

A **corpus** (plural: *corpora*) is simply a **highly structured collection of texts**, which allows **researchers** to carry out extremely sophisticated searches to see what is happening in the language (for example between **genres, dialects, and over time**) in ways that would never be possible with other simple search engines like Google. Corpora also allow **learners and teachers** to easily find a wide range of data on words, phrases, and grammatical constructions – far beyond what would be found in a textbook or dictionary.

Corpus	Overview	Download	# words	Dialect	Time period	Genre(s)
News on the Web (NOW)			18.1 billion+	20 countries	2010-yesterday	Web: News
iWeb: The Intelligent Web-based Corpus			14 billion	6 countries	2017	Web
Global Web-Based English (GloWbE)			1.9 billion	20 countries	2012-13	Web (incl blogs)
Wikipedia Corpus			1.9 billion	(Various)	2014	Wikipedia
Coronavirus Corpus			1.5 billion	20 countries	Jan 2020-Dec 2022	Web: News
Corpus of Contemporary American English (COCA)			1.0 billion	American	1990-2019	Balanced
Corpus of Historical American English (COHA)			475 million	American	1820-2019	Balanced
The TV Corpus			325 million	6 countries	1950-2018	TV shows
The Movie Corpus			200 million	6 countries	1930-2018	Movies
Corpus of American Soap Operas			100 million	American	2001-2012	TV shows
<hr/>						
Hansard Corpus			1.6 billion	British	1803-2005	Parliament
Early English Books Online			755 million	British	1470s-1690s	(Various)
Corpus of US Supreme Court Opinions			130 million	American	1790s-present	Legal opinions
TIME Magazine Corpus			100 million	American	1923-2006	Magazine
British National Corpus (BNC) *			100 million	British	1980s-1993	Balanced
Strathy Corpus (Canada)			50 million	Canadian	1920s-2000s	Balanced
CORE Corpus			50 million	6 countries	2014	Web

The corpora from [English-Corpora.org](https://www.english-corpora.org) are **used more than any other corpora** – with more than 74,000 users each month. Limited, basic access is free, but **hundreds of universities** have purchased **academic licenses** for expanded access, especially by classes. The corpora are used by:

- Tens of thousands of **researchers** from universities throughout the world, for **thousands of publications**
- Hundreds of thousands of **learners and teachers**
- Many **companies**, especially in the fields of technology (e.g. Google, Microsoft, Amazon, IBM, Adobe, Intel, Samsung), as well as language teaching (e.g. Duolingo, Grammarly, Merriam-Webster, Sketch Engine, Oxford University Press)

Teaching and learning

The corpora from English-Corpora.org (especially the one billion word Corpus of Contemporary American English; **COCA**) are the focus of **almost every book on corpora** and language teaching in the last 10-15 years (examples: [1](#) [2](#) [3](#) [4](#) [5](#)).

Teachers and learners can **search for words** by word form, part of speech, frequency (1-60,000), meaning (for example, words in a definition), synonyms, more specific or more general words, and even pronunciation.

Word form:

Meaning: + DEFINITION SYNONYM SPECIFIC GENERAL

You can now search for words by meaning. For example, words with the following words in the definition: sugar, [] for the dictionary entry, e.g. herb OR herbs (herb* would include the perhaps unwanted herbivore as well), comp search by synonym (noun: festival, disaster; adjective: harsh, kind; verb: groan, laugh), find more specific words (noun: frisbee, tequila; shriek, sashay) (both for just nouns/verbs), or combine these (e.g. walk, scare, screen, crystal

Part of speech: NOUN VERB ADJ ADV OTHER ALL

Range: -

Pronunciation: Rhymes with Type: EXACT

Syllables / stress:

