Cambridge English Corpus

# Cambridge Sketch Engine

## Advanced Help (v.2.0)

Last updated: February 2012

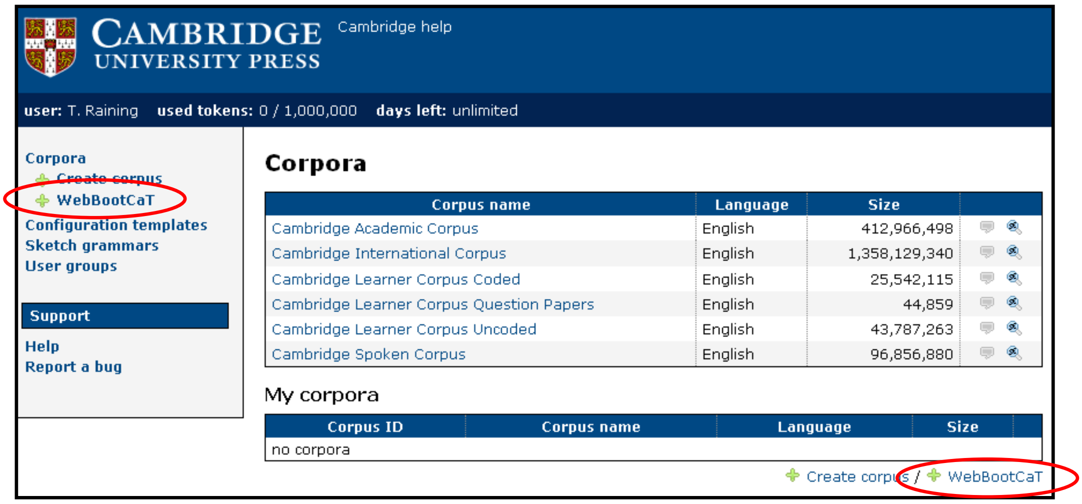
## Contents:

## 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:



**user:** T. Raining **used tokens:** 0 / 1,000,000 **days left:** unlimited

**Corpora**

| Corpus name | Language | Size |
|---|---|---|
| Cambridge Academic Corpus | English | 412,966,498 |
| Cambridge International Corpus | English | 1,358,129,340 |
| Cambridge Learner Corpus Coded | English | 25,542,115 |
| Cambridge Learner Corpus Question Papers | English | 44,859 |
| Cambridge Learner Corpus Uncoded | English | 43,787,263 |
| Cambridge Spoken Corpus | English | 96,856,880 |

My corpora

| Corpus ID | Corpus name | Language | Size |
|---|---|---|---|
| no corpora | | | |

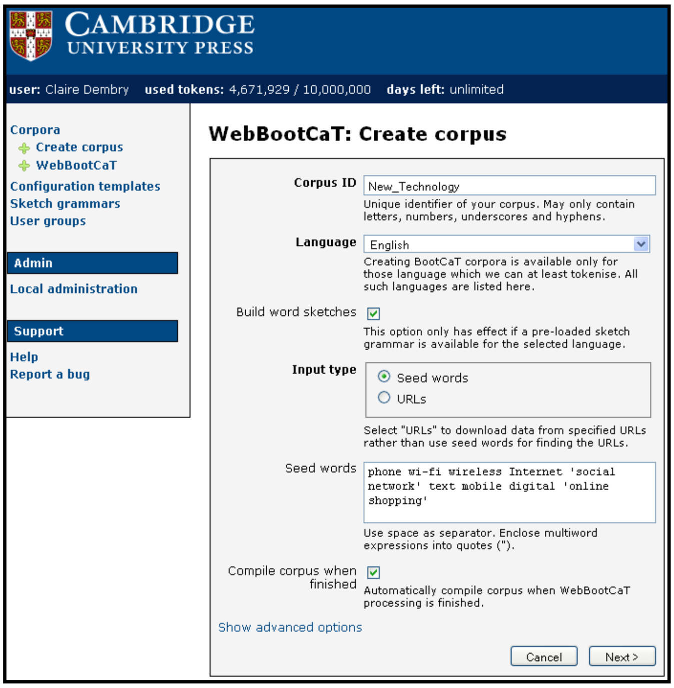
This takes you to a screen where you can enter the name of your corpus, the language of your corpus (English) and your seed words.

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.



**user:** Claire Dembry **used tokens:** 4,671,929 / 10,000,000 **days left:** unlimited

**WebBootCaT: Create corpus**

**Corpus ID** New_Technology
Unique identifier of your corpus. May only contain letters, numbers, underscores and hyphens.

**Language** English
Creating BootCaT corpora is available only for those language which we can at least tokenise. All such languages are listed here.

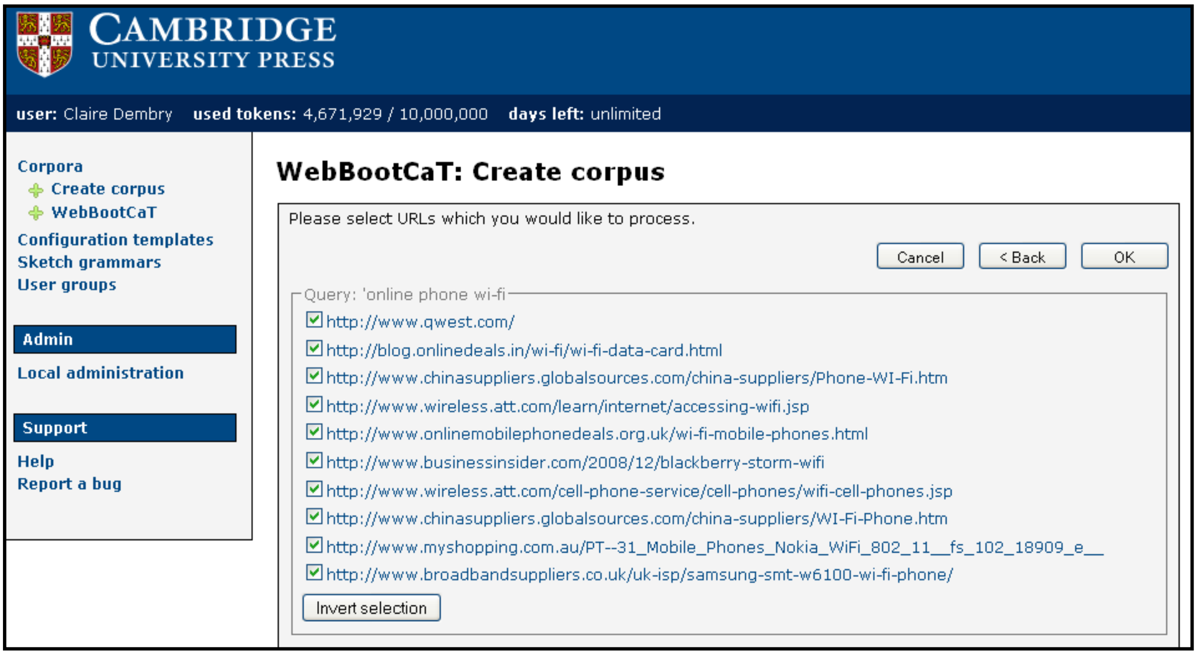
Build word sketches ☑
This option only has effect if a pre-loaded sketch grammar is available for the selected language.

