

Cambridge Sketch Engine

Advanced Help (v.2.0)

Contents:

1	Creating corpora from the web – WebBootCaT	р.3
2	View options - advanced	р.б
3	Simple vs. multilevel sort	p.7
4	Filter	p.9
5	Frequency	p.9
6	Building a subcorpus	p.11
7	Working with a subcorpus	p.11
8	Building wordlists	p.12
9	Comparing corpora using keywords	p.12
10	Word Sketch – Advanced options	p.13
11	Uploading your own text files to Sketch Engine	p.13

1. Creating corpora from the web - WebBootCaT

The WebBootCaT function allows you to build corpora from the web. WebBootCaT can be used to complement the existing Cambridge Corpus resources – it can be used to compile a corpus on a particular topic or subject not currently found within the Corpus (for example, on *new technology, international law* or *retail.*)

WebBootCaT requires only a list of seed words (terms that are expected to be typical of the domain of interest) as an input. For example, to generate a *new technology* corpus, you might use seed words such as *phone, wi-fi, email, wireless, Internet*, etc. BootCaT then generates a corpus based on searches for these seed words.

To build your own corpus, click on WebBootCaT (shown in red below) from the Sketch homepage:

orpora 🍝 Create corpus	Corpora				
+ WebBootCaT	Corpus	name	Language	Size	
Configuration templates	Cambridge Academic Corpus		English	412,966,498	,
Sketch grammars	Cambridge International Corp	us	English	1,358,129,340	, 🔍 🔍
user groups	Cambridge Learner Corpus Co	ided	English	25,542,115	
Current	Cambridge Learner Corpus Qu	uestion Papers	English	44,859	
Support	Cambridge Learner Corpus Ur	icoded	English	43,787,263	
Help Second a bus	Cambridge Spoken Corpus		English	96,856,880	
ceport a bug					
	My corpora				
	Corpus ID	Corpus pame	Lang	1300 61	

This takes you to a screen where you can enter the name of your corpus, the language of your corpus (English) and your seed words.

For example, if we wanted to make a *new technology* corpus, our screen might look like this:

Seed words should be separated from each other by a space. Seed words that consist of more than one word (such as *social network* in the example on the right) should be enclosed in quotation marks.

The more seed words you enter, the larger your corpus will be.

Once you have entered all of your seed words click on *Next* at the bottom of the page.

UNIVERSITY PRESS					
user: Claire Dembry used tok	ens: 4,671,929 / 10,000,00	00 days left: unlimited			
Corpora	WebBootCaT:	Create corpus			
Configuration templates	Corpus ID	New Technology			
Sketch grammars User groups		Unique identifier of your corpus. May only contain letters, numbers, underscores and hyphens.			
	Language	English			
Admin Local administration		Creating BootCaT corpora is available only for those language which we can at least tokenise. All such languages are listed here.			
	Build word sketches				
Support		This option only has effect if a pre-loaded sketch grammar is available for the selected language.			
Report a bug	Input type	Seed words URLs Select "URLs" to download data from specified URLs rether than use seed words for finding the URLs			
	Seed words	phone wi-fi wireless Internet 'social network' text mobile digital 'online shopping' Use space as separator. Enclose multiword expressions into quotes (").			
	Compile corpus when finished	✓ Automatically compile corpus when WebBootCaT processing is finished.			
	Show advanced options	5			
		Cancel Next>			

This takes you to a screen that shows the websites

that BootCaT has found that relate to your seed words. Here you can browse this list of URLs and untick any sources that you do not wish to use - all of the boxes are ticked by default.

CAMBRIDGE UNIVERSITY PRESS						
user: Claire Dembry used to	user: Claire Dembry used tokens: 4,671,929 / 10,000,000 days left: unlimited					
Corpora 💠 Create corpus	WebBootCaT: Create corpus					
WebBootCaT	Please select URLs which you would like to process.					
Configuration templates Sketch grammars	Cancel < Back OK					
User groups	Query: 'online phone wi-fi					
Admin	Http://www.qwest.com/					
Local administration	✓ http://blog.onlinedeals.in/wi-fi/wi-fi/data-card.html					
Lucal duministration	M http://www.chinasuppliers.globalsources.com/china-suppliers/Phone-WI-Fi.htm					
Support	Mhttp://www.wireless.att.com/learn/internet/accessing-wifi.jsp					
Help	Inttp://www.onlinemobilephonedeals.org.uk/wi-ti-mobile-phones.ntml Inttp://www.busineesingides.com/0000(10/blackbarry.sterm.wifi					
Report a bug	Intep://www.businessinsider.com/2008/12/blackberry-scom-will Http://www.businessinsider.com/2008/12/blackberry-scom-will					
	Inter://www.chinasuppliers.globalsources.com/china.suppliers/WI-Ei-Phone.tom					
	☑ http://www.myshonping.com.au/PT31_Mobile_Dhones_Nokia_WiFi_802_11fs_102_18009_e					
	Month Strate Str					

Once you've selected the URLs you wish to use, click OK (at the top, and also at the bottom of the page) to download data from these sources.

Sketch Engine will now process the text so that it can be used with the same functionality as the Cambridge Corpus resources - the toolbar at the top of the screen keeps you updated on BootCaT's progress.