	RANK	FREQ	Word	PoS	Audio	Video	Image
★	2448	33519	cake	NOUN			
1. a block of solid substance (such as soap or wax) 2. made from or based on a mixture of flour and sugar and eggs 3. small flat mass of chopped food							
★	2819	28147	chocolate	NOUN			
1. made from baking chocolate or cocoa powder and milk and sugar 2. a medium to dark brown color 3. made from roasted ground cacao beans							
★	3370	21987	candy	NOUN			
1. a rich sweet made of flavored sugar and often combined with fruit or nuts							
★	6323	8248	cane	NOUN			
1. a stick that people can lean on to help them walk 2. a stiff switch used to hit students as punishment 3. a strong slender often flexible stem as of bamr rattans , or sugar cane							
★	9361	4193	glucose	NOUN			
1. a monosaccharide sugar that has several forms							
★	9778	3858	cocoa	NOUN			
1. powder of ground roasted cacao beans with most of the fat removed 2. made from baking chocolate or cocoa powder and milk and sugar							
★	9824	3818	sweet	NOUN			
1. the property of containing sugar 2. the taste experience when sugar dissolves in the mouth 3. a food rich in sugar 4. (British) dessert							
★	10353	3483	sweetness	NOUN			
1. the taste experience when sugar dissolves in the mouth 2. the quality of giving pleasure 3. a pleasingly sweet olfactory property 4. the property of co							

immunity (NOUN) #5702

BLOG WEB TV/M SPOK FIC MAG NEWS ACAD

1. the state of not being susceptible 2. an act exempting someone 3. the quality of being unaffected by something
D M O C G E

YouGlish PlayPhrase Yarn

ES: Google WordRef Reverso Linguee

SYNONYMS (▶ CONCEPT) **NEW:** DEFIN +SPEC +GENL
 [exemption] exemption, freedom, immunity, liberation, liberty
 [invulnerability] freedom, immunity, invulnerability, protection
 [protection] invulnerability, protection, resistance

CLUSTERS (more)
 immunity • immunity from • immunity to • immunity for • immunity in • immunity idol • immunity challenge • immunity against • immunities clause
 • immunity herd immunity • sovereign immunity • diplomatic immunity • for immunity • have immunity • wins immunity • granted immunity • hidden
 immunity • immunity from prosecution • immunities of citizens • immunity in exchange • immunity is back • immunity to it • immunity and reward • ir

TOPICS (more)
 immune, infection, disease, vaccine, infect, virus, prosecute, outbreak, induce proceedings, allege, eg, antibody, epidemic, tribe, flu, prosecution, suppress, !

COLLOCATES (more)
 NOUN privilege, herd, challenge, idol, prosecution, doctrine, tax, vaccine
 VERB grant, win, enjoy, boost, develop, protect, acquire, entitle
 ADJ sovereign, diplomatic, natural, absolute, qualified, hidden, intergovver
 ADV naturally, expressly, ie, vivo, facially, constitutionally, qualitatively, une

RELATED WORDS
 immune, immunization, immunize

And then they can see **detailed “word sketches”** for each of the top 60,000 words in English, including definition, frequency by genre (for example, academic or spoken), synonyms, more specific and more general words, collocates (nearby words), related topics (which appear anywhere in the text), clusters (2, 3, 4 word strings), concordance lines, and links to external resources like dictionary entries, pronunciation, images, videos, and translations to 100+ languages.

EDIT TEXT SAVE TEXT WORD PHRASE (CLICK ANY WORD FOR FULL WORD SKETCH)

FREQ RANGE	1-500	501-3000	> 3000
1699 WORDS	53 %	10 %	23 %

CLICK ON ANY WORD BELOW FOR A FULL WORD SKETCH

This time last year, the **brand** new, **stunningly effective** Covid-19 **vaccines** were rolling out across the country. **Injecting** a strong note of **optimism** into the United States' once **fumbling pandemic** response. Millions of people were lining up **daily** to get their **shots**. **Instead** of the **steady drumbeat** of cases, **hospitalizations** and deaths, we were **tracking** a new number: the percentage of Americans who had been **vaccinated**. This number, we believed, was our best **chance** to **beat** the **virus**. HINES, ILLINOIS - APRIL 01: Army **veteran** Robert Hall waits the **recommended** 15 minutes to see if he will have any **adverse** reactions after receiving his second COVID-19 **booster shot** at Edward Hines Jr. VA Hospital on April 01, 2022 in Hines, Illinois. **Earlier** this week the CDC **updated** its **recommendations** to **encourage** a second COVID-19 **booster shot** for certain **immunocompromised** individuals and people over the age of 50 who received an **initial booster dose** at least 4 months ago. Should you get your second **booster shot** now? The US was **caught up** in a **fever** **dream** of reaching **herd immunity**, a **threshold** we might **cross** where **vulnerable** individuals - including those too young to be **vaccinated** or those who didn't **respond** well