**Input type** ● Seed words ○ URLs
Select "URLs" to download data from specified URLs rather 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.

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.



**WebBootCaT: Create corpus**

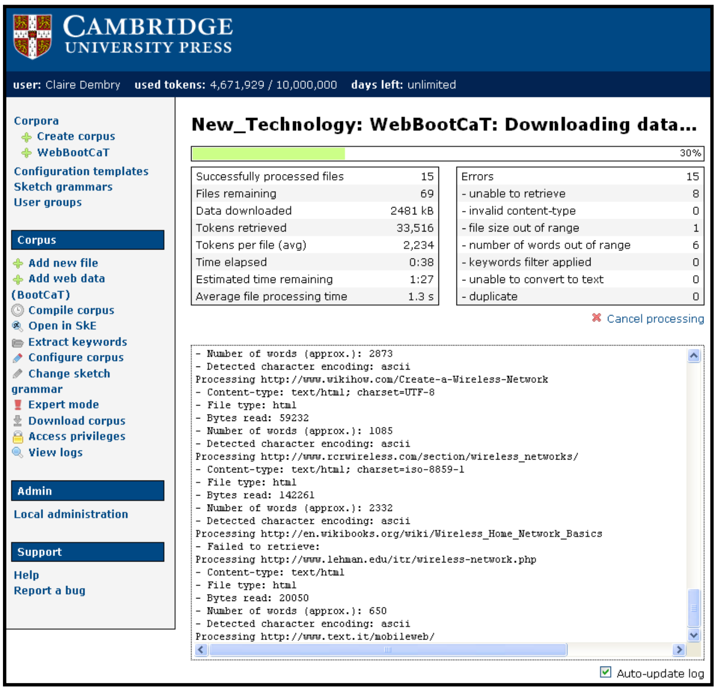
Please select URLs which you would like to process.

Query: 'online phone wi-fi

- ☑ http://www.qwest.com/
- ☑ http://blog.onlinedeals.in/wi-fi/wi-fi-data-card.html
- ☑ http://www.chinasuppliers.globalsources.com/china-suppliers/Phone-WI-Fi.htm
- ☑ http://www.wireless.att.com/learn/internet/accessing-wifi.jsp
- ☑ http://www.onlinemobilephonedeals.org.uk/wi-fi-mobile-phones.html
- ☑ http://www.businessinsider.com/2008/12/blackberry-storm-wifi
- ☑ http://www.wireless.att.com/cell-phone-service/cell-phones/wifi-cell-phones.jsp
- ☑ http://www.chinasuppliers.globalsources.com/china-suppliers/WI-Fi-Phone.htm
- ☑ http://www.myshopping.com.au/PT--31_Mobile_Phones_Nokia_WiFi_802_11__fs_102_18909_e__
- ☑ http://www.broadbandsuppliers.co.uk/uk-isp/samsung-smt-w6100-wi-fi-phone/

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.



**New_Technology: WebBootCaT: Downloading data...**

| | |
|---|---|
| Successfully processed files | 15 |
| Files remaining | 69 |
| Data downloaded | 2481 kB |
| Tokens retrieved | 33,516 |
| Tokens per file (avg) | 2,234 |
| Time elapsed | 0:38 |
| Estimated time remaining | 1:27 |
| Average file processing time | 1.3 s |

| | |
|---|---|
| Errors | 15 |
| - unable to retrieve | 8 |
| - invalid content-type | 0 |
| - file size out of range | 1 |
| - number of words out of range | 6 |
| - keywords filter applied | 0 |
| - unable to convert to text | 0 |
| - duplicate | 0 |

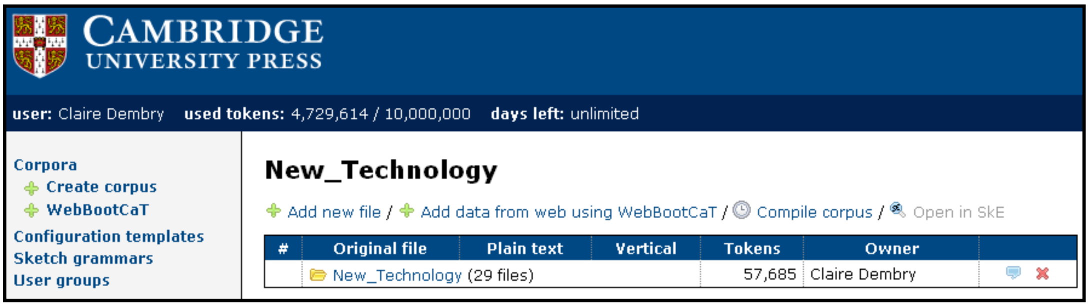
```
- Number of words (approx.): 2873
- Detected character encoding: ascii
Processing http://www.wikihow.com/Create-a-Wireless-Network
- Content-type: text/html; charset=UTF-8
- File type: html
- Bytes read: 59232
- Number of words (approx.): 1085
- Detected character encoding: ascii
Processing http://www.rcrwireless.com/section/wireless_networks/
- Content-type: text/html; charset=iso-8859-1
- File type: html
- Bytes read: 142261
- Number of words (approx.): 2332
- Detected character encoding: ascii
Processing http://en.wikibooks.org/wiki/Wireless_Home_Network_Basics
- Failed to retrieve:
Processing http://www.lehman.edu/itr/wireless-network.php
- Content-type: text/html
- File type: html
- Bytes read: 20050
- Number of words (approx.): 650
- Detected character encoding: ascii
Processing http://www.text.it/mobileweb/
```

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:



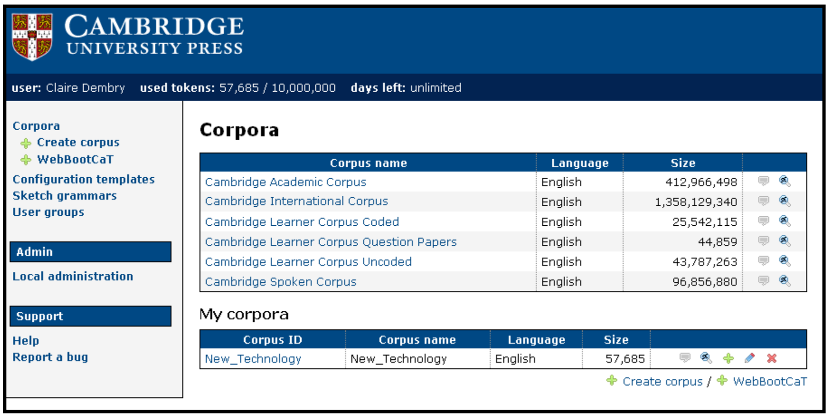
user: Claire Dembry used tokens: 4,729,614 / 10,000,000 days left: unlimited

New_Technology

| # | Original file | Plain text | Vertical | Tokens | Owner | |
|---|---|---|---|---|---|---|
| | 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 compile 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*.



user: Claire Dembry used tokens: 57,685 / 10,000,000 days left: unlimited

Corpora

| Corpus name | Language | Size |
|---|---|---|
| Cambridge Academic Corpus | English | 412,966,498 |
| Cambridge International Corpus | English | 1,358,129,340 |
| Cambridge Learner Corpus Coded | English | 25,542,115 |
| Cambridge Learner Corpus Question Papers | English | 44,859 |
| Cambridge Learner Corpus Uncoded | English | 43,787,263 |
| Cambridge Spoken Corpus | English | 96,856,880 |

My corpora

| Corpus ID | Corpus name | Language | Size |
|---|---|---|---|
| New_Technology | New_Technology | English | 57,685 |

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.