user: Clare Dembry used tokens: 4.671,929 / 10,000.000 days left: unlimited Corpus	CAMBRI UNIVERSITY	DGE PRESS			
Corpora ● Create corpus WebbootCaT: Configuration templates Sketch grammars User groups Add new file Add new file Add new file Add new file Compute Compute Compute Compute Compute Compute Add new file Successfully processed files Compute Compute Compute corpus Compute corpus Condiguration corpus Download corpus Access privileges View logs Support Help Report a bug Paper to du Postes corpus: text/htall Protes corpus: text/htall Protes corpus: text/htall Protes corpus: text/htall Download corpus Access privileges </th <th>user: Claire Dembry used to</th> <th>kens: 4,671,929 / 10,000,000 da</th> <th>ys left: unlimit</th> <th>ed</th> <th></th>	user: Claire Dembry used to	kens: 4,671,929 / 10,000,000 da	ys left: unlimit	ed	
• Webbootcat 30% Configuration templates Successfully processed files 15 Sketch grammars 15 Files remaining 69 Dat downloaded 2.491 kB	Corpora	New_Technology:	WebBoo	otCaT: Downloading da	ata
Configuration templates Sketch grammars User groups Corpus Add new file Add web data (BootCaT) Compus Comple corpus Configure corpus Content-type: text/htal; characet-UTF-8 Pile type: thal Porte cost words (approx.): 2873 Content-type: text/htal; characet-UTF-8 Pytes read: 192261 Mamber of words (approx.): 2055 Maber of words (approx.): 2055 Maber of words (approx.): 2055 Pytes read: 192261 Dytes read: 192261 Support Help Report a bug Pile type: thal Pytes read: 202061 Pile type: thal Pytes read: 202061 Pile type: thal	WebBootCaT				30%
Sketch grammars User groups Corpus Add new file Add new file Add new file Add new file Add web data (BootCar) Compus Compus Compus Compus Compus Add web data (BootCar) Compus Configure corpus Contigure corpus Pownload corpus Download corpus Addmin Local administration Support Help Report a bug Pibles creat: 142261 Pibles creat: 20201 Add min Pible Download corpus Admin Pible tor certivex: hall Pibl	Configuration templates	Successfully processed files	15	Errors	15
User groups User groups Data downloaded 2481 kB Tokens retrieved - invalid content-type 0 Add web dat 0.381 - invalid content-type 0 Add web dat 0.381 - invalid content-type 0 Add web dat 0.381 - invalid content-type 0 Mad web dat 0.381 - invalid content-type 0 Complie corpus 0.081 0.381 - invalid content-type 0 Complie corpus - Add web dat 0.381 - invalid content-type 0 Complie corpus - Amerage file processing time 1.3 s - invalid content-type 0 Configure corpus - Maber of words (approx.): 2873 - onable convert to text 0 Configure corpus - Maber of words (approx.): 2873 - Detected character encoding: ascii - Detected character encoding: ascii Processing http://www.inblow.com/Create-a-Wireless-Network - Content-type: thylind: characteriso-8859-1 - Detected character encoding: ascii Processing http://www.inbl.characteriso-8859-1 - Site type: thai - Site type: thai Processing http://www.inbla.characteriso-8859-1 - File type: thai - File type:	Sketch grammars	Files remaining	69	- unable to retrieve	8
Corpus - file size out of range 1 Add new file - file size out of range 1 Add web data 0:39,516 - number of words out of range 6 Add web data 0:39,516 - number of words out of range 6 Comple corpus 0:39,516 - number of words out of range 6 Comple corpus 0:39,516 - duplicate 0 Comfigure corpus - Number of words (approx.): 2073 - duplicate - duplicate Configure corpus - Number of words (approx.): 2073 - Cancel processing - duplicate - duplicate Download corpus - Number of words (approx.): 1085 - betected character encoding: ascii - Detected character encoding: ascii - Detected character encoding: ascii Processing http://www.wikhowk.com//treate-a-Wireless_Network/ - Content-toppe: text/htul - Detected character encoding: ascii - Detected character encoding: ascii Processing http://www.ikhokoks.org/wiki/Wireless_networks/ - Download corpus - Detected character encoding: ascii - Download corpus - Detected character encoding: ascii Admin - Suber of words (approx.): 2332 - Detected character encoding: ascii - Folied to retri:wei - Folied to retri:wei </td <td>User groups</td> <td>Data downloaded</td> <td>2481 kB</td> <td>- invalid content-type</td> <td>0</td>	User groups	Data downloaded	2481 kB	- invalid content-type	0
Corpus 10kens per file (avg) 2,234 Add web date		Tokens retrieved	33,516	- file size out of range	1
 Add new file Add web data (GootCaT) (Compile corpus (Compile corpus	Corpus	Tokens per file (avg)	2,234	- number of words out of range	6
Add web data Add web data (BootCar) Compile corpus Open in SkE Estimated time remaining 1.33 Compile corpus Oningure corpus Configure corpus Configure corpus Change sketch Processing http://www.wikinow.com/ficeate-a-Wireless-Network Download corpus Download corpus Access privileges View logs Admin Local administration Support Help Report a bug Postes text/html - Trile type: html - Bytes read: 142601 - State type: html - Synchronic tripe: text/html - Processing http://www.lehnan.edu/itr/wireless_networks/ - Content-type: text/html - Filed to retrifieve: - Filed to retrifieve: Processing http://www.lehnan.edu/itr/wireless_Home_Network_Basics - Filed to retrifieve: Processing http://www.lehnan.edu/itr/wireless-network.php - Detected character encoding: ascii Processing http://www.lehnan.edu/itr/wireless-Home_Wetwork_	🔶 êdd pew file	Time elanced	0:38	- keywords filter annlied	0
Average file processing time 1.3 ()	Add web data	Estimated time remaining	1:27	- unable to convert to text	0
Admin - Number of words (approx.): 2873 Admin - Present mode Admin - Present mode View logs - Detected character encoding: ascii Processing http://www.viki/wichow.com/freete-e-Wireless-Network - Contenttype: text/http://scharet-ise-8659-1 - Transperiment Support Help Report a bug - Filed to retrieve: Processing http://www.ikinboks.org/wiki/Wireless_New_Wetworks/ - Transperiment - Transperiment - Support Help Report a bug	(BootCaT)	Average file presessing time	1.27	dualiante	0
Open in SkE Extract keywords Configure corpus Configure corpus Change sketch grammar Expert mode Download corpus Access privileges View logs Access privileges View logs Muber of vords (approx.): 2873 Aumber of words (approx.): 2873 Aumber of words (approx.): 2873 Aumber of words (approx.): 1085 Detected character encoding: ascil Processing http://www.retx.it/whiteless_home_Networks/ Content-type: text/html Pytes read: 142261 Nubber of words (approx.): 2332 Puber html Pytes read: 142261 Nubber of words (approx.): 2332 Puber html Pytes read: 142261 Nubber of words (approx.): 2332 Puber html Pytes read: 142261 Nubber of words (approx.): 2332 Puber html Pytes read: 142261 Nubber of words (approx.): 630 Nubber of words (approx.): 650 Petected character encoding: ascil Processing http://www.text.it/wbleweb/ Support	Compile corpus	Average life processing time	1.5 5	- duplicate	0
 Extract keywords Configure corpus Change sketch grammar Expert mode Expert mode Download corpus Actions Expert mode Download corpus Actions Actions	Onen in SkE			🗶 Cancel p	rocessing
Configure corpus - Number of words (spprox.): 2373 Change sketch - Detected character encoding: ascil Processing http://www.wikinow.com//Create-a-Wireless-Network Content-type: text/html Download character encoding: ascil Admin Local administration Support Help Report a bug Processing http://www.leinan.edu/itr/wireless_Hetworks/ - Content-type: text/html - Support Help Report a bug - Nuber of words (sprox.): 630 - Detected character encoding: ascil Processing http://www.leinan.edu/itr/wireless_Hetwork_Basics - Pailed to retrieve: Processing http://wwww.leinan.edu/itr/wireless_	Extract keywords				
 Change sketch grammar Detected character encoding: sacii Frocessing http://www.thr.it/scate-a-Wireless-Network Content-type: text/html; charact=UTF-8 Pile type: thal Bytes read: 53232 Number of words (approx.): 1085 Detected character encoding: sacii Processing http://www.twr.iterka.com/section/wireless_networks/ Content-type: text/html; charact=iso-8859-1 File type: thal Bytes read: 142261 Number of words (approx.): 2332 Detected character encoding: sacii Processing http://www.tholooks.org/witk/Wireless_None_Network_Basics Faled to retrieve: Processing http://www.text.it/wolleweb/ Support File type: thal Bytes read: 20050 Detected character encoding: sacii Processing http://www.text.it/wolleweb/ 	Configure corpus	- Number of words (approx.):	2873		~
