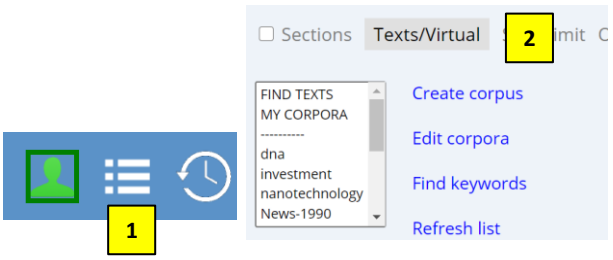
Creating and using Virtual Corpora at English-Corpora.org (see video)

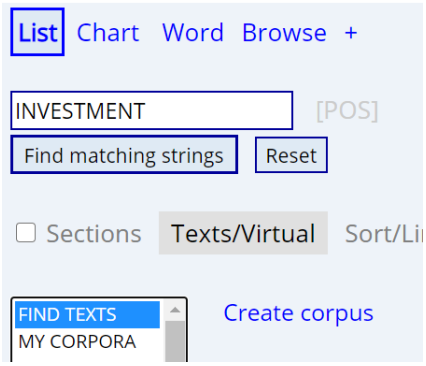
Virtual Corpora (VC) allow you to create a sub-corpus of the full corpus (often, to focus on a specific topic), and then to search just within that Virtual Corpus, compare different Virtual Corpora, and find keywords for a Virtual Corpus. This help file shows how to:

1. Create a VC based on words or phrases in the text
2. Create a VC based on information about the text – date, genre, author, country, etc
3. Organize your VC
4. Delete, add, modify texts in VC
5. Search within your VC
6. Compare across VC
7. See the keywords for your VC

You can access your Virtual Corpora via the link at the top of the corpus [1] or via the search form [2]



1. Create a Virtual Corpus based on words or phrases in the text

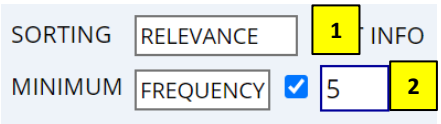


Click on **Texts/Virtual** in the search form, and then **Find Texts**. (In some of the corpora, it may be Find Articles or Find Websites).

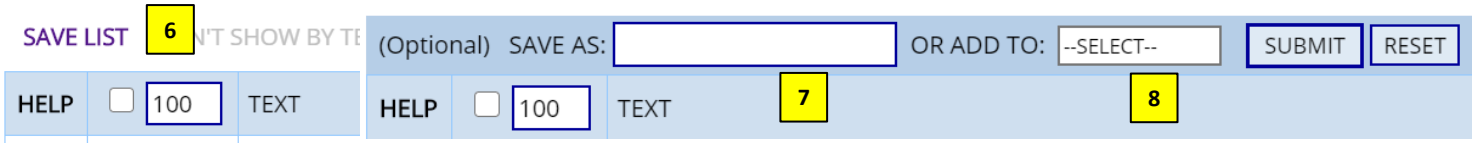
Enter the word or phrase on which the texts in the VC are at will occur in the texts in the VC. For example, INVESTMENT (all forms of investment), nuclear power, or REFUGEE.

Click on Find matching strings.

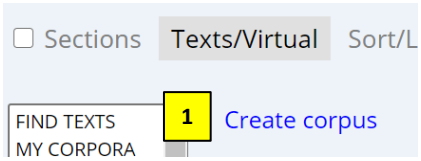
As is shown below, the next page shows what the corpus thinks are the best texts for your VC. By default, it finds the texts where your word or phrase occurs the most. But since this might favor longer texts (where there is more of anything), you can select **SORT/LIMIT** in the search form [1], and select **Relevance**. The MINIMUM field [2] shows the minimum number of times that the word or phrase must occur in the text.



Select the texts that you want to have in the VC [3]. You can (de-)select all by clicking on the checkbox at the top [4] and can decide how many texts that action applies to [5]. Click on SAVE LIST [6] and then name the VC [7]. You can also add these texts to a pre-existing VC [8]. That's all there is to creating a Virtual Corpus. It takes just 3-4 seconds and a few clicks!