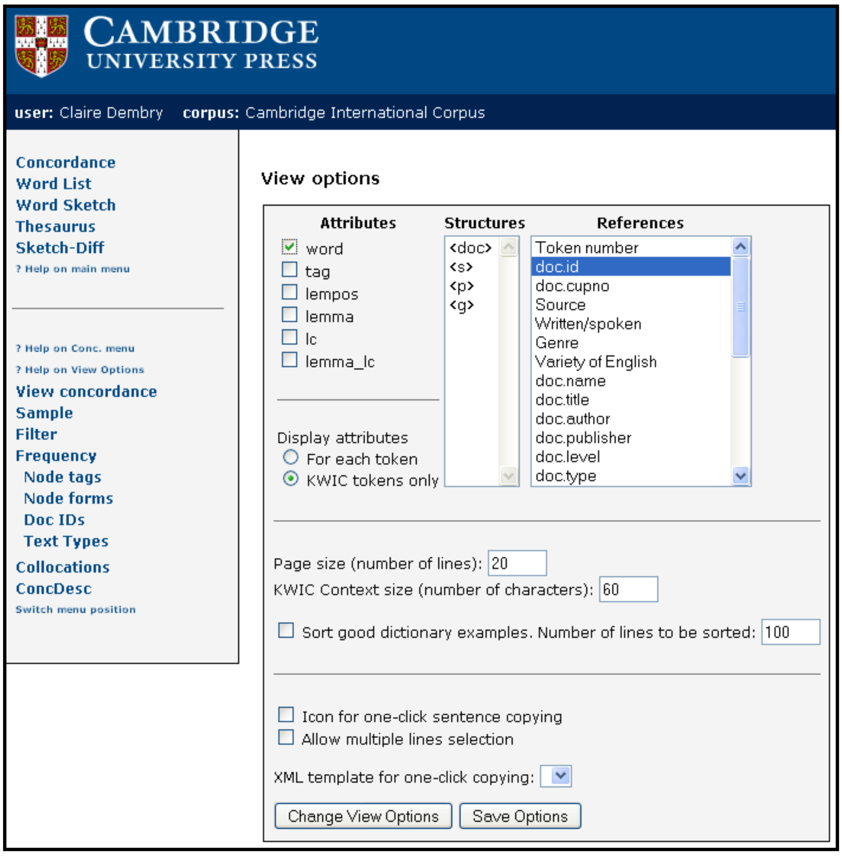
## 2. View options – advanced

*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!)



user: Claire Dembry corpus: Cambridge International Corpus

View options

| Attributes | Structures | References |
|---|---|---|
| ☑ word | <doc> | Token number |
| ☐ tag | <s> | doc.id |
| ☐ lempos | <p> | doc.cupno |
| ☐ lemma | <g> | Source |
| ☐ lc | | Written/spoken |
| ☐ lemma_lc | | Genre |
| | | Variety of English |
| | | doc.name |
| | | doc.title |
| | | doc.author |
| | | doc.publisher |
| | | doc.level |
| | | doc.type |

Display attributes
○ For each token
● KWIC tokens only

Page size (number of lines): 20
KWIC Context size (number of characters): 60

☐ Sort good dictionary examples. Number of lines to be sorted: 100

☐ Icon for one-click sentence copying
☐ Allow multiple lines selection

XML template for one-click copying:

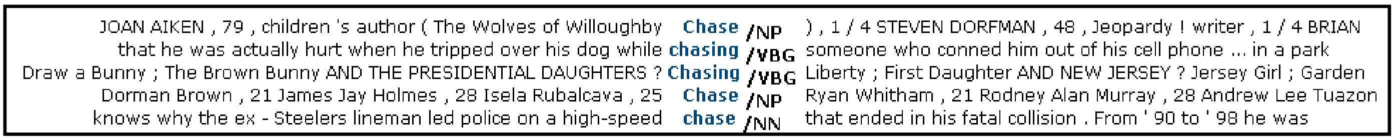
Once selected, the attributes appear in the concordance separated from the words by a forward slash, as shown in the examples below:

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

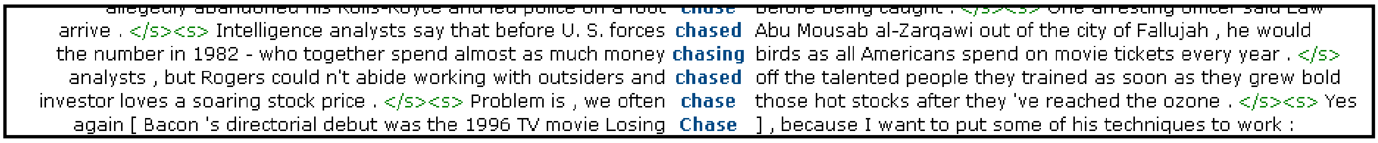
You can choose to display attributes for each token in your results (by selecting *display attributes for each token*) or for your search term only (by selecting *display attributes KWIC tokens only*).

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



```
      JOAN AIKEN , 79 , children 's author ( The Wolves of Willoughby  Chase /NP    ) , 1 / 4 STEVEN DORFMAN , 48 , Jeopardy ! writer , 1 / 4 BRIAN
       that he was actually hurt when he tripped over his dog while chasing /VBG   someone who conned him out of his cell phone ... in a park
Draw a Bunny ; The Brown Bunny AND THE PRESIDENTIAL DAUGHTERS ? Chasing /VBG   Liberty ; First Daughter AND NEW JERSEY ? Jersey Girl ; Garden
      Dorman Brown , 21 James Jay Holmes , 28 Isela Rubalcava , 25  Chase /NP    Ryan Whitham , 21 Rodney Alan Murray , 28 Andrew Lee Tuazon
   knows why the ex - Steelers lineman led police on a high-speed  chase /NN    that ended in his fatal collision . From ' 90 to ' 98 he was
```

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:



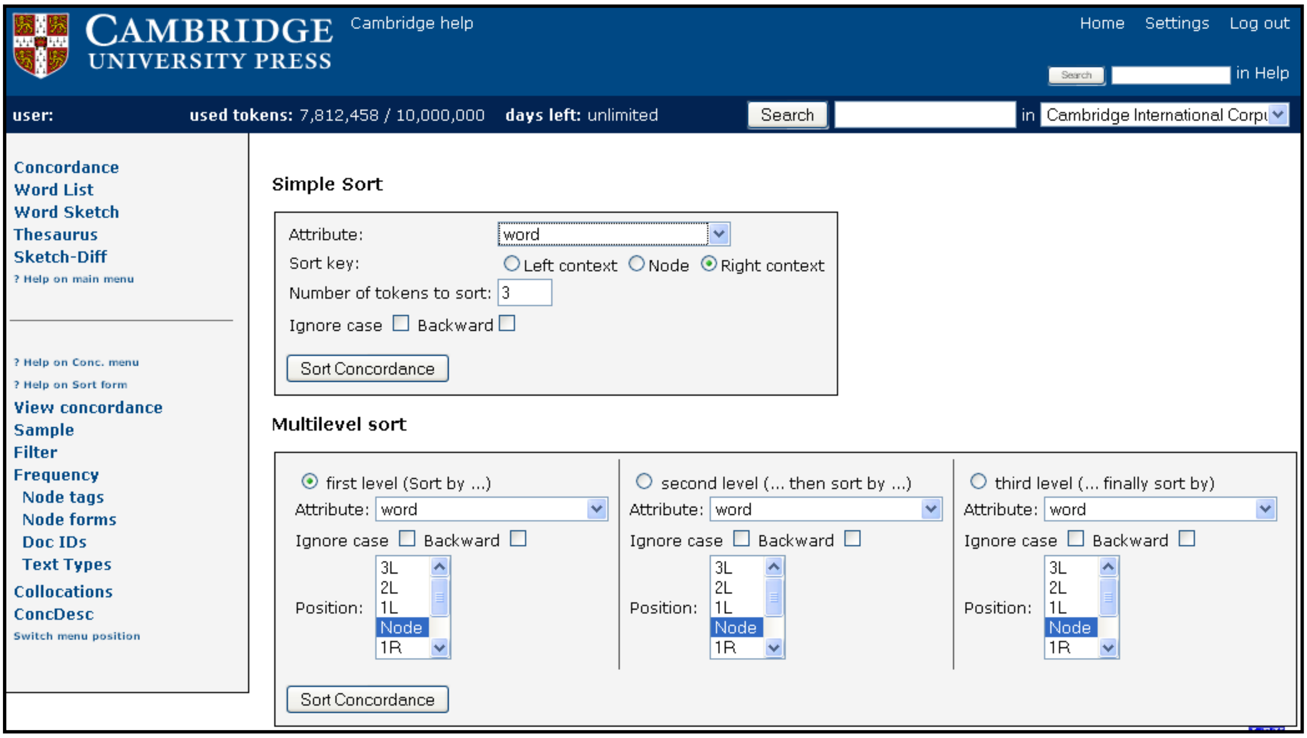
```
      allegedly abandoned his Rolls-Royce and led police on a foot  chase   before being caught . </s><s> One arresting officer said Law
   arrive . </s><s> Intelligence analysts say that before U. S. forces  chased   Abu Mousab al-Zarqawi out of the city of Fallujah , he would
  the number in 1982 - 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
investor loves a soaring stock price . </s><s> Problem is , we often  chase    those hot stocks after they 've reached the ozone . </s><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:



Simple Sort

Attribute: word

Sort key: ○ Left context ○ Node ● Right context

Number of tokens to sort: 3

Ignore case □ Backward □

Sort Concordance

Multilevel sort

● first level (Sort by ...)
Attribute: word
Ignore case □ Backward □
Position: 3L 2L 1L Node 1R

○ second level (... then sort by ...)
Attribute: word
Ignore case □ Backward □
Position: 3L 2L 1L Node 1R

○ third level (... finally sort by)
Attribute: word
Ignore case □ Backward □
Position: 3L 2L 1L Node 1R

Sort Concordance

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)

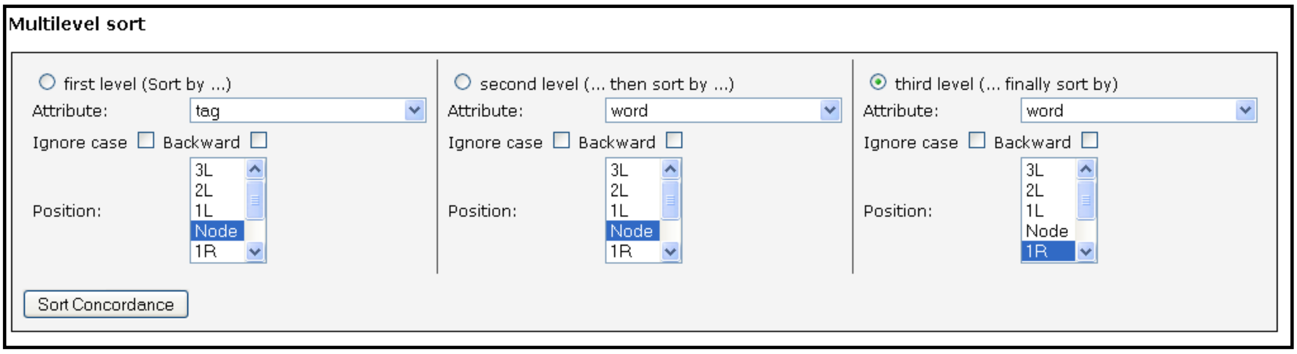
Note! If you choose to sort by e.g. tag, in order to see the actual tags in the results you must choose this option from the view options menu (look back to Section 2. for more information on this).

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**

○ first level (Sort by ...)
Attribute: tag
Ignore case □ Backward □
Position: 3L, 2L, 1L, Node (selected), 1R

○ second level (... then sort by ...)
Attribute: word
Ignore case □ Backward □
Position: 3L, 2L, 1L, Node (selected), 1R

● third level (... finally sort by)
Attribute: word
Ignore case □ Backward □
Position: 3L, 2L, 1L, Node, 1R (selected)

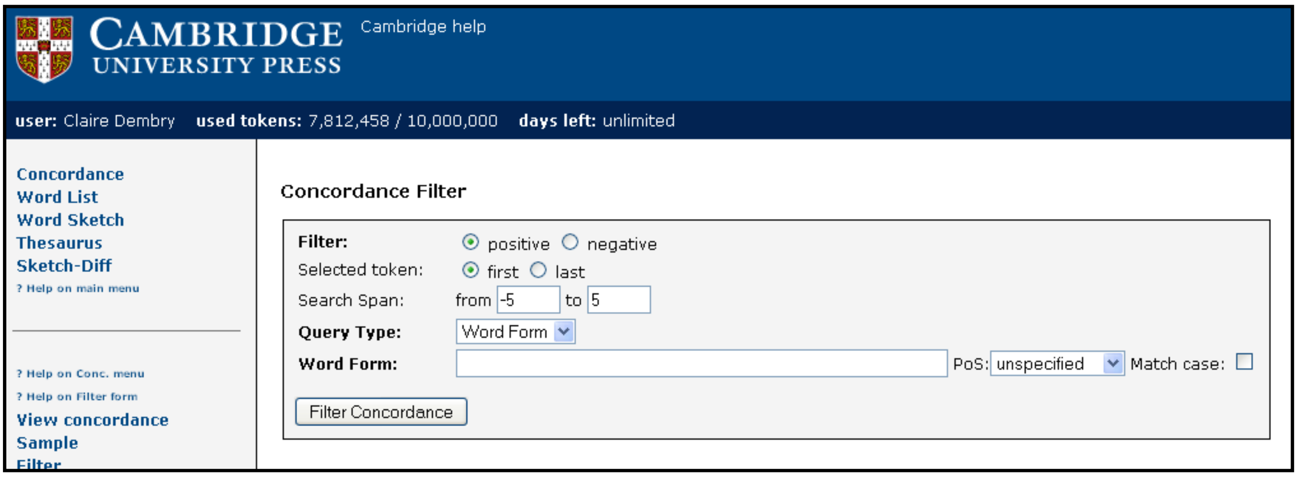
If you're sorting by more than one level, you should click the button at the highest level, i.e. to search by first and second level, click the second level button.