21: immunity
19: virus
15: vaccines
12: herd
10: vaccinated, vaccine
9: measles
8: contagious, transmission
6: vaccination
5: booster
4: infection, viruses
3: spray, sterilizing, variants
2: antibodies, coronavirus, dose, durability, induced, infected, lifelong, mutate, nasal, variant, virtually, vulnerable
1: acceptance, adverse, advertises, advisers, ambitious, arise, assist, asymptotically, brand, childhood, circulate, circulating,

6: prevent
5: disease, s
4: effective
3: according
2: beat, cert
1: account, s
article, bey
campaign, c
change, ch
concept, co
coverage, cr
defenses, d
depends, de
develop, de
differences

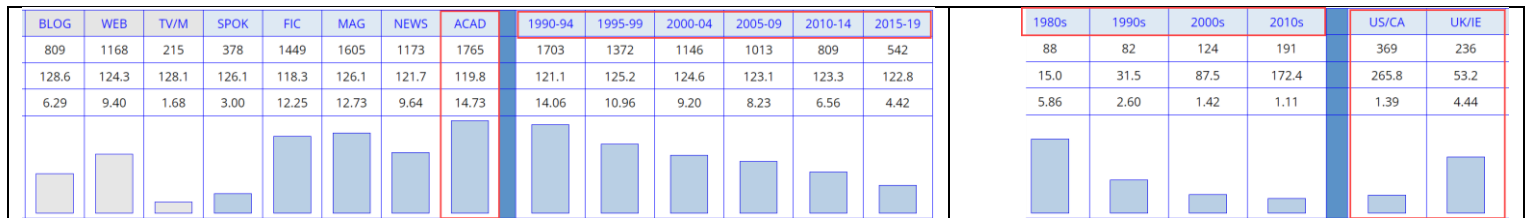
Students can enter **entire texts** that they have **written**, and then quickly and easily highlight phrases in the text to **find related phrases** in COCA, which will allow them to edit their writing to make it sound more natural.

They can also **enter entire texts from the Web**, to find the **keywords** in the text (to understand better what it's about), and also **click on any word or phrase** in the text to see a wide range of information, such as in the "Word Sketches" section above.

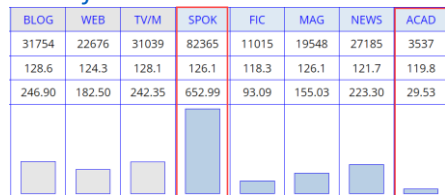
Writing (assisting non-native writers and speakers, including professors writing for publication)

Thousands of professors (from a wide range of academic fields) **use the corpora** on a regular basis, to help improve their writing and **edit papers for publication**. This is because the corpora provide information on "nuances" in English that aren't available in standard dictionaries or style guides.

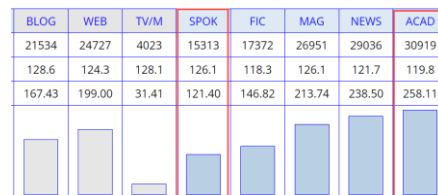
To give a simple example, the **following chart** (left) from COCA shows that the word **seldom** (as in "they seldom go there anymore") is used much more in formal genres like academic, and that it is decreasing in frequency over time. The data from the 325 million word **TV Corpus** (right) **also shows** that **seldom** is decreasing over time, and that it is much more common in British English than in American English. In other words, in American English **seldom** sounds very formal, quite old-fashioned, and somewhat British. Again, this is the type of data that a dictionary or style guide could probably never provide.



a lot of NOUN



several NOUN



Or suppose that someone is writing an academic paper, and she wants to know which sounds more formal – *a lot of NOUN* or *several NOUN*. In less than one second, she can search through one billion words in COCA and see that *several NOUN* is much more common in academic writing.