grammar Fr.Cossing intop://www.inimo.cou/itests-metuoin Expert mode Content-copy: text/html; charset/UT-8 Download corpus Processing intop://www.inimo.cou/itests-metuoin Acoms Processing intop://www.inimo.cou/itests-metuoin View logs Processing intop://www.inimo.cou/itests-metuoin Admin Processing introp://www.inimo.cou/itests-metuoins Local administration File type: thal Support Processing introp://www.inimo.cou/it/it/vireless_Mae_Wetwork_Basics Help Processing http://www.inima.cdu/it/vireless_network.php - Content-type: text/html Processing http://www.inima.cdu/it/vireless_network.php - Doutent-type: text/html Processing http://www.inima.cdu/itr/wireless_network.php - Doutent-type: text/html Processing http://www.inima.cdu/itr/wireless_network.php - Doutent-type: text/html Pile type: thal - Filet vords (approx.): 650 Detected character encoding: ascii Processing http://www.inima.cdu/itr/wireless-network.php Processing http://www.text.it/bilewe//	🖉 Change sketch	- Detected character encoding	r: ascii	. Uivelees Vetweek	
Expert mode - File type: fnal Download corpus - Fytes read: 5932 Access privileges - Number of words (opprox.): 1005 View logs - Dected character encoding: ascil Admin - Content-type: text/html; characteriso-0859-1 Local administration - File type: html Support - Dected character encoding: ascil Help - File type: html Report a bug - File twr/html - File twr/second - File twr/second - Nuber of words (opprox.): 630 - Dected character encoding: ascil - Nuber of words (opprox.): 650 - Dected character encoding: ascil - Words encoding: ascil - File type: html - Bytes read: 20050 - Nuber of words (opprox.): 650 - Dected character encoding: ascil - Dected character encoding: ascil Processing http://www.text.it/ableweb/ - Vector Content-type: html - Bytes read: 20050 - Detected character encoding: ascil - Deceted character encoding: ascil	grammar	- Content-type: text/html: ch	arset=UTF-8	a-wireless-Network	
 Download corpus Bytes read: 59223 Access privileges View logs Pocessing http://www.text.it/wolleweb/ Content-type: text/httal; charset=iso-8859-1 File type: htal Bytes read: 142261 Number of words (approx.): 2332 Detected character encoding: ascil Processing http://www.text.it/wolleweb/ 	📱 Expert mode	- File type: html			
 Access privileges Nuber of words (approx.): 1005 Detected character encoding: ascil Admin Content-type: text/html; charact+ise.setworks/ Content-type: text/html; charact+ise.setworks/ Torcessing http://www.text/html Bytes read: 14260 Nuber of words (approx.): 2332 Nuber of words (approx.): 2332 Processing http://www.text/html Forcessing http://www.text/html Forcessing http://www.text.html File type: html Bytes read: 20050 Nuber of words (approx.): 650 Detected character encoding: ascil Processing http://www.text.it/abileweb/ 	🛓 Download corpus	- Bytes read: 59232			
View logs View logs Admin Locasing http://www.crwiteless.com/section/wireless_networks/ - Content-type: text/html; characteriso-8859-1 - Content-type: text/html - Processing http://www.leman.edu/itr/wireless_network.php - Content-type: text/html - File type: thal - Soutent-type: text/html - Detected character encoding: sacia Processing http://www.text.it/mobilewe/ Content-type: text/html - File type: thal - File type: thal - File type: thal - Processing http://www.text.it/mobilewe/ Content-type: text/html - Soutent-type: text.it/mobilewe/	🔒 Access privileges	- Number of words (approx.):	1085		
Admin Local administration Support Help Report a bug - Tile type: thai - Suptort - Support Help Report a bug - Tile type: thai - Bytes read: 14261 - Nuaber of words (approx.): 2332 - Detected character encoding: ascil Processing http://www.lehnan.edu/itr/wireless_hame_Hetwork_Basics - Failed to retrieve: Processing http://www.lehnan.edu/itr/wireless-network.php - Ontent-type: text/htal - File type: thai - Bytes read: 20050 - Nuaber of words (approx.): 650 - Detected character encoding: ascil Processing http://www.text.it/holleweb/ C	🔍 View logs	Processing http://www.rcrwire	less.com/sect	tion/wireless networks/	
Admin Local administration Support Help Report a bug Vertexted: character encoding: ascil Processing http://www.text.it/abileweb/		- Content-type: text/html; ch	arset=iso-88	59-1	
- Bytes read: 142261 - Bytes read: 142261 - Number of words (approx.): 2332 - Butter of words (approx.): 2332 - Betexted character encoding: saci1 Frocessing http://en.wiki/Wireless_Home_Wetwork_Basics - Field to retrieve: Processing http://en.wiki/Wireless_network.php - Content-type: text/html - File type: html - File type: thml - Words (approx.): 650 - Detected character encoding: saci1 Processing http://www.text.it/mobilewe/ C	Admin	- File type: html			
Local administration - Detected character percoding: ascil Frocessing http://www.text.it/wolleweb/ - Falled to retrieve/ Help Report a bug - Content-type: thal - Bytes read: 20050 - Number of words (epprox.): 650 - Detected character encoding: ascil Processing http://www.text.it/wolleweb/ C - Content-type: that - Bytes read: 20050 - Detected character encoding: ascil - Content-type: that - Bytes read: 20050 - Detected character encoding: ascil - Content-type: that - Bytes read: 20050 - Detected character encoding: ascil - Content-type: that - Bytes read: 20050 - Detected character encoding: ascil - Content-type: that - Bytes read: 20050 - Detected character encoding: ascil - Content-type: that - Bytes read: 20050 - Detected character encoding: ascil - Content-type: that - Bytes read: 20050 - Detected character encoding: ascil - Content-type: that - Bytes read: 20050 - Detected character encoding: ascil - Content-type: that - Bytes read: 20050 - Detected character encoding: ascil - Content-type: that - Bytes read: 20050 - Detected character encoding: ascil - Content-type: that - Bytes read: 20050 - Detected character encoding: ascil - Content-type: that - Bytes read: 20050 - Detected character encoding: ascil - Content-type: that - Bytes read: 20050 - Detected character encoding: ascil - Content-type: that - Bytes read: 20050 - Detected character encoding: ascil - Content-type: that - Bytes read: 20050 - Detected character encoding: ascil - Content-type: that - Bytes read: 20050 - Detected character encoding: ascil - Content-type: that - Bytes read: 20050 - Detected character encoding: ascil - Content-type: that - Bytes read: 20050 - Detected character encoding: ascil - Content-type: that - Bytes read: 20050 - Detected character encoding: ascil - Content-type: that - Bytes read: 20050 - Detected character encoding: ascil - Content-type: that - Bytes read: 20050 - Detected character encoding: ascil - Content-type: that - Bytes read: 20050 - Detected character encoding: ascil - Content-type: that - Bytes read:	Admin	- Bytes read: 142261	2222		
Support Support Help Processing http://www.icki.ooks.org/wiki/Wireless_Home_Wetwork_Basics - Failed to retrizeve: Processing http://www.lehman.edu/itr/wireless-network.php - Content-type: text/html - File type: thal - Bytes read: 20000 N - File type: html - Bytes read: 20000 N - Bytes	Local administration	- Detected character encoding	r: ascii		
Support Help Report a bug Processing http://www.leman.edu/it/vireless-network.php - Content-type: text/html - Sytes read: 20050 - Number of words (approx.): 650 - Detected character encoding: sscii Processing http://www.text.it/mobileweb/ C		Processing http://en.wikibook	s.org/wiki/W	ireless_Home_Network_Basics	
Help Report a bug - Content-Cype: text/html - File type: html - File type: html - Bytes read: 20050 - Number of words (approx.): 650 - Detected character encoding: ascii Processing http://www.text.it/nobileweb/	Support	- Failed to retrieve:			
Heiji - File type: thal - File type: thal - Bytes tesd: 2000 - Number of words (approx.): 650 - Beteted character encoding: soci - Processing http://www.text.it/mobileweb/ >	11-1-	- Content-type: text/html	eau/itr/wire	less-network.php	
Report a buy - Bytes read: 2005 - Number of words (approx.): 650 - Detected character encoding: ascii Processing http://www.text.it/abileweb/	Help Depart a hug	- File type: html			
- Number of words (approx.): 650 - Detected character encoding: ascii Processing http://www.text.it/mobileweb/	Report a buy	- Bytes read: 20050			
- vetetea character encoaing: ascil Processing http://www.text.it/hobileveb/		- Number of words (approx.):	650		
		J - Detected character encoding Processing http://www.text.it	f: ascii :/mohileweb/		~
		<	, additeeed)		>
		<u>r</u>			