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 include or exclude 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:



user: Claire Dembry used tokens: 7,812,458 / 10,000,000 days left: unlimited

Concordance Filter

Filter: ● positive ○ negative
Selected token: ● first ○ last
Search Span: from -5 to 5
Query Type: Word Form
Word Form: PoS: unspecified Match case: □

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.

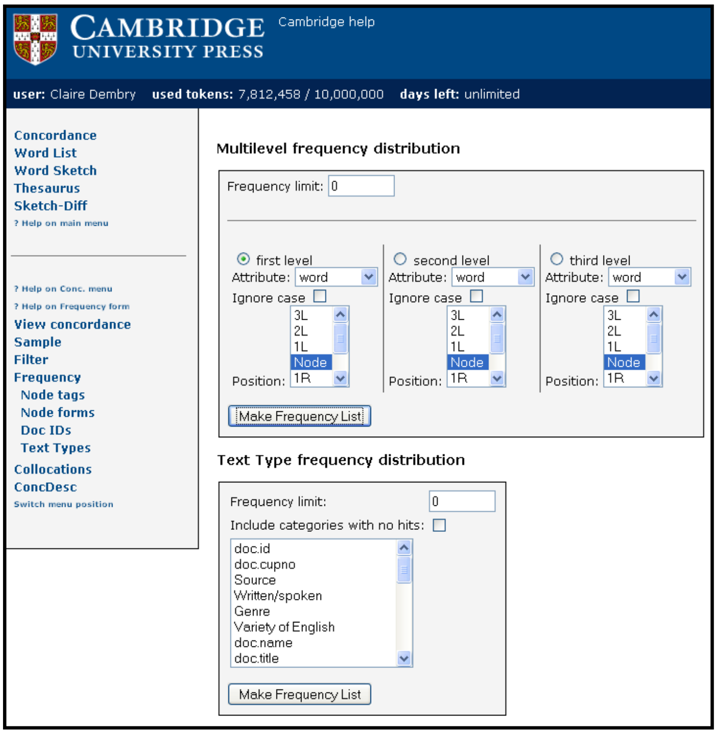
The *query type* box works in the same way as a 'normal' concordance query – you are able to run simple, lemma, word form, phrase and CQL searches. Here you should enter the word or phrase that you would like to include or exclude in your filter. You can also match the PoS and case of the word you enter.

## 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:



**user:** Claire Dembry **used tokens:** 7,812,458 / 10,000,000 **days left:** unlimited

**Multilevel frequency distribution**

Frequency limit: 0

| ● first level | ○ second level | ○ third level |
|---|---|---|
| Attribute: word | Attribute: word | Attribute: word |
| Ignore case □ | Ignore case □ | Ignore case □ |
| 3L<br>2L<br>1L<br>Node | 3L<br>2L<br>1L<br>Node | 3L<br>2L<br>1L<br>Node |
| Position: 1R | Position: 1R | Position: 1R |

**Text Type frequency distribution**

Frequency limit: 0

Include categories with no hits: □

doc.id
doc.cupno
Source
Written/spoken
Genre
Variety 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.

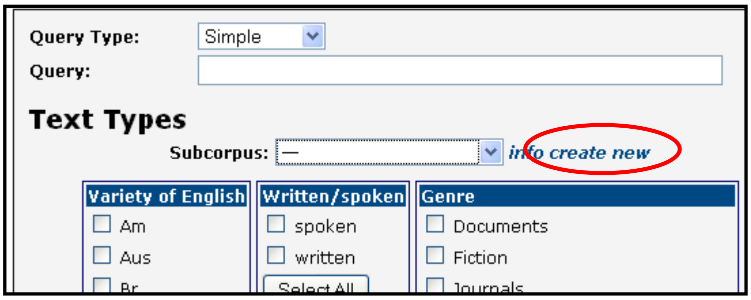
This multilevel approach works on the same basis as multilevel sorting (described in Section 3. earlier). We can explore multilevel frequency distribution further by looking at an example:

- 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 following 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 following 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 and second level, 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 your subcorpus by ticking the appropriate boxes.