	ALL	BLOG	WEB	TV/M	SPOK	FIC	MAG	NEWS	ACAD
COMPPELLING ARGUMENT	340	86	71	23	42	8	34	27	49
STRONG ARGUMENT	331	83	57	3	54	8	38	25	63
POWERFUL ARGUMENT	148	19	20	2	28	4	17	17	41
PERSUASIVE ARGUMENT	137	21	23	12	16	5	15	14	31
GREAT ARGUMENT	76	23	19	11	11	3	4	1	4
SOLID ARGUMENT	54	20	12	4	5	2	7		4
EFFECTIVE ARGUMENT	39	6	7	2	12		5	2	5
SOUND ARGUMENT	51	21	16	1	2		1	3	7

And finally, suppose that someone wants to know which **synonyms of strong** sound better with *argument* in academic English. *Strong argument* is possible, but *compelling argument* or *powerful argument* are also common in academic English, while *great argument* or *solid argument* is common in less formal English (like on the web). The writer can click on any of these phrases to see the phrase in context.

1	1991	ACAD	TheologStud	🔍	🔍	🔍	can be brought to bear in examining conflicting goals: " It remains a compelling argument against a proposed cognitive aim if the primary theories
2	1996	ACAD	Bioscience	🔍	🔍	🔍	codes also mention obligations to superiors and funders. Nevertheless, there is a compelling argument for scientists to recognize the greater good
3	1994	ACAD	ArmedForces	🔍	🔍	🔍	detailed public case dismissing the alleged new Soviet threats: " The underlying and compelling argument " was " fully understood and never fully
4	2016	ACAD	Political Research Q	🔍	🔍	🔍	Martin and Vanberg 2011; Thies 2001). # Comparative research provides a compelling argument that parties participating in multiparty governance
5	2000	ACAD	AmerIndianQ	🔍	🔍	🔍	concerning the " facts of the case. " # The exclusivists' most compelling argument against the comparability of the two acts of genocide has been t
6	1997	ACAD	AfricaToday	🔍	🔍	🔍	to these mobile landscapes of group identity as " ethnoscapes " and makes a compelling argument for a new cosmopolitan ethnography to unrav
7	1998	ACAD	ScandinavStud	🔍	🔍	🔍	now and then. # On the one hand, we have Systembolaget's compelling argument to convince the public. On the other, we have knowledge about

In summary, the corpus data allows non-native speakers and writers to easily and quickly examine the frequency and use of words, phrases, and grammatical constructions in ways that are not possible anywhere else, including any other online corpora.

Research

As mentioned, the corpora from English-Corpora.org provide a much wider **range of searches** (and are much **faster** and easier to use as well), which has resulted in **thousands of academic articles** that are based on the corpora. At the most basic level, users can see the frequency of words or phrases by section – as with genres in **COCA** , or time period in **COHA** , or dialect in **GloWbE**  (two billion words from 20 countries).