This can take a few moments – particularly if you've entered lots of seed words.

You can navigate away from this screen (e.g. by clicking on the Cambridge logo at the top left of the page) and leave BootCaT running in the background while you complete other corpus searches.

You can return to check on the progress of BootCaT at any time by clicking on the name of your corpus from the opening menu page (navigate there by clicking on the Cambridge logo.)

Once the data has downloaded, the following screen is shown:

CAMBRIDGE UNIVERSITY PRESS						
user: Claire Dembry used to	user: Claire Dembry used tokens: 4,729,614 / 10,000,000 days left: unlimited					
Corpora	New_Technology		-			
♦ WebBootCaT ♦ Add new file / ♦ Add data from web using WebBootCaT / ③ Compile corpus / ④ Open in SkE				SkE		
Contiguration templates Sketch grammars	# Original file Plain text Vertical To	kens	Owner			
User groups	😑 New_Technology (29 files)	57,685	Claire Dembry	,	×	

In order to use all of the functions available in Sketch with your new corpus, (e.g. Word Sketch, Sketch Diff) you now need to <u>compile</u> your corpus. To do this, click on *Compile corpus*, and then on *Compile*. This adds the information needed to use all of the functions.

Once the processing is complete, click on *OK*. Your corpus is now ready to use in Sketch Engine – to do this click on *Open in SkE*.

orpora 🄄 Create corpus	Corpora				
🔶 WebBootCaT		Corpus name	Language	Size	
onfiguration templates	Cambridge Academic	Corpus	English	412,966,498	
etch grammars	Cambridge Internatio	nal Corpus	English	1,358,129,340	
er groups	Cambridge Learner C	orpus Coded	English	25,542,115	
	Cambridge Learner C	orpus Question Papers	English	44,859	
lmin	Cambridge Learner C	orpus Uncoded	English	43,787,263	
cal administration	Cambridge Spoken Co	orpus	English	96,856,880	

Your corpus will now also appear on the opening page, underneath the other Cambridge corpora that are available to use.

Sketch users can create their own corpora up to a maximum of 1m words (it may be possible to increase this amount, on request – contact cdembry@cambridge.org for more info).

It is also possible to **share** corpora that you have built using WebBootCaT with other Sketch Engine users:

- click on the name of your new corpus from the main screen (shown above)
- click on *access privileges*
- select the level of access you would like the other person to have
- type their name into the box then select it from the list that appears and click *ok*.
- this user can now access the corpus you have built.

A note of caution! WebBootCaT gathers data from the web in an indiscriminate way. Because of this, occasionally unusual words or characters may appear in your concordance lines - you should bear this in mind when drawing conclusions from any results you find.

2. View options – advanced

View options allows you to change the type and amount of information that is displayed in your concordance lines. For a discussion of changing the references column and page size, please see Sketch Engine – Getting Started, available here: http://www.cambridge.org/sketch/help/

From time to time you may want to display additional information in your concordance lines.

For example, you may wish to see the Part of Speech tag that a particular word has been assigned.