- Click on *create subcorpus*



Query Type: Simple

Query:

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.

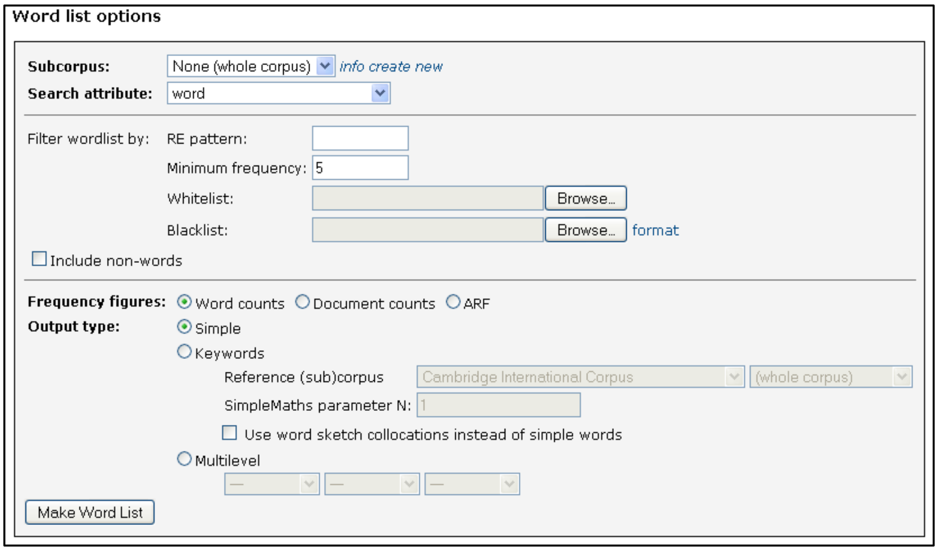
To run a **Word sketch** using only your subcorpus, click on *Word sketch* then on *advanced options* from the lower left hand menu. Then select your subcorpus from the dropdown list and then run the Word sketch as usual.

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.

## 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:



**Word list options**

**Subcorpus:** None (whole corpus) *info create new*
**Search attribute:** word

Filter wordlist by: RE pattern:
Minimum frequency: 5
Whitelist: Browse...
Blacklist: Browse... format
□ Include non-words

**Frequency figures:** ● Word counts ○ Document counts ○ ARF
**Output type:** ● Simple
○ Keywords
Reference (sub)corpus: Cambridge International Corpus (whole corpus)
SimpleMaths parameter N: 1
□ Use word sketch collocations instead of simple words
○ Multilevel
— — —

Make Word List

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: http://cup.sketchengine.co.uk/stopwords/english/)

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

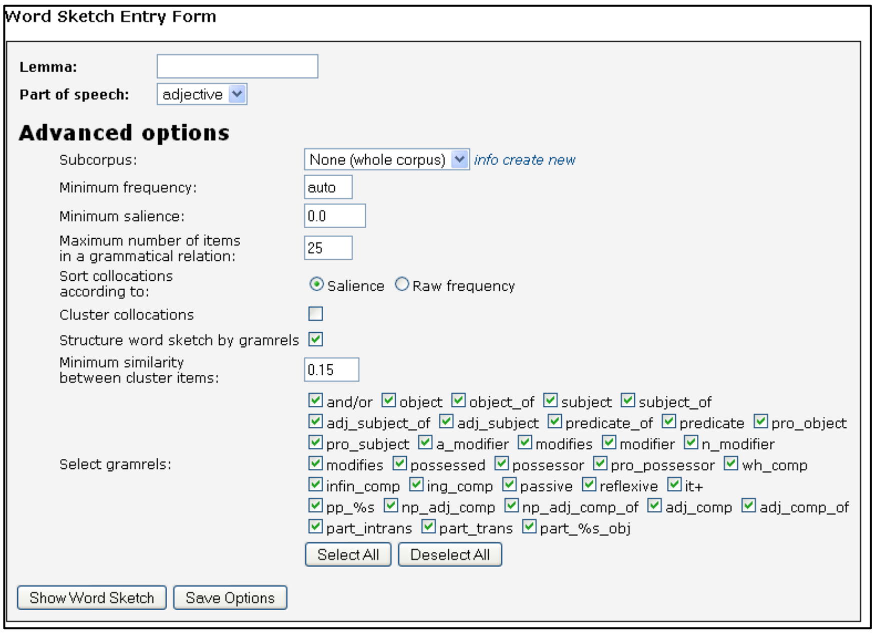
You can create a keyword list in order to compare one corpus to another. A keyword list displays those words in your corpus that are most different to the other corpus you compare it to (the reference corpus).

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.



Word 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: 25

Sort collocations 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 ☑ modifies ☑ modifier ☑ n_modifier ☑ modifies ☑ possessed ☑ possessor ☑ pro_possessor ☑ wh_comp ☑ infin_comp ☑ ing_comp ☑ passive ☑ reflexive ☑ it+ ☑ pp_%s ☑ np_adj_comp ☑ np_adj_comp_of ☑ adj_comp ☑ adj_comp_of ☑ part_intrans ☑ part_trans ☑ part_%s_obj

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

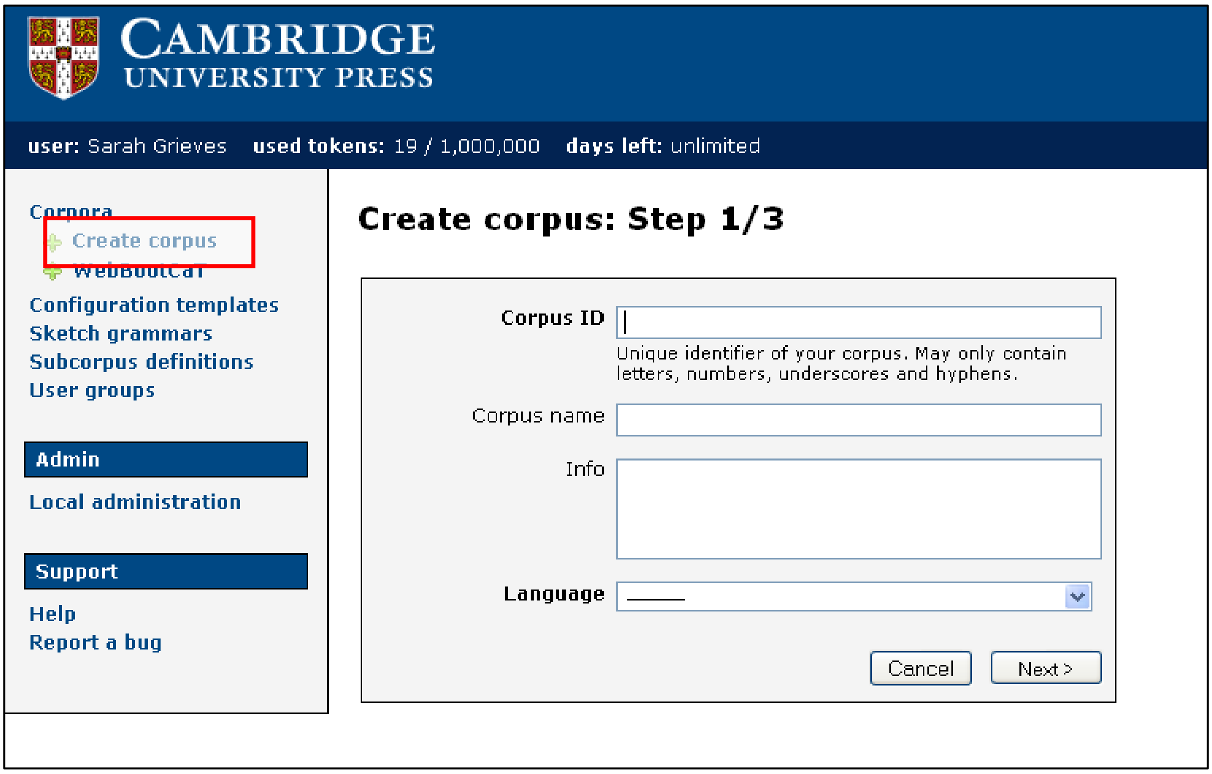