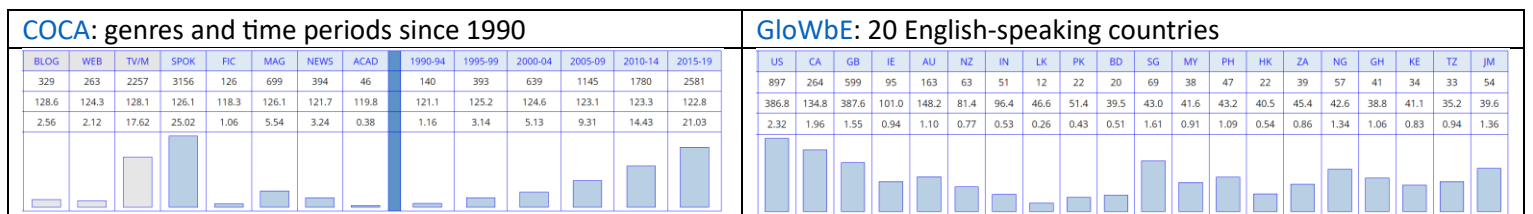
COHA: ADJ society


	ALL	1820	1830	1840	1850	1860	1870	1880	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000	2010
AMERICAN SOCIETY	1382	2	14	28	18	49	3									147	105	166	157	124	74
HUMAN SOCIETY	817	15	39	64	39	49	6									29	34	31	22	13	17
HISTORICAL SOCIETY	670	27	29	11	35	20	3									19	29	34	89	67	62
MODERN SOCIETY	595	3	5	8	18	42	4									35	33	23	43	13	37
ROYAL SOCIETY	540	6	15	20	34	59	2									19	8	24	20	15	72
CIVIL SOCIETY	525	21	19	35	18	65	2									6	3	14	56	85	147
GOOD SOCIETY	388	12	41	34	41	41	4									14	2	11	11	8	10
MEDICAL SOCIETY	352	17	17	1	5	6										39	32	24	13	11	5
GREAT SOCIETY	347	1	1	3	1	2										136	39	33	68	14	17
DEMOCRATIC SOCIETY	338															56	37	41	46	20	16
CIVILIZED SOCIETY	318	15	43	35	19	16										19	9	8	18	9	9

GloWbE: *ism

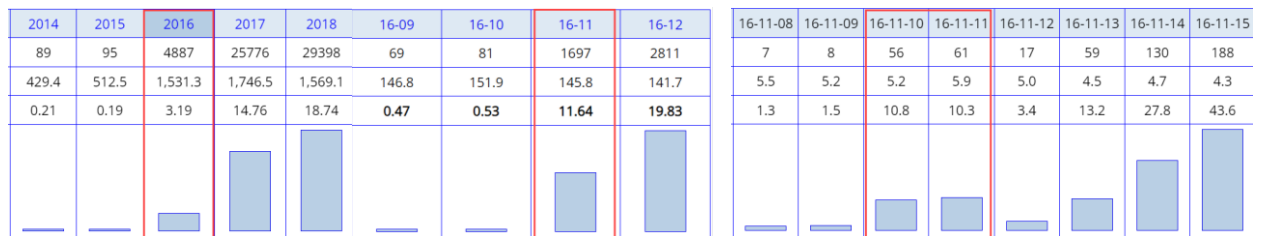
	ALL	US	CA	GB	IE	AU	NZ	IN	LK	PK	BD	SG	MY	PH	HK	ZA	NG	GH	KE	TZ	JM
TOURISM	66231	2862	3177	7376	3290	4237	3871	3564	3718	922	1706	2138	2								
CRITICISM	62753	14465	3646	15809	3165	4984	2298	3018	1841	2200	1148	811	10								
MECHANISM	44354	8851	2554	8022	2293	3576	1793	3275	1737	1107	1178	886	9								
TERRORISM	42215	8783	1912	6845	732	2102	882	2941	5427	5530	1570	317	4								
JOURNALISM	41483	10282	2879	10441	1591	3954	1090	1695	998	746	929	522	3								
CAPITALISM	37344	9466	2269	10261	1944	2835	1551	1358	683	603	874	461	2								
RACISM	36556	11535	1896	8545	1860	2988	1052	797	1082	579	332	503	8								
BUDDHISM	21816	1830	310	1437	351	757	390	1791	9064	324	829	846	11								
AUTISM	20350	7250	1514	5285	1590	2211	264	715	76	58	274	73	5								
SOCIALISM	19851	6427	792	4292	1020	1732	734	746	292	284	536	192	1								
OPTIMISM	15144	2950	1251	3767	767	990	533	678	265	375	324	347	2								
NATIONALISM	14409	1523	880	3053	1022	851	270	1033	1474	887	773	143	1								

Researchers can also see the overall frequency of a word, phrase, or **grammatical construction**, as with the “like construction” (*and he’s like, no way*) in COCA and GloWbE in the charts below. Hundreds of papers on syntactic variation and change have been published – based on data from the corpora from English-Corpora.org – have been published in the last 10-15 years.



And in the **NOW Corpus**, they can even see the frequency by year, month, and day. (NOW grows by about **5-7 million words each day**.) For example, researchers could see the frequency of *fake news* increases markedly right after the US elections on 8 November 2016. **No other corpus provides this level of detail**, and that is why researchers have used the NOW Corpus to look at a **wide range of phenomena** , including politics, economics, entertainment, and other cultural and societal shifts.

fake news in
18.1 billion word
NOW Corpus
(updated daily with
5-7 million words
of new data)




Researchers can even see the frequency of **collocates** (nearby words) in different genres, time periods, and dialects. This can signal **differences in meaning or usage**, such as with *gay* changing from “happy” to “sexual orientation” in COHA (historical), or changing collocates of *food* in the 1800s and 1970s-2010s. Similar searches can compare meaning in genres (such as in COCA) or in different countries (GloWbE). All of this allows researchers to carry out one fast, simple search to see a wide range of **information on culture and society**, and this data has resulted in hundreds of academic papers on these topics.

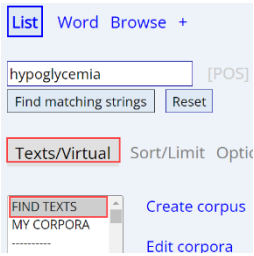
COHA: collocates of *gay*, 1820s-2010s

	ALL	1820	1830	1840	1850	1860	1870	1880	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000	2010
LESBIAN	258						1										1	7	70	81	98
GAY	216		4	2	2	10		8	8	6	2	4	22	24	4	14	8	8	18	34	38
RIGHTS	216																6	20	49	60	81
MARRIAGE	205				1	1											1		8	85	107
BRIGHT	185	5	8	10	14	13	23	12	15	12	4	12	12	15	12	8	6	3			
FLOWERS	154	5	13	10	17	9	20	16	7	13	10	11	7	5	6	1	3		1		
LAUGH	138	2	7	5	15	13	13	14	7	11	14	8	12	4	4	5	3				1
GRAVE	132	6	15	14	10	14	8	13	13	18	9	5	4	1	1						1
COLORS	125	3	6	3	9	12	8	9	10	5	7	10	6	17	8	5	6	1			
LAUGHTER	90	5	5	6	6	8	3	6	15	9	11	4	2	3	2	3	1				1
GALLANT	87	7	11	12	4	9	7	9	6	6	1	9	4	1	1						
BISEXUAL	83																	8	10	15	50
BRILLIANT	74	3	8	6	10	8	7	5	3	5	4	3	3	5	4						

COHA: ADJ *food*, 1800s (left) vs 1970s-2010s (right)

	SEC 1 (1820, 1830, 1840, 1850, 186...): 129,254,741 WORDS	SEC 2 (1970, 1980, 1990, 2000, 2010): 162,104,741 WORDS			
WORD/PHRASE	TOKENS 1	TOKENS 2	WORD/PHRASE	TOKENS 2	TOKENS 1
1 SPIRITUAL FOOD	26	1	1 FAST FOOD	281	0
2 MENTAL FOOD	27	2	2 CHINESE FOOD	221	0
3 COARSE FOOD	25	2	3 REAL FOOD	110	1
4 INTELLECTUAL FOOD	20	0	4 MEXICAN FOOD	96	0
5 LIGHT FOOD	12	1	5 AMERICAN FOOD	68	1
6 UNWHOLESOME FOOD	18	0	6 FREE FOOD	78	0
7 CHOICE FOOD	11	1	7 ITALIAN FOOD	77	0
8 DAINTY FOOD	16	0	8 ORGANIC FOOD	67	0
9 INSUFFICIENT FOOD	18	2	9 CANNED FOOD	44	1
10 WHOLESOME FOOD	62	7	10 LOCAL FOOD	56	0
11 DAILY FOOD	93	11	11 FROZEN FOOD	52	0

Virtual corpora . A feature that is extremely useful for both researchers and language learners is the ability to create custom-designed “corpora within a corpus”. These **Virtual Corpora** can be created with just a few clicks in just a few seconds, and can then be used anytime in the future.



For example, users can create Virtual Corpora **based on a given word or phrase** (for example, *hypoglycemia*, *investments*, *basketball*, or *nuclear energy*), or based on **information about the texts** (for example, works by a particular author, or subtitles from a given TV show, or related Wikipedia entries, or something as complex as articles from the *Guardian* newspaper in England from 1 Sep 2015 – 31 Dec 2015, with *refugees* in the title).

In less than one second, the corpus will create a “Virtual Corpora” of these texts, even in corpora like NOW, which have tens of millions of texts in more than 18 billion words of text.



HELP	<input type="checkbox"/>	100	WEBSITE	TEXT	# WORDS	# HITS ↓	RELEVANCE ↓	PER MILLION WORDS
1	<input checked="" type="checkbox"/>		DIABETES.ORG	HOW TO AMELIORATE THE PROBLEM OF HYPOGLYCEMIA IN INTENSIVE AS WELL ...	6286	123	19,567.3	<div style="width: 10%;"></div>
2	<input checked="" type="checkbox"/>		GBHEALTHWATCH.COM	THE FACTS ABOUT HYPOGLYCEMIA. - GB HEALTHWATCH	3129	107	34,196.2	<div style="width: 15%;"></div>
3	<input checked="" type="checkbox"/>		DIABETESSELFMANAGEMENT.COM	HYPOGLYCEMIA SYMPTOMS - DIABETES SELF-MANAGEMENT	3640	77	21,153.8	<div style="width: 10%;"></div>
4	<input checked="" type="checkbox"/>		DIABETESINCONTROL.COM	DIABETIC EMERGENCIES: HYPOGLYCEMIA CAUSED BY INSULIN, PART 3	2674	62	23,186.2	<div style="width: 15%;"></div>
5	<input checked="" type="checkbox"/>		DIABETESINCONTROL.COM	PRACTICAL DIABETES CARE, 3RD ED., EXCERPT #1: DIABETES IN THE ...	6065	60	9,892.8	<div style="width: 5%;"></div>
6	<input checked="" type="checkbox"/>		ENCOGNITIVE.COM	HOW SWEET IT IS? ENCOGNITIVE.COM	3447	59	17,116.3	<div style="width: 10%;"></div>
7	<input checked="" type="checkbox"/>		AHC MEDIA.COM	OVERDOSE OF ORAL ANTIDIABETIC MEDICATIONS AND INSULIN IN 2012	5124	56	10,007.7	<div style="width: 5%;"></div>

And users can then **search within** these Virtual Corpora. Or, in just 1-2 seconds more, they can **extract keywords**, such as these words from a [hypoglycemia] Virtual Corpus from the 14 billion word **iWeb corpus**:

HYPOGLYCEMIA [277,654 WORDS, 300 TEXTS] **NOUN** VERB ADJ ADV N+N ADJ+N [ALL VIRTUAL CORPORA] [SAVE LIST](#)

HELP	ENTRY	SAVE	WORD (CLICK FOR CONTEXT) TRANSLATE ALL ENTRIES	FREQ	# TEXTS	SPECIFIC FREQ 45 30 TEXTS	ENTIRE CORPUS	EXPECTED
1			HYPOGLYCEMIA	3636	300	12,243.6	14,974	0.3
2			GLUCAGON	350	78	3,494.6	5,050	0.1
3			UNAWARENESS	109	41	2,366.9	2,322	0.0
4			HYPERGLYCEMIA	160	63	1,576.3	5,118	0.1
5			GLUCOSE	2416	255	791.6	153,883	3.1
6			INSULIN	2240	237	633.1	178,395	3.5

This is very useful for non-native speakers who are studying, for example, aeronautical engineering or molecular biology or corporate law, and who just need to find out about the language of that narrow domain.

Other tools and features. There are many other features that cannot be fully described in this short overview. For example, just two of these are 1) the ability to create personalized **word and phrase lists** , to save words and phrase for further study, including grouping by topic, and 2) extensive links from the corpora to **external resources** , such as images, videos, pronunciation, translations, and so on.

Downloadable data. In addition to accessing the corpora via the online web interface, users can also download corpus data for use on their own computer. This includes **full-text data**, and **word frequency**, **n-grams** (word strings), and **collocates** data. This data has been used extensively by many technology-related companies, and it has also served as the backbone for thousands of academic publications.

In summary, the corpora from English-Corpora.org are (by far) the most widely used corpora in existence. Hundreds of thousands of researchers, teachers, and students use the online corpora every year, and many **universities throughout the world** have purchased **academic licenses** that provide expanded access to the corpora.

For more information, please contact us at admin@english-corpora.org.