Under *Attributes* a selection of information that can be displayed is shown. You can choose to show as many of these as you wish (although adding in more than one at a time can make your results difficult to read!)

	ius. Cambridge International	Corpus	
Concordance Word List Word Skotch	View options		
Thesaurus	Attributes	Structures	References
Sketch-Diff	word	<doc> 🗠</doc>	Token number
? Help on main menu	🗖 tag	<s></s>	doc.id
	🗌 lempos	<	doc.cupno
	emma 🗌 🔲 lemma	'gr	Written/spoken
? Help on Conc. menu			Genre
? Help on View Options	lemma_lc		Variety of English
View concordance		_	doc.title
Sample			doc.author
Filter	Display attributes		doc.publisher
Frequency Node tags	O For each token		doc.level
Node forms	KWIC tokens on	у	doc.ype
Doc IDs			
Text Types			
Collocations	Page size (number of	f lines): 20	
ConcDesc	KWIC Context size (r	number of cha	aracters): 60
Switch menu position			
	Sort good diction	ary examples	Number of lines to be sorted: 100
	🔲 Icon for one-click	sentence cop	oying
	🗌 Allow multiple line	es selection	
	XML template for one	e-click copying	a: 🔼 💌

Once selected, the attributes appear in the concordance separated from the words by a forward slash, as shown in the examples below:

- <u>Word</u> displays the words in the concordance.
- <u>Tag</u> displays the Part of Speech that has been assigned to each word.
- For example: *Girls/NNS are/VBP*
- <u>Lempos</u> displays both the lemma and the Part of Speech that a word belongs to. For example: *Girls/girl-n* are/be-v
- <u>Lemma</u> displays the lemma (or headword) that each word belongs to. For example: *Girls/girl are/be*
- <u>Lc</u> indicates the lowercase forms.
- For example: *Girls/girls are/are*
- <u>Lemma lc</u> displays the lowercase form of the lemma that each word belongs to. For example: *Girls/girl are/be*

A full list of the PoS tags can be found in the Cambridge Help: <u>http://www.cambridge.org/sketch/help/</u>

You can choose to display attributes for <u>each token</u> in your results (by selecting *display attributes for each token*) or for your <u>search term only</u> (by selecting *display attributes KWIC tokens only*).

Tags for the KWIC tokens only in a simple search for *chase* will return results like this:

It is also possible to display markers that indicate boundaries of different sorts using the *Structures* column. For example, sentence boundaries, <s>, can be seen in green an excerpt below:

arrive . </s><s> Intelligence analysts say that before U. S. forces **chased** Abu Mousab al-Zarqawi out of the city of Fallujah , he would the number in 1982 - who together spend almost as much money **chasing** birds as all Americans spend on movie tickets every year . </s> analysts , but Rogers could n't abide working with outsiders and **chased** off the talented people they trained as soon as they grew bold those hot stocks after they 've reached the ozone . </s> yes again [Bacon 's directorial debut was the 1996 TV movie Losing **Chase**], because I want to put some of his techniques to work :

3. Simple vs. multilevel sort

Sorting allows you to change the order in which your concordance is displayed. You might want to use *sort* to look at patterns within your results.

To sort the words in the concordance alphabetically to the **left**, **right**, on the **node**, or by **references** choose these option from the concordance menu.

To run more complex sorts, click on the *sort* button itself, situated on the lower left menu. This takes you to the screen shown below:

CAM	CAMBRIDGE Cambridge help Home Settings Log o				
UNIVE	RSITY PRESS	Search in Help			
user:	used tokens: 7,812,458 / 10,000,000 days left: unlimited Search	in Cambridge International Corpl 💙			
Concordance Word List Word Sketch Thesaurus Sketch-Diff ? Help on main menu ? Help on Conc. menu ? Help on Sort form View concordance Sample Filter	Simple Sort Attribute: word Sort key: OLeft context Node Right context Number of tokens to sort: 3 Ignore case Backward Sort Concordance Multilevel sort				
Frequency Node tags Node forms Doc IDs Text Types Collocations ConcDesc Switch menu position	Ignore case Backward Ignore case Backward Ignore case Backward Desition: 3L 2L Position: 1L Node 1R Sort Concordance	 third level (finally sort by) Attribute: word Ignore case Backward Backward 2L Position: 1L Node 1R 			

There are two types of searches that you can run:

- **Simple Sort** allows you to sort your concordance results by a feature that you select (e.g. by word, tag, genre or variety etc).
- **Multilevel Sort** allows you to sort in the same way, but to sort for more than one parameter, e.g. you can sort by word then by PoS tag within the same results.

To run a **Simple Sort**, select options from the top section of the sort screen. Use the dropdown box next to *Attribute* to select the feature you would like to sort by. For example, you can sort by *word, tag, source, or genre*.

You can then select whether you'd like to sort the concordance to the *left, right* or on the *node* (i.e. your initial search term). Choose the option you'd like to select from the buttons next to *Sort key*

You can select how many tokens are sorted e.g. choose 3, if you wanted to sort 3 words to the left. You can select to ignore case and you can also select to sort backwards (reverse the sort)

Note! If you choose to sort by e.g. tag, in order to see the actual tags in the results you must choose this option from the view options menu (look back to Section 2. for more information on this).

To run a **Multilevel Sort** you need to fill in the boxes in the bottom half of the sort screen. The choices are the same as the simple sort, but with a multilevel sort you can sort on 1, 2 or 3 levels – this means that you can sort within a sort.

For example, you can run a simple search for *fire* (i.e. all inflected forms, all PoS). Then, by using Multilevel Sort you can then sort on 3 levels. For example you could sort by:

- First level part-of-speech of the node (so, sorting *fire* as a noun from *fire* as a verb).
- Second level the word (i.e. the actual word form of the node, so sorting the verbs *fire* from *fired*, etc).
- Third level word to the left of the node (1L).

These options would look like this:

Multilevel sort					
O first level (Sort by) Attribute: tag	○ second ✓ Attribute:	level (then sort by) word	 third Attribute 	level (finally sort by) : word	>
Ignore case 🗆 Backward 🗖	Ignore case	Ignore case 🔲 Backward 🔲		Ignore case 🗆 Backward 🗖	
3L ▲ 2L Position: 1L Node 1R ✓	Position:	3L ▲ 2L 1L Node 1R ✓	Position:	3L A 2L 1L Node 1R	
Sort Concordance					

If you're sorting by more than one level, you should click the button at the highest level, i.e. to search by first <u>and second level</u>, click the second level button.

Note! If you choose to sort by e.g. *tag*, or *lempos* you must choose this option from the view options menu in order to see the actual tags or lemma-pos displayed in the results.

4. Filter

The **filter** function allows you to narrow down your search to <u>include</u> or <u>exclude</u> a particular word or phrase in your analysis.