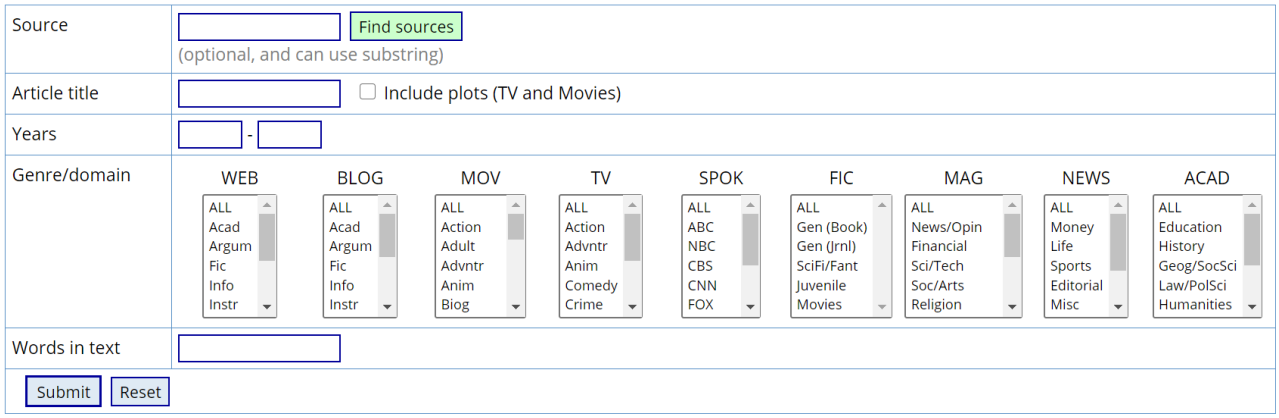


2. Create a Virtual Corpus based on information about the text

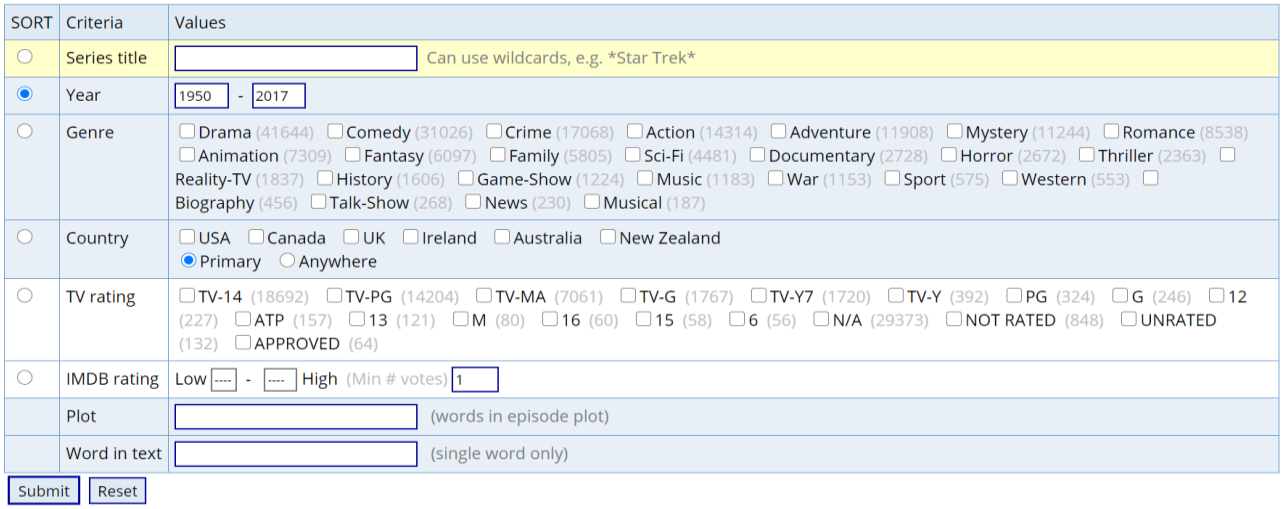


You can also create a VC based on information about the text – its date, author, country, genre, and so on. Click on **Create Corpus** [1] in the search form to create a VC in this way.

The "fields" that are available to you depend on the particular corpus. Below are the Create Corpus pages for COCA, TV, NOW, and COHA, but similar pages are available for the other corpora as well.