Most typically, Sketch Engine is used in order to examine Cambridge corpus resources. However, it is possible to upload your own text files and analyse them in the same way.

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:



**user:** Sarah Grieves **used tokens:** 19 / 1,000,000 **days left:** unlimited

**Create corpus: Step 1/3**

**Corpus ID**

Unique identifier of your corpus. May only contain letters, numbers, underscores and hyphens.

Corpus name

Info

**Language** ———

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.

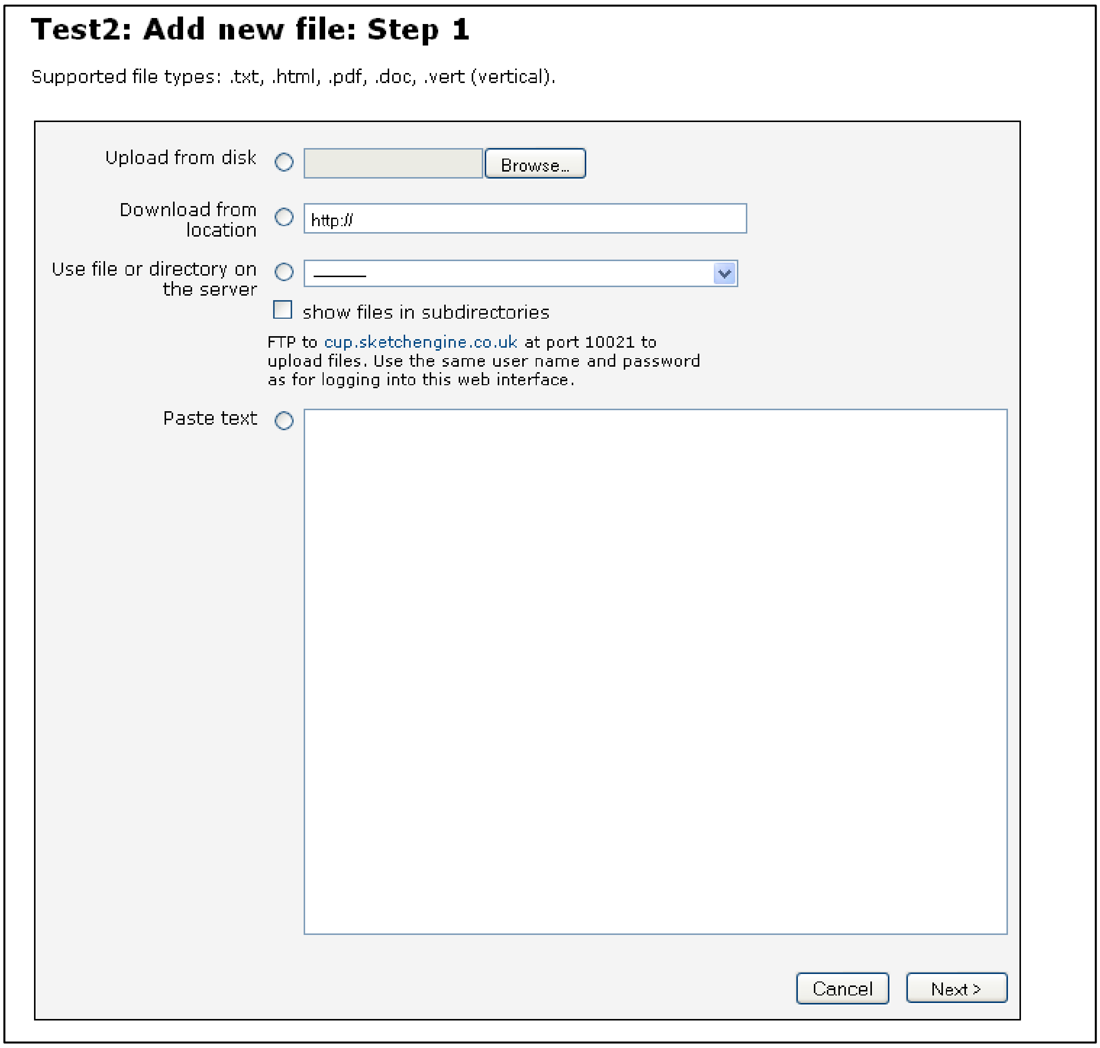
The Info box is optional, but it may be helpful to add more detail about your data for yourself and/or other users. Select the language of your corpus (typically English) from the drop-down box, and click Next.

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:



**Test2: Add new file: Step 1**

Supported file types: .txt, .html, .pdf, .doc, .vert (vertical).

Upload from disk ○ Browse...

Download from location ○ http://

Use file or directory on the server ○ ———

□ show 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 ○

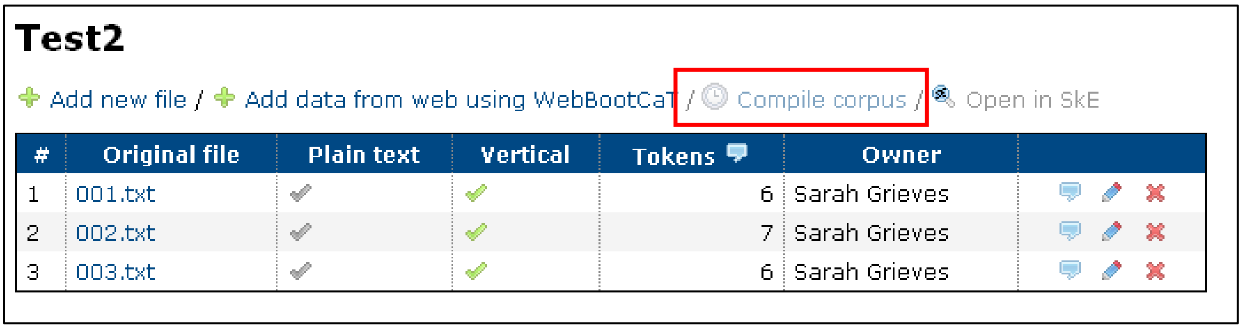
This is where you can select the files you want to upload. If you want to upload text files saved in a folder on your computer, select Upload from disk and click on Browse... to select the file.

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.



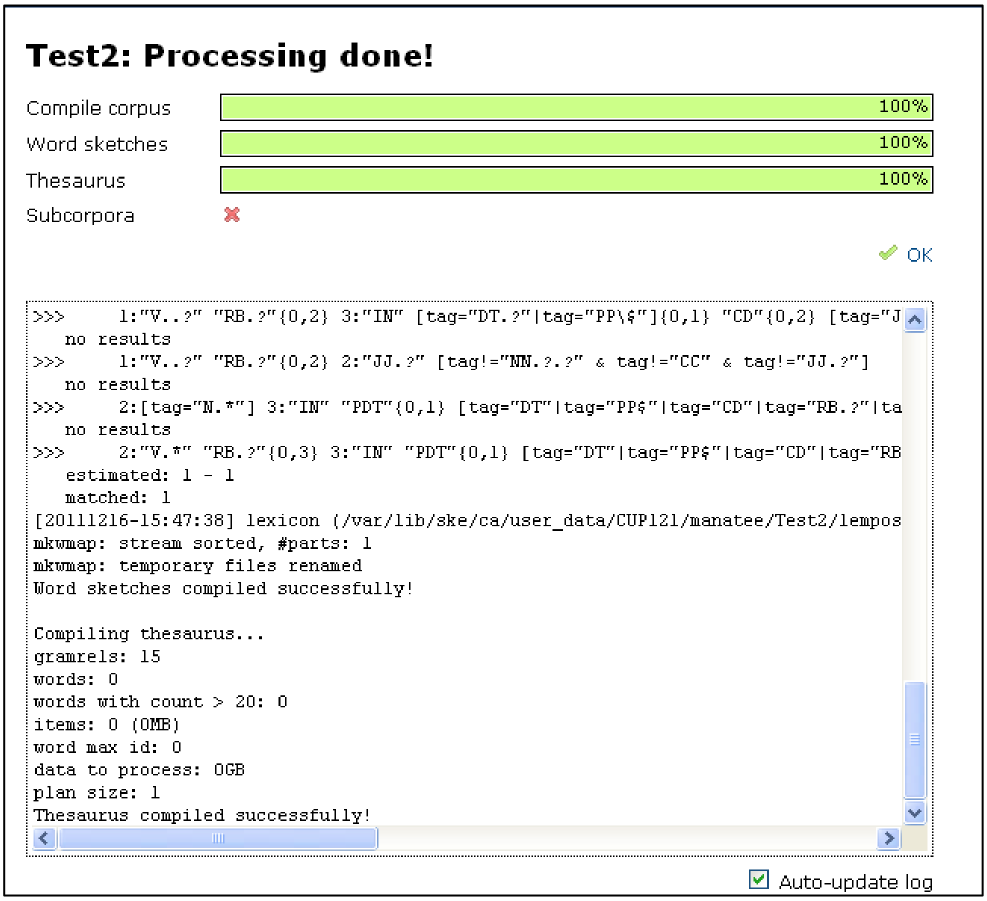
**Test2**

✚ Add new file / ✚ Add data from web using WebBootCaT / Compile corpus / Open in SkE

| # | Original file | Plain text | Vertical | Tokens | Owner | |
|---|---|---|---|---|---|---|
| 1 | 001.txt | ✓ | ✓ | 6 | Sarah Grieves | |
| 2 | 002.txt | ✓ | ✓ | 7 | Sarah Grieves | |
| 3 | 003.txt | ✓ | ✓ | 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.



**Test2: Processing done!**

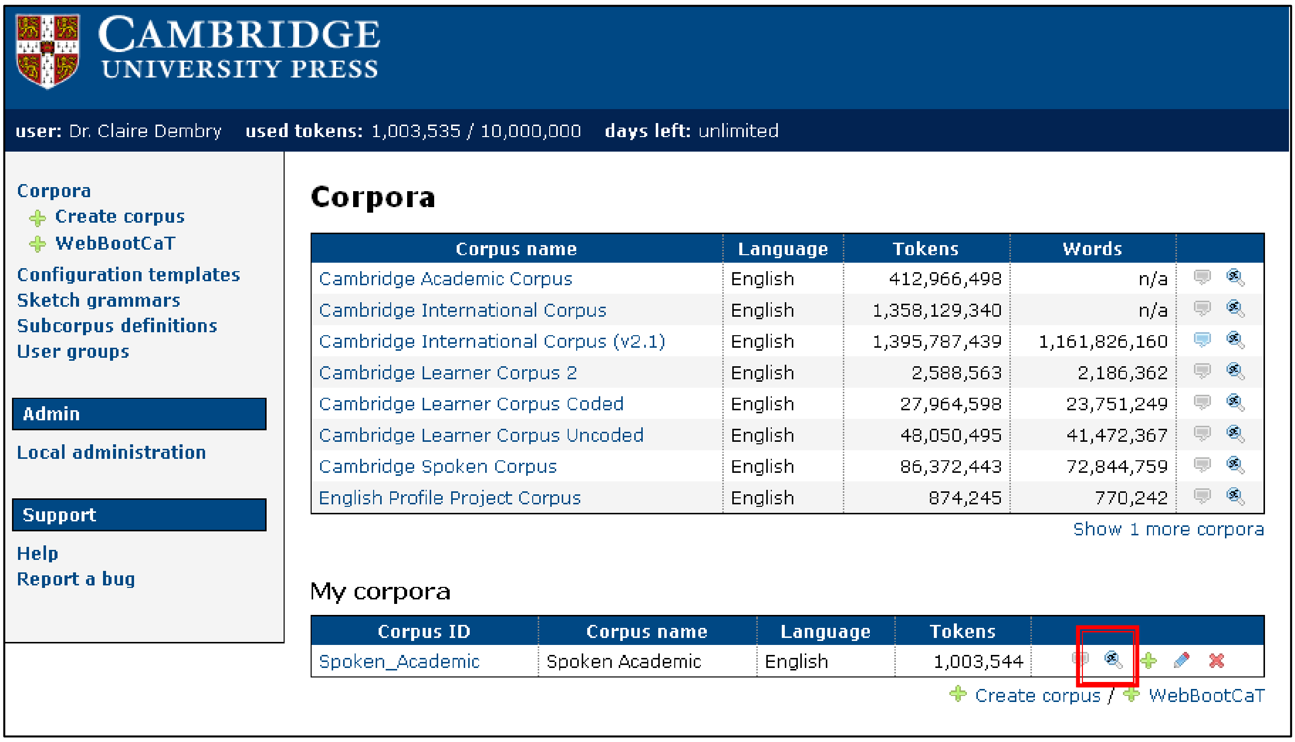
| | |
|---|---|
| Compile corpus | 100% |
| Word sketches | 100% |
| Thesaurus | 100% |
| Subcorpora | ✗ |