To run a filter, first make a concordance of your search term. From the lower left hand side menu, click on the *filter* button. This takes you to the screen shown below:

CAMBRIDGE Cambridge help UNIVERSITY PRESS					
user: Claire Dembry used tok	ens: 7,812,458 / 10,000,000 days left: unlimited				
Concordance Word List Word Sketch	Concordance Filter				
Thesaurus	Filter: O positive O negative				
Sketch-Diff	Selected token:				
Prep of fian field	Search Span: from -5 to 5				
	Query Type: Word Form 🔽				
? Help on Conc. menu	Word Form: PoS: unspecified 💌 Match case: 🗆				
? Help on Filter form	Filter Concordance				
View concordance					
Sample Filter					

In the main panel you have various options for filtering your concordance.

You have the following choices:

- **Positive:** includes a word or phrase in the subsequent results.
- **Negative:** excludes a word or phrase from the subsequent results.
- **First** and **Last:** use this option if your original query was for more than one word you can determine which token (first or last) is used when calculating distance using the Search Span.
- Search Span: set how many tokens before or after your node word that the filter should apply to.

The *query type* box works in the same way as a 'normal' concordance query – you are able to run simple, lemma, word form, phrase and CQL searches. Here you should enter the word or phrase that you would like to include or exclude in your filter. You can also match the PoS and case of the word you enter.

5. Frequency

Frequency allows you to investigate the most common features of your search (e.g. most frequent words, tags, genres etc).

To look at frequency results for the **node tags**, **node forms**, **doc IDs** and **Text Types**, choose these options from the concordance menu.

To obtain more complex frequency results, click on the *frequency* button itself, situated on the lower left menu. This takes you to the screen shown below:

UNIVERSITY PRESS					
user: Claire Dembry used	tokens: 7,812,458 / 10,000,000 days left: unlimited				
Concordance Word List Word Sketch Thesaurus Sketch-Diff ? Help on main monu	Multilevel frequency distribution				
? Help on Conc. menu ? Help on frequency form View concordance Sample Filter Frequency Node tags Node forms Doc IDs	 				
Text Types Collocations ConcDesc	Text Type frequency distribution				
Switch menu position	Frequency limit: 0 Include categories with no hits:				
	doc.id doc.id doc.cupno Source Written/spoken Genre Vanety of English doc.name doc.title				

There are two types of frequency distributions that you can run:

- **Text Type frequency distribution** is straightforward to use it allows you to show how your search term is distributed through the texts in the Corpus. You may find, for example, that a word like *police* appears significantly more often in newspaper texts than in other text types. To display results for more than one category, hold ctrl and clicking.
- **Multilevel frequency distribution** allows you to look at the frequency of more than one parameter, e.g. you can sort by the most frequent word then by the most frequent PoS tag within the same results.

This multilevel approach works on the same basis as multilevel sorting (described in Section 3. earlier). We can explore multilevel frequency distribution further by looking at an example:

- Run a simple search for *travel*.
- From the concordance select *Frequency* from the left hand menu
- In the Multilevel frequency distribution section, choose *word* and select *node*. This will display the frequencies of e.g. *travels, travelling, travelled*.
- To search for the most frequent words <u>following</u> these forms, we need to add a second level:
- Under *second level* set the attribute at *word* and set the position at *1R* (word one position to right of node word)
- To search for the most frequent words <u>following</u> both the node and the token after the node, we need to add a third level:
- Under *third level* set the attribute at *word* and set the position at *2R*.
- Remember that if you're sorting by more than one level, you should click the button at the highest level, i.e. to search by first <u>and second level</u>, click the second level button.
- Press *Make frequency list* to show your results.

6. Building a subcorpus

You can use the *subcorpus* function to explore only specific parts of the corpus. To create a subcorpus:

- Click on *text types* to display the text types options
- Next to Subcorpus, click on *create new* (shown in red)
- (Subcorpora you have already built appear in the dropdown menu)
- In the next screen, choose a name for your subcorpus and enter it into the box.
- Select the text types you wish to include in upper subcorpus by ticking the appropriate boxes.
- Click on *create subcorpus*

Your subcorpus is now added to your list. There is no limit to the number of subcorpora that you can build.

7. Working with subcorpora

To run a **Concordance** using only your subcorpus, select *text types* and choose your subcorpus from the dropdown menu. Then complete your concordance search in the usual way – the data returned will be from your subcorpus only.

To generate a **Frequency list** using only your subcorpus, click on *frequency list* and then choose your subcorpus from the dropdown list. The most frequent words in your subcorpus will then be listed.

To run a **Word sketch** using only your subcorpus, click on *Word sketch* then on *advanced options* from the lower left hand menu. Then select your subcorpus from the dropdown list and then run the Word sketch as usual.

You may also want to use your subcorpus to run a **Keyword analysis**. This allows you to compare two corpora or subcorpora in order to find out the words that are comparatively the most different. For example, a keyword analysis that compared a 'business' subcorpus to, for example, the Cambridge International Corpus would indicate the words that are most frequently found in the business subcorpora but not in the CIC (and are therefore more typical in that domain).

To run a **Keyword analysis** using your subcorpus, click on *frequency list*. Under *Keywords* select your subcorpus from the first drop down menu and choose a corpus to compare it to from the second drop down menu.

Query Type: Query:	Simple 💌	
Text Types Sub	ocorpus:	v into create new
Variety of Er	nglish Written/spoken	Genre
🗖 Am	🗖 spoken	Documents
🗖 Aus	🗖 written	Fiction
Br	Select All	lournals

8. Building wordlists

It is possible to build word frequency list using Sketch Engine. You can do this by corpus or by subcorpus. To run a frequency list, click 'word list' from the left hand side menu. The word list screen is shown below:



If you would like to make a word list from a subcorpus that you have already defined, select it from the drop down menu, or click 'create new' to create a new subcorpus to work with.

You can search for all words, or you can enter patterns, e.g. all words ending in "ing" or starting with "pre". These patterns should be entered using the

format: . *ing or pre. *

You can also change the minimum frequency level of your list, in order to exclude low frequency items.

Whitelists and Blacklists allow you to include and exclude words from your wordlist.

- White lists search only for those words in the list, and ranks them with respect to how frequent they are in the corpus.
- Black lists exclude words from the corpus frequency list. Blacklists can be used to exclude stopwords (for a list of commonly used stopwords that you may wish to use, click here: <u>http://cup.sketchengine.co.uk/stopwords/english/</u>)

Once you're happy with your selections, click on make wordlist. You wordlist will then be displayed. You can then choose 'save' to e.g. save it to Excel.

9. Comparing corpora using keywords

You can create a keyword list in order to compare one corpus to another. A keyword list displays those words in your corpus that are most different to the other corpus you compare it to (the reference corpus).

