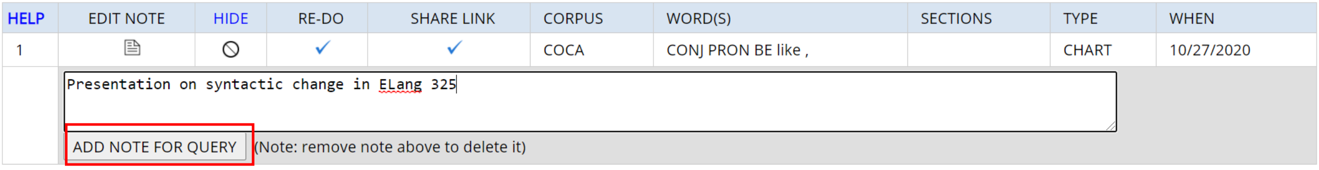
Copy the following web address into a web page, email, or other document, to see the same results from the corpus as when you did the search yourself.

https://www.english-corpora.org/glowbe/?c=glowbe&q=92716465



HELP	ADD NOTE	HIDE	RE-DO	SHARE LINK	CORPUS	WORD(S)	SECTIONS	TYPE	WHEN
1	🗎	⊘	✓	✓	GLOWBE	CONJ PRON BE like ,		CHART	10/29/2020
2	🗎	⊘	✓	✓	COCA	CONJ PRON BE like ,		CHART	10/28/2020
3	🗎	⊘	✓	✓	COCA	VERB likely VERB		CHART	10/28/2020
4	🗎	⊘	✓	✓	COCA	BE likely the		CHART	10/28/2020
5	🗎	⊘	✓	✓	COCA	CONJ PRON BE like ,		CHART	10/27/2020
6	🗎	⊘	✓	✓	COCA	CONJ PRON BE like ,		TABLE	10/27/2020

They can also "annotate" their searches by adding notes or comments, and then search through these annotations for all matching queries (e.g. all searches for a particular class lecture, or for a paper they are writing).

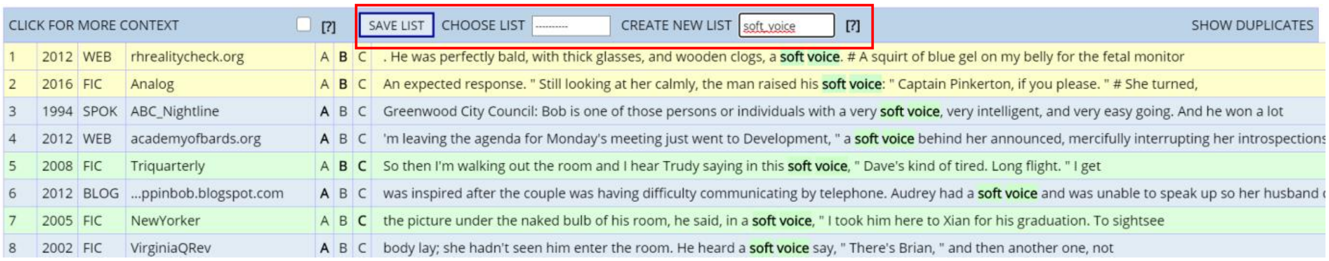


HELP	EDIT NOTE	HIDE	RE-DO	SHARE LINK	CORPUS	WORD(S)	SECTIONS	TYPE	WHEN
1	🗎	⊘	✓	✓	COCA	CONJ PRON BE like ,		CHART	10/27/2020

Presentation on syntactic change in ELang 325

(Note: remove note above to delete it)