```
>>>     1:"V..?" "RB.?"{0,2} 3:"IN" [tag="DT.?"|tag="PP\$"]{0,1} "CD"{0,2} [tag="J
   no results
>>>     1:"V..?" "RB.?"{0,2} 2:"JJ.?" [tag!="NN.?.?" & tag!="CC" & tag!="JJ.?"]
   no results
>>>     2:[tag="N.*"] 3:"IN" "PDT"{0,1} [tag="DT"|tag="PP$"|tag="CD"|tag="RB.?"|ta
   no results
>>>     2:"V.*" "RB.?"{0,3} 3:"IN" "PDT"{0,1} [tag="DT"|tag="PP$"|tag="CD"|tag="RB
   estimated: 1 - 1
   matched: 1
[20111216-15:47:38] lexicon (/var/lib/ske/ca/user_data/CUP121/manatee/Test2/lempos
mkwmap: stream sorted, #parts: 1
mkwmap: temporary files renamed
Word sketches compiled successfully!

Compiling thesaurus...
gramrels: 15
words: 0
words with count > 20: 0
items: 0 (0MB)
word max id: 0
data to process: 0GB
plan size: 1
Thesaurus compiled successfully!
```

☑ Auto-update log

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:



CAMBRIDGE UNIVERSITY PRESS

**user:** Dr. Claire Dembry **used tokens:** 1,003,535 / 10,000,000 **days left:** unlimited

**Corpora**

| Corpus name | Language | Tokens | Words |
|---|---|---|---|
| Cambridge Academic Corpus | English | 412,966,498 | n/a |
| Cambridge International Corpus | English | 1,358,129,340 | n/a |
| Cambridge International Corpus (v2.1) | English | 1,395,787,439 | 1,161,826,160 |
| Cambridge Learner Corpus 2 | English | 2,588,563 | 2,186,362 |
| Cambridge Learner Corpus Coded | English | 27,964,