To create a keyword list:

- click on 'wordlist' from the left hand side menu options. You will then see the screen shown in 8 above.
- Choose your corpus or subcorpus, and then under 'output type' (in the lower half of the screen) select 'keywords'
- Select your reference corpus (and, if you wish, subcorpus). In order to get a picture of how your corpus is different to others, if may be useful to compare it to a very general corpus in this case, the CIC is a good choice.
- Click *make Word List* to display your results

10. Word Sketch – Advanced options

Along with using Word Sketch in the usual way (as outlined in the Getting Started guide), a number of 'expert options' are available.

vora sketch Entry Form	
Lemma:	
Part of speech: adjective 💙	
Advanced options	
Subcorpus:	None (whole corpus) 💌 info create new
Minimum frequency:	auto
Minimum salience:	0.0
Maximum number of items in a grammatical relation: Sort collocations	25
according to:	Salience ○ Raw frequency
Cluster collocations	
Structure word sketch by gramrels	
Minimum similarity between cluster items:	0.15
Select gramrels:	♥ and/or ♥ object ♥ object_of ♥ subject ♥ subject_of ♥ adj_subject_of ♥ adj_subject ♥ predicate_of ♥ predicate ♥ pro_object ♥ pro_subject ♥ a_modifier ♥ modifier ♥ modifier ♥ modifier ♥ modifies ♥ possessed ♥ possessor ♥ pro_possessor ♥ wh_comp ♥ infin_comp ♥ ing_omp ♥ passive ♥ reflexive ♥ it+ ♥ pp_%s ♥ np_adj_comp ♥ np_adj_comp of ♥ adj_comp ♥ adj_comp_of ♥ patt_intrans ♥ part_trans ♥ part_%s_obj
Show Word Sketch Save Options	

Open these options by selecting 'word sketch' then clicking 'expert options' from the lower left hand menu:

Here, the main difference is that you can search by a particular subcorpus, by selecting it from the dropdown.

You can also choose to see the grammatical relations you wish, by selecting/deselecting them from the lower part of the screen.

11. Uploading your own text files to Sketch Engine

Most typically, Sketch Engine is used in order to examine Cambridge corpus resources. However, it is possible to upload your own text files and analyse them in the same way.

Your own data will be POS tagged thus allowing you to use many of the same functions usually can in Sketch (such as e.g. Work Sketch). However, you are not able to use any metadata associated with you text files.

Creating a Corpus:

Before adding any text files, you need to create a new location for them in Sketch Engine.

Go to the main Sketch Engine homepage, and select Create Corpus (underneath the My Corpora heading). The following screen will appear:

CAMBRIDGE UNIVERSITY PRESS								
user: Sarah Grieves used to	cens: 19 / 1,000,000 days left: unlimited							
Corpora Create corpus WebBootCat	Create corpus: Step 1/3							
Configuration templates Sketch grammars Subcorpus definitions User groups	Corpus ID Unique identifier of your corpus. May only contain letters, numbers, underscores and hyphens.							
Admin Local administration	Info							
Support Help Report a bug	Language Cancel Next >							

Type in a name for your corpus in the Corpus ID and Corpus name boxes. The Corpus name will appear on your main Sketch Engine screen.

The Info box is optional, but it may be helpful to add more detail about your data for yourself and/or other users. Select the language of your corpus (typically English) from the drop-down box, and click Next.

On the next screen, choose TreeTagger for English and click Next.

On the next screen, choose English PennTB-TreeTagger MCD 2.1 and click Finish.

Uploading text files:

You can now upload text files to your new corpus. Choose Add new file (underneath the corpus name, and also in the left hand menu). The following screen will appear:

Upload from disk	O Browse	
Download from location	O http://	
lse file or directory on the server	O	
	Snow files in subdirectories FTP to cup.sketchengine.co.uk at port 10021 to upload files. Use the same user name and password as for logging into this web interface.	
Paste text	0	

This is where you can select the files you want to upload. If you want to upload text files saved in a folder on your computer, select Upload from disk and click on Browse... to select the file.

You can also paste text directly into the box shown above, by selecting the Paste text button.

Once you have selected your option and selected/pasted your file, click Next, and when the file preview appears click Finish. This will take you back to the original screen for your corpus, where you can select Add new file again and repeat the process.

Each file has to be uploaded separately.

Compiling your corpus:

Once you have added all of your files, click on Compile corpus. You are unable to work with your corpus until it has been compiled.

Test2									
+ A	dd new file / 💠 Ad	d data from we	b using WebB	ootCa1 / 🕓 Com	ipile corpus / 🔍 Ope	n in SkE			
#	Original file	Plain text	Vertical	Tokens 🛡	Owner				
1	001.txt	4	«	6	Sarah Grieves	🤜 🖉 🗙			
2	002.txt	4	«	7	Sarah Grieves	🤜 🖉 🗙			
з	003.txt	1	 Image: A set of the set of the	6	Sarah Grieves	🤍 🖉 🗶			

On the next screen, make sure English PennTB-TreeTagger MCD 2.1 is selected (it is the default). Then click on Compile at the bottom of the screen.

When the message at the top of the screen changes from Corpus is busy to Processing done!, click on OK.

Complie corpus		100%
Word sketches		100%
Thospurus		100%
niesaurus Cubaamaana	· · · · · · · · · · · · · · · · · · ·	100 /
Subcorpora	*	
		🖌 O
>>> 1:"V?"	' "RB.?"{0,2} 3:"IN" [tag="DT.?" tag="PP\\$"]{0,1} "CD"{0,2	2} [tag="J 🔨
no results		
>>> 1:"V?'	' "RB.?"{0,2} 2:"JJ.?" [tag!="NN.?.?" & tag!="CC" & tag!="	'JJ.?"]
>>> 2:[tam='	=nst["CD"=nst["\$999"=nst["TC"=sst] {[.0}"TC9" "WT":5 ["*.W	"RB.2"Ita
no results	III] O. IN IDI (0,1) [oug- DI (oug- IIV (oug- ob (oug-	1011 100
>>> 2:"\.*"	"RB.?"{0,3} 3:"IN" "PDT"{0,1} [tag="DT" tag="PP\$" tag="CD)" tag="RB
estimated: 1	- 1	
matched: 1		
[20111216-15:47:	38] lexicon (/var/lib/ske/ca/user_data/UUP121/manatee/les	st2/lempos
mkumap: scream : mkuman: temporal	v files renamed	
Word sketches co	mpiled successfully!	
Compiling thesau	urus	
gramrels: 15		
words: U words with count	- N 20• 0	
items: 0 (OMB)	, , 20. 0	
word max id: 0		
data to process	OGB	
data to process.		
plan size: 1		

Your corpus is now ready to use. You can open it from the Sketch Engine homepage under My corpora; click on the small magnifying glass logo next to the corpus name:

CAMBRI UNIVERSITY	DGE PRESS						
user: Dr. Claire Dembry use	d tokens: 1,003,535 / 10,00	00,000 days left: ur	nlimited				
Corpora 💠 Create corpus	Corpora						
🔶 WebBootCaT	Corpus	name	Language	Tokens	Words		
Configuration templates	Cambridge Academic C	orpus	English	412,966,498	n/a	Ţ	۹
Sketch grammars	Cambridge Internation	Cambridge International Corpus			n/a	Ţ	۹
subcorpus definicions Iser arouns	Cambridge Internation	English	1,395,787,439	1,161,826,160		۲	
soon groups	Cambridge Learner Cor	rpus 2	English	2,588,563	2,186,362	Ţ	۹
Admin	Cambridge Learner Cor	pus Coded	English	27,964,598	23,751,249	Ţ	۹
	Cambridge Learner Cor	English	48,050,495	41,472,367	Ţ	۹	
ocal administration	Cambridge Spoken Cor	pus	English	86,372,443	72,844,759	Ţ	۲
Current	English Profile Project C	Corpus	English	874,245	770,242	Ţ	۲
lelp					Show 1 mor	e co	rpoi
kepurt a buy	My corpora						
	Corpus ID	Corpus name	Langua	ge Tokens			
	Spoken_Academic	Spoken Academic	English	1,003,544	1 🔍 🔍 🔶 🖉	> >	6
			·	🕈 Crea	te corpus / 💠 We	вво	otCa

Sharing your corpus with other Sketch Engine users:

It is possible to share any corpus you upload with other Sketch Engine users.

To do this, from the main screen click on the name of the corpus you have uploaded.

The files that you have uploaded will then appear on the screen, along with a number of options displayed in the left hand menu, as shown below:

CAMBRIDGE UNIVERSITY PRESS									
user: Dr. Claire Dembry used tokens: 1,003,535 / 10,000,000 days left: unlimited									
Corpora	Spo + Ac	oken Academic	n web using V	VebBootCa	T / 🕲 Compile	e corpus / 🔍 Open in	ı SkE		
Sketch grammars	#	Original file	Plain text	Vertical	Tokens 🔻	Owner			
Subcorpus definitions	1	LNGLEC08.txt	×	×	10,707	Dr. Claire Dembry	~		*
User groups	2	LNGLEC10.txt	×	×	/,4/1	Dr. Claire Dembry	~		*
	3	LNGLEC09.txt	×	×	7,883	Dr. Claire Dembry	~	Ø	*
Corpus	4	NC000002_MARGE_1.txt	4	×	30,113	Dr. Claire Dembry	-	<i>.</i>	×
💠 Add new file	5	NC00004.txt	4	<	26,650	Dr. Claire Dembry	-	0	×
💠 Add web data	6	NC0000072txt	1	«	20,211	Dr. Claire Dembry		<i>.</i>	×
(BootCaT)	7	NC0000082txt	4	 Image: A set of the set of the	6,818	Dr. Claire Dembry	,	<i>></i>	×
Complie Corpus	8	EAPLEC01.txt	1	«	8,621	Dr. Claire Dembry	,	٨	×
Extract keywords	9	NC0000086txt	st and a start of the start of	~	7,349	Dr. Claire Dembry	,	٨	×
Configure corpus	10	EAPLEC02.txt	Ś	~	7,779	Dr. Claire Dembry	,	٨	×
🖋 Change sketch	11	NC000093txt	1	v	21,095	Dr. Claire Dembry	,	٨	×
grammar	12	EAPLEC03.txt	1	v	8,504	Dr. Claire Dembry	,	٨	×
Set subcorpus	13	NC0000932.txt	1	 Image: A set of the set of the	21,095	Dr. Claire Dembry	,	٨	×
definitions	14	EAPLEC04.txt	1	v	12,189	Dr. Claire Dembry	,	٨	×
Download corpus	15	EAPLEC05.txt	1	√	7,288	Dr. Claire Dembry		٨	×
Access privileges		LEC06.txt	1	√	11,169	Dr. Claire Dembry	,	٨	×
🔍 View logs	17	EAPLEC08.txt	1	~	6,352	Dr. Claire Dembry	,	٥	×

To share your corpus, click on Access privileges, as shown above.

There are three different levels of access that you can assign to users. These access levels are explained on the screen.

To add a user start to type their name in the appropriate box and then select them from the dropdown list. Then click ok. The user(s) you have selected will now have access to your corpus. Your corpus will appear underneath any corpus that they have added themselves, as shown below:

CAMBRI UNIVERSITY	DGE PRESS					
user: Dr. Claire Dembry used	i tokens: 1,003,535 / 10,000	0,000 days left: u	nlimited			
Corpora 🔶 Create corpus	Corpora					
+ WebBootCaT	Corpus n	ame	Language	Tokens	Words	
Configuration templates	Cambridge Academic Co	rpus	English	412,966,498	n/a	,
Sketch grammars Subcornus definitions	Cambridge International	l Corpus	English	1,358,129,340	n/a	🔍 🔍
User groups	Cambridge International	l Corpus (v2.1)	English	1,395,787,439	1,161,826,160	🛡 🍕
	Cambridge Learner Corp	ous 2	English	2,588,563	2,186,362	🔍 🍕
Admin	Cambridge Learner Corp	ous Coded	English	27,964,598	23,751,249	,
Local administration	Cambridge Learner Corp	English	48,050,495	41,472,367	🔍 🔍	
LUCAI auministration	Cambridge Spoken Corpus		English	86,372,443	72,844,759	🔍 🍕
Support	English Profile Project Co	orpus	English	874,245	770,242	🔍 🍕
Help Report a bug	My corpora				Show 1 mor	e corpora
	Corpus ID	Corpus name	Langu	age Tokens		
	Spoken_Academic	Spoken Academic	English	1,003,54	4 🛡 🍭 🔶 🕯	1 ×
				💠 Crea	ite corpus / 💠 We	bBootCaT
	Other users' corpo	ora				
	Corpus ID	Corpus name	Language	Size Ow	ner	
	CUP_business_emails	Business email	English	18,771 Laura Mu	rkin 🔍 🤍	+ 🖉
	TimesOfIndia	TimesOfIndia	English	380,750 Dominic (ilennon 🛛 🛡 🍭	🔶 🤌