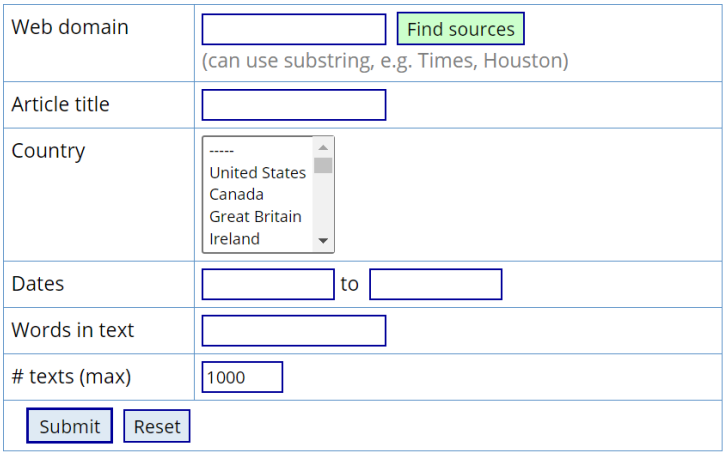


COCA

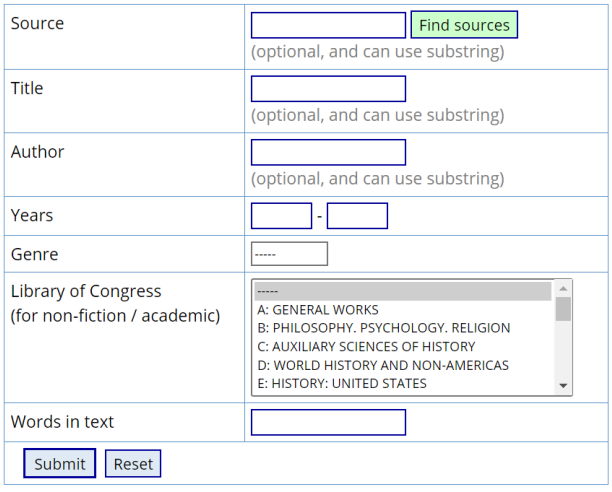


TV (similar for Movies)

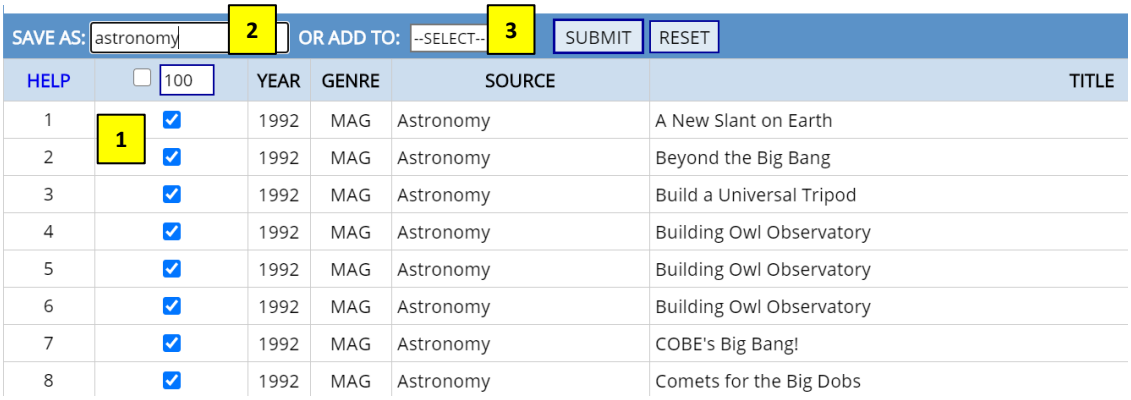
NOW



COHA

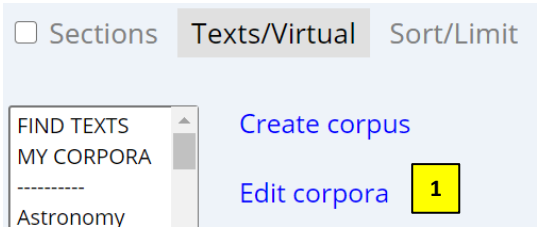


As with the VC that are based on words and phrases in the text (see Section 1 above), you can (de-)select texts [1], and then name your new VC [2], or add the texts to an existing VC [3]. (Note: this list of texts is from COCA, where the [Source] was the magazine [Astronomy].)



HELP	100	YEAR	GENRE	SOURCE	TITLE
1	☑	1992	MAG	Astronomy	A New Slant on Earth
2	☑	1992	MAG	Astronomy	Beyond the Big Bang
3	☑	1992	MAG	Astronomy	Build a Universal Tripod
4	☑	1992	MAG	Astronomy	Building Owl Observatory
5	☑	1992	MAG	Astronomy	Building Owl Observatory
6	☑	1992	MAG	Astronomy	Building Owl Observatory
7	☑	1992	MAG	Astronomy	COBE's Big Bang!
8	☑	1992	MAG	Astronomy	Comets for the Big Dobs

3. Organize your Virtual Corpora



You can see a list of all of your VC by clicking on **Edit Corpora** in the search form. You can also group VC into categories, delete VC, and move/add texts between VC. (The following list of texts comes from VC that we have created in the NOW Corpus.)