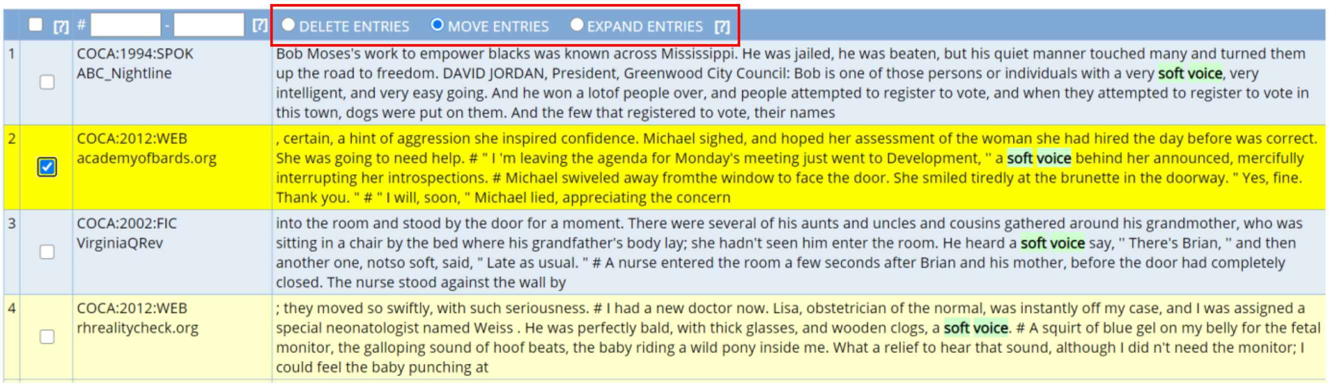
Users can also **save concordance lines** from a search, and categorize the lines into different groups (note the three different colors below):



soft voice

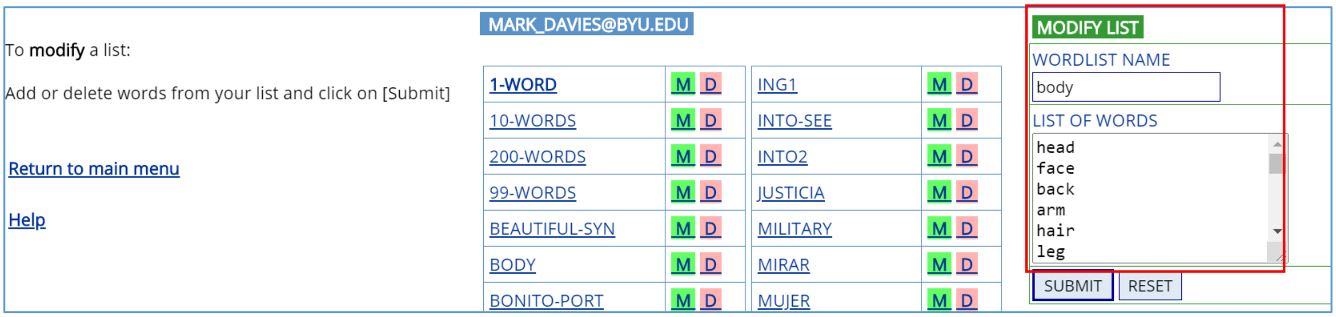
1	2012	WEB	rhrealitycheck.org	A	B	C	. He was perfectly bald, with thick glasses, and wooden clogs, a **soft voice**. # A squirt of blue gel on my belly for the fetal monitor
2	2016	FIC	Analog	A	B	C	An expected response. " Still looking at her calmly, the man raised his **soft voice**: " Captain Pinkerton, if you please. " # She turned,
3	1994	SPOK	ABC_Nightline	A	B	C	Greenwood City Council: Bob is one of those persons or individuals with a very **soft voice**, very intelligent, and very easy going. And he won a lot
4	2012	WEB	academyofbards.org	A	B	C	'm leaving the agenda for Monday's meeting just went to Development, " a **soft voice** behind her announced, mercifully interrupting her introspections
5	2008	FIC	Triquarterly	A	B	C	So then I'm walking out the room and I hear Trudy saying in this **soft voice**, " Dave's kind of tired. Long flight. " I get
6	2012	BLOG	...ppinbob.blogspot.com	A	B	C	was inspired after the couple was having difficulty communicating by telephone. Audrey had a **soft voice** and was unable to speak up so her husband
7	2005	FIC	NewYorker	A	B	C	the picture under the naked bulb of his room, he said, in a **soft voice**, " I took him here to Xian for his graduation. To sightsee
8	2002	FIC	VirginiaQRev	A	B	C	body lay; she hadn't seen him enter the room. He heard a **soft voice** say, " There's Brian, " and then another one, not

Later, they can expand, delete, and move these lines:



□ [?] # ______ - ______ [?]			○ DELETE ENTRIES ● MOVE ENTRIES ○ EXPAND ENTRIES [?]
1	□	COCA:1994:SPOK ABC_Nightline	Bob Moses's work to empower blacks was known across Mississippi. He was jailed, he was beaten, but his quiet manner touched many and turned them up the road to freedom. DAVID JORDAN, President, Greenwood City Council: Bob is one of those persons or individuals with a very soft voice, very intelligent, and very easy going. And he won a lotof people over, and people attempted to register to vote, and when they attempted to register to vote in this town, dogs were put on them. And the few that registered to vote, their names
2	☑	COCA:2012:WEB academyofbards.org	, certain, a hint of aggression she inspired confidence. Michael sighed, and hoped her assessment of the woman she had hired the day before was correct. She was going to need help. # " I 'm leaving the agenda for Monday's meeting just went to Development, " a soft voice behind her announced, mercifully interrupting her introspections. # Michael swiveled away fromthe window to face the door. She smiled tiredly at the brunette in the doorway. " Yes, fine. Thank you. " # " I will, soon, " Michael lied, appreciating the concern
3	□	COCA:2002:FIC VirginiaQRev	into the room and stood by the door for a moment. There were several of his aunts and uncles and cousins gathered around his grandmother, who was sitting in a chair by the bed where his grandfather's body lay; she hadn't seen him enter the room. He heard a soft voice say, " There's Brian, " and then another one, notso soft, said, " Late as usual. " # A nurse entered the room a few seconds after Brian and his mother, before the door had completely closed. The nurse stood against the wall by
4	□	COCA:2012:WEB rhrealitycheck.org	; they moved so swiftly, with such seriousness. # I had a new doctor now. Lisa, obstetrician of the normal, was instantly off my case, and I was assigned a special neonatologist named Weiss . He was perfectly bald, with thick glasses, and wooden clogs, a soft voice. # A squirt of blue gel on my belly for the fetal monitor, the galloping sound of hoof beats, the baby riding a wild pony inside me. What a relief to hear that sound, although I did n't need the monitor; I could feel the baby punching at

Users can create "**customized wordlists**" for any set of words that they want to use in a search, such as words relating to the body, or to emotions, or a certain class of verbs:



To **modify** a list:

Add or delete words from your list and click on [Submit]

Return to main menu

Help

MARK_DAVIES@BYU.EDU

1-WORD	M D	ING1	M D
10-WORDS	M D	INTO-SEE	M D
200-WORDS	M D	INTO2	M D
99-WORDS	M D	JUSTICIA	M D
BEAUTIFUL-SYN	M D	MILITARY	M D
BODY	M D	MIRAR	M D
BONITO-PORT	M D	MUJER	M D

