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:

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

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>FREQ</td>
<td>8562</td>
<td>809</td>
<td>1168</td>
<td>215</td>
<td>378</td>
<td>1449</td>
<td>1605</td>
<td>1173</td>
<td>1765</td>
<td>1703</td>
<td>1372</td>
<td>1146</td>
<td>1013</td>
<td>809</td>
<td>542</td>
</tr>
<tr>
<td>WORDS (M)</td>
<td>993</td>
<td>128.6</td>
<td>124.3</td>
<td>128.1</td>
<td>126.1</td>
<td>118.3</td>
<td>126.1</td>
<td>121.7</td>
<td>119.8</td>
<td>139.1</td>
<td>147.8</td>
<td>146.6</td>
<td>144.9</td>
<td>145.3</td>
<td>144.7</td>
</tr>
<tr>
<td>PER MIL</td>
<td>8.62</td>
<td>6.29</td>
<td>9.40</td>
<td>1.68</td>
<td>3.00</td>
<td>12.25</td>
<td>12.73</td>
<td>9.64</td>
<td>14.73</td>
<td>12.25</td>
<td>9.28</td>
<td>7.82</td>
<td>6.99</td>
<td>5.57</td>
<td>3.74</td>
</tr>
</tbody>
</table>

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.
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.

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).

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.
Phrases and collocations (strings of words)  

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.

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).
**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*:

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>account</td>
<td>nearly always followed by for</td>
</tr>
</tbody>
</table>

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.

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).
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.

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):
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:

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 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):

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).

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”.
Synonyms also vary by genre. For example, consider the synonyms of \textit{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 (\text{"strong\}) \textit{beefy, burly, strapping lumberjacks} in fiction, but (\text{"strong\}) \textit{effective, compelling, persuasive arguments} in academic writing.

**Historical change** \hspace{1cm} \textit{(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 \textit{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 \textit{steamship} by decade, and \textit{Reds} by decade and even by year (note 1953, the year of the McCarthy hearings in the US Senate). As the search for \textit{a most ADJ NOUN} shows, researchers can also look for phrases, including part of speech.
Researchers can find the frequency of all matching strings 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.
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.

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.
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.

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).
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).

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).
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*).

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”.

The NOW Corpus 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.

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).

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.
Dialectal variation

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).

![Table of Frequency](image)

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.
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).

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.
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*).

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).
Virtual Corpora

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).

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.

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

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

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:

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.

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.

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.
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.
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.

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

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.

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

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).

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

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

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

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.

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.