HELP			LIST NAME	# TEXTS	# WORDS	FIND KEYWORDS ◉ SPECIFIC ○ FREQ	CREATED
1	🗑	Sc	ASTRONOMY	327	249,396	NOUN VERB ADJ ADV N+N ADJ+N	1551 d
2	🗑		ASYLUM	11	8,009	NOUN VERB ADJ ADV N+N ADJ+N	877 d
3	🗑		BELTANDROAD	100	121,103	NOUN VERB ADJ ADV N+N ADJ+N	307 d
4	🗑	Sc	CLIMATECHANGE	100	409,007	NOUN VERB ADJ ADV N+N ADJ+N	1302 d
5	🗑	Me	CORONAVIRUS	498	4,077,197	NOUN VERB ADJ ADV N+N ADJ+N	139 d
6	🗑	Me	COVID	400	3,620,882	NOUN VERB ADJ ADV N+N ADJ+N	137 d
7	🗑		ELECTRON	125	87,460	NOUN VERB ADJ ADV N+N ADJ+N	1568 d
8	🗑	Po	IMMIGRANT	146	186,769	NOUN VERB ADJ ADV N+N ADJ+N	999 d
9	🗑	Fi	INVESTMENT	98	40,037	NOUN VERB ADJ ADV N+N ADJ+N	1569 d

Note: nearly every page at the corpus has a HELP link [1 above], which gives context-sensitive instructions for what you can do on that page. Also, you can click on most of the columns to sort your corpora.

[5] lists all of your Virtual Corpora. You can click on any VC to delete, add, or move texts (from one VC to another).
[6] shows the number of texts and the size of the VC, and [9] shows how many days ago you created the VC.
[2] deletes the VC (it will first prompt you for confirmation). [3] doesn't delete the VC, but it "ignores" it so that it doesn't appear in the list of VC in main search form, and the VC will not be used when comparing the frequency of words in different VC (see Section 6).
[7-8] allow you to see the keywords from the VC (see Section 7 below for more information)
[4] allows you to create a category for the VC (e.g. above Fi = Financial, Sc = Science, etc). You can then group your VC by clicking on the header for this column.

4. Delete, add, modify texts in a Virtual Corpus

You can easily modify the list of texts for your VC. Select the desired texts [1] and then delete the texts [2], add them to another VC [3], move them to another VC [4] (and specify the other VC via [5])

TELESCOPE (RENAME) DELETE ADD TO MOVE TO --SELECT-- (SEE ALL VIRTUAL CORPORA)

HELP	100	2 AR	G 3 E	4	SOU 5	TITLE
1	1	1990	ACAD	Mercury		The Space Telescope: Eyes Above the Atmosphere....
2		1990	MAG	ScienceNews		Dawn of a big telescope. (cover story)...
3		1990	MAG	ScienceNews		Space Telescope: A Saga of Setbacks. (cover story)...
4		1990	NEWS	NYTimes		New European Telescope Called Equal to Hubble...
5		1990	SPOK	ABC_Nightline		The Hubble Telescope Launch
6		1990	SPOK	PBS_Newshour		Newshour 900409
7		1990	SPOK	PBS_Newshour		Newshour 900628
8		1991	ACAD	Mercury		Discovering the invisible universe. (cover story)...

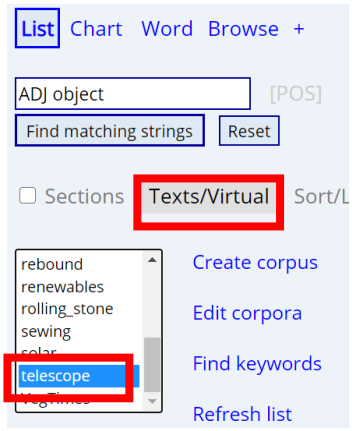
5. Search within a Virtual Corpus

The real power of Virtual Corpora is that we can just focus on the part of the overall corpus that is of interest to us. This is useful if we have a 10 billion or 14 billion word corpus, and we are only interested in searching texts for a particular topic, like nuclear power, or Buddhism, or astronomy. For example, we can search the one billion word COCA corpus for ADJ object and we see the results in [1]. Or we can create a Virtual Corpus that we have created, dealing with telescopes, and then just search those texts (see [2]), and the results are specific to astronomy.