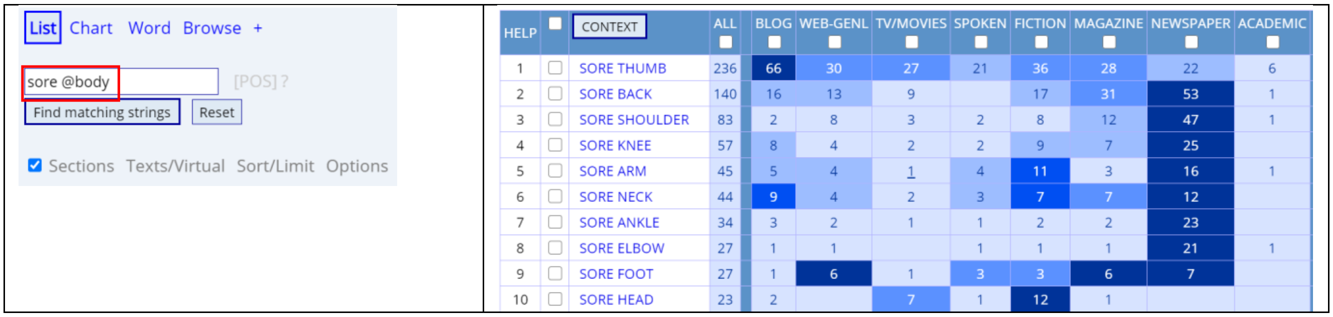
MODIFY LIST

WORDLIST NAME: body

LIST OF WORDS:

```
head
face
back
arm
hair
leg
```

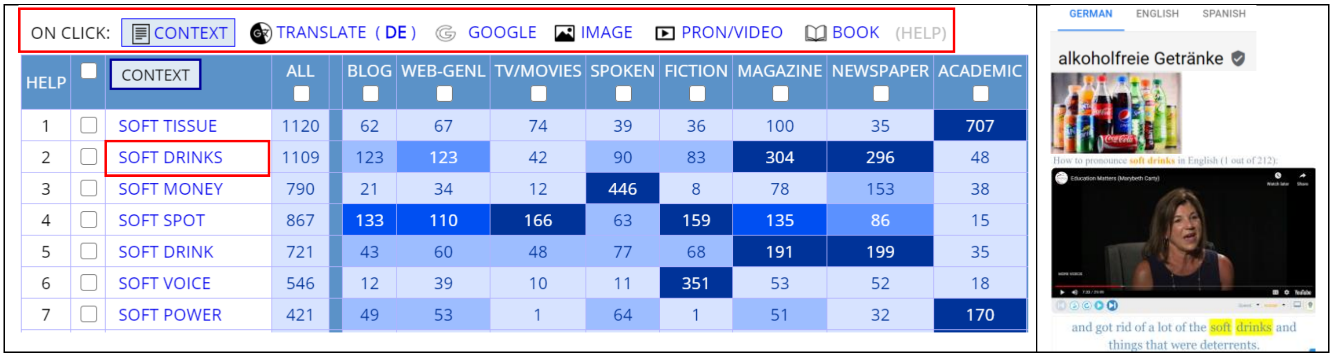
They can then use these words directly as part of any search, and thus **search the corpus "semantically"**:



sore @body

HELP	□	CONTEXT	ALL	BLOG	WEB-GENL	TV/MOVIES	SPOKEN	FICTION	MAGAZINE	NEWSPAPER	ACADEMIC
1	□	SORE THUMB	236	66	30	27	21	36	28	22	6
2	□	SORE BACK	140	16	13	9		17	31	53	1
3	□	SORE SHOULDER	83	2	8	3	2	8	12	47	1
4	□	SORE KNEE	57	8	4	2	2	9	7	25	
5	□	SORE ARM	45	5	4	1	4	11	3	16	1
6	□	SORE NECK	44	9	4	2	3	7	7	12	
7	□	SORE ANKLE	34	3	2	1	1	2	2	23	
8	□	SORE ELBOW	27	1	1		1	1	1	21	1
9	□	SORE FOOT	27	1	6	1	3	3	6	7	
10	□	SORE HEAD	23	2		7	1	12	1		

In the "results" page of any search, there are **links to a wide range of external resources**, such as translations (to more than 100 languages), Google searches for web, images, and books; and pronunciation and videos.



ON CLICK: CONTEXT TRANSLATE (DE) GOOGLE IMAGE PRON/VIDEO BOOK (HELP)
GERMAN ENGLISH SPANISH
Education Matters (Marybeth Carty)

HELP	□	CONTEXT	ALL	BLOG	WEB-GENL	TV/MOVIES	SPOKEN	FICTION	MAGAZINE	NEWSPAPER	ACADEMIC
1	□	SOFT TISSUE	1120	62	67	74	39	36	100	35	707
2	□	SOFT DRINKS	1109	123	123	42	90	83	304	296	48
3	□	SOFT MONEY	790	21	34	12	446	8	78	153	38
4	□	SOFT SPOT	867	133	110	166	63	159	135	86	15
5	□	SOFT DRINK	721	43	60	48	77	68	191	199	35
6	□	SOFT VOICE	546	12	39	10	11	351	53	52	18
7	□	SOFT POWER	421	49	53	1	64	1	51	32	170

alkoholfreie Getränke

How to pronounce soft drinks in English (1 out of 212):

and got rid of a lot of the soft drinks and things that were deterrents.

Finally, researchers can download for offline use a wide range of data that is based on the online corpora, such as full text data (www.corpusdata.org), word frequency data (www.wordfrequency.info), collocates (www.collocates.info), and n-grams (www.ngrams.info).

Summary

The corpora from English-Corpora.org are the **most widely used corpora in the world**, and they are used by 130,000+ distinct researchers, teachers, and learners each month. The corpora are used as the basis for thousands of research **articles** each year, as well as being an integral part of **classrooms** throughout the world.

The corpora allow researchers to look at **variation in English** (e.g. genre-based, historical, and dialectal variation) in ways that are not even remotely possible with any other collection of corpora. They allow researchers in fields like history, cultural studies, and legal studies to look at **societal and cultural issues** through the lens of huge collections of texts. They provide **non-native researchers (in academic fields)** with tools to analyze their English in ways that standard dictionaries and thesauruses never could. And they offer a wealth of possibilities in terms of **language learning and teaching** that are completely and totally unique to these corpora.