1

PHYSICAL OBJECT	231
OTHER OBJECT	183
INANIMATE OBJECT	182
FOREIGN OBJECT	130
BLUNT OBJECT	126
SINGLE OBJECT	125
SHINY OBJECT	119
DIRECT OBJECT	116
NEW OBJECT	115
IMMOVABLE OBJECT	107
SMALL OBJECT	107

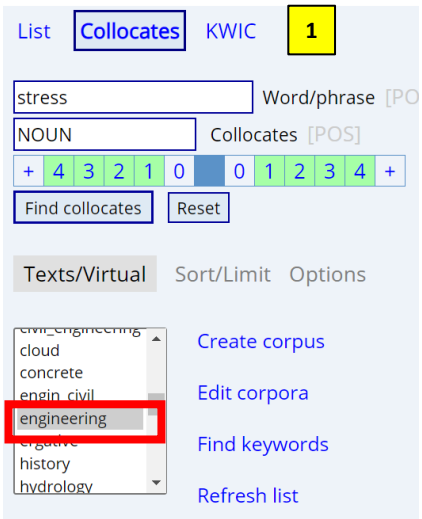
2



3

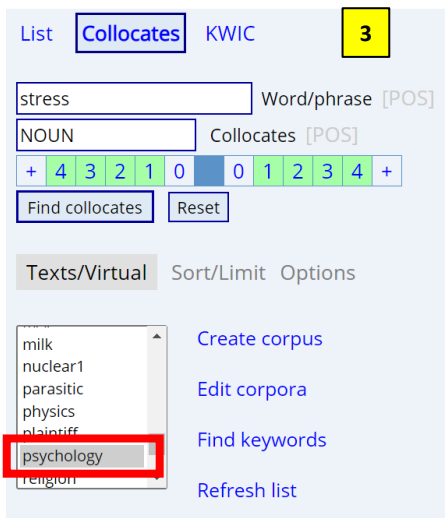
FAINT OBJECT	20
CELESTIAL OBJECT	12
DEEP-SKY OBJECT	9
DISTANT OBJECT	6
ASTRONOMICAL OBJECT	5
BRIGHT OBJECT	4
FUZZY OBJECT	4
OTHER OBJECT	3
MASSIVE OBJECT	3
DESIRED OBJECT	3
MESSIER OBJECT	2

We can also use Virtual Corpora to focus on the meaning of a given word in a particular semantic domain. For example, *stress* has different means depending on whether we are talking about engineering or psychology. In the Wikipedia corpus, we can limit our search to an engineering VC that we have created [1] and then we see the collocates [2], or we can limit the search to a VC dealing with psychology [3] and we see the collocates [4].



2

TENSOR
SLIP
FATIGUE
DISLOCATIONS
YIELD
STRENGTH
RELATIONSHIP
LIMIT
AMOUNT
TESTS
MAXIMUM



4

REDUCTION
CORTISOL
PATHWAYS
PREGNANCY
POST
INCIDENCE
EUSTRESS
AMOUNTS
DISORDER
MANAGEMENT
DIVORCE

6. Comparing across Virtual Corpora

Once you have created multiple Virtual Corpora, you can comparing the frequency of word, phrases (or even a given grammatical construction) in different VC. (Before doing this, you might want to "disable" VC that you don't want to compare, as in #3 in Section 3 above). For example, suppose that we have created four different VC by searching for the words quran (=Islam), bible (=Christianity), SUTRA (=Buddhism), and atheism. We can then compare the frequency of the following words in these four VC, and this might tell us something about the frequency with which these four belief systems discuss certain topics.

prophet

HELP	☐ 100	TEXT	# WORDS	# HITS ↕	RELEVANCE ↕	PER MILLION WORDS
1	☑	QURAN	662499	469	707.9	
2	☑	BIBLE	2065862	224	108.4	
3	☑	ATHEISM	2458001	65	26.4	
4	☑	SUTRA	455808	9	19.7	

reason

HELP	☐ 100	TEXT	# WORDS	# HITS ↕	RELEVANCE ↕	PER MILLION WORDS
1	☑	ATHEISM	2458001	1809	736.0	
2	☑	BIBLE	2065862	1027	497.1	
3	☑	QURAN	662499	270	407.5	
4	☑	SUTRA	455808	142	311.5	

compassion

HELP	☐ 100	TEXT	# WORDS	# HITS ↕	RELEVANCE ↕	PER MILLION WORDS
1	☑	SUTRA	455808	52	114.1	
2	☑	QURAN	662499	30	45.3	
3	☑	BIBLE	2065862	76	36.8	
4	☑	ATHEISM	2458001	69	28.1	

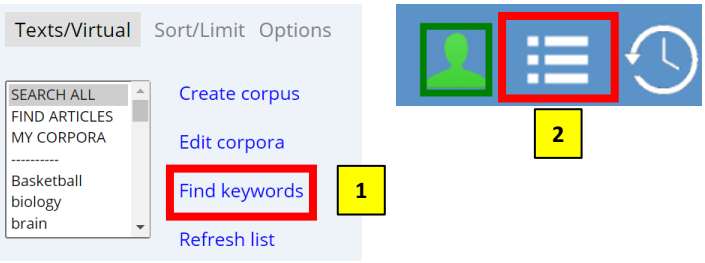
salvation

HELP	☐ 100	TEXT	# WORDS	# HITS ↕	RELEVANCE ↕	PER MILLION WORDS
1	☑	BIBLE	2065862	335	162.2	
2	☑	QURAN	662499	56	84.5	
3	☑	SUTRA	455808	30	65.8	
4	☑	ATHEISM	2458001	102	41.5	

Other examples might be comparisons of:
-- newspapers in NOW (e.g. a presumed "progressive" newspaper and a more "conservative" one)
-- scientific disciplines in Wikipedia (e.g. which uses the words *empirical* or *arguments* the most)
-- genres of TV or Movies in those two corpora (e.g. dramas vs sitcoms on TV)
-- speakers from different political parties in the Hansard corpus
-- different authors in COHA
-- topics (e.g. religion or political philosophy or science) in EEBO

7. Finding keywords

Perhaps the best use of Virtual Corpora (at least for language learners) is the ability to quickly and easily generate lists of "keywords" for a given topic. Examples might be words related to biology in Wikipedia, words from the magazine Astronomy in COCA, websites in iWeb dealing with endocrinology or solar power, or articles referring to refugees in NOW.



To see the keywords list for a given VC, click on Find Keywords under Texts/Virtual in the search form [1], or click on the Virtual Corpora icon at the top of the corpus [2]. Then click on the desired part of speech (e.g. NOUN or ADJ) for the desired VC [3]

HELP				LIST NAME	# ARTICLES	# WORDS	3 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

If [3] is set to SPECIFIC (the default), it will show the words that are much more frequent in the VC than in the corpus as a whole. If it is set to FREQ(UENCY), then it will show the most frequent nouns, adjectives, etc in the VC (so that overall high-frequency words like *people, time*, or *good* may be at the top of the "keyword "list).

After selecting a VC and a part of speech, you will see the keywords list (this one is taken from a biolog* list from the Wikipedia corpus). The HELP link [1] on that page explains very well the different columns in the table.

BIOLOGY [142,355 WORDS, 100 TEXTS] NOUN VERB ADJ ADV N+N ADJ+N [ALL CORPORA] SAVE LIST

HELP 1	WORD (CLICK FOR CONTEXT) 2	FREQ 3	# TEXTS	SPECIFIC 4 FREQ 30 12 TEXTS	ALL WIKIPEDIA 6	EXPECTED 7
1	ORGANISM	329	52	346.8	12,327	0.9
2	NEURON	42	13	133.2	4,097	0.3
3	BIOLOGIST	89	27	123.3 5	9,379	0.7
4	BIOLOGY	445	50	116.1	49,803	3.8
5	CHROMOSOME	68	14	89.9	9,834	0.8
6	MOLECULE	105	29	86.6	15,753	1.2
7	MEMBRANE	146	18	85.8	22,106	1.7
8	CELL	736	50	83.8	114,074	8.8
9	GENOME	88	25	81.4	14,055	1.1
10	EGG	183	12	80.5	29,546	2.3

As the HELP link [1] explains:
[2] are the list of keywords, and [3] is the number of texts and the number of tokens for the word.
[6] is total number of tokens for the word in the entire corpus (not just the Virtual Corpus) and [7] is the "expected" frequency of the word in the VC. [5] is a number representing how much more frequent the word is than would be expected (again, see HELP for more details on the exact formula used, which is similar to Log Likelihood).
[4] allows you to create a more or less specific keyword list. If you click on [-], you will decrease how specific the words are to the VC, and [+] will increase this. Make sure you click on SPECIFIC again after [-] or [+] to change the list. See the HELP page online [1] for more details and examples.

Again, the KEYWORDS feature can be a great way to find the specific vocabulary for a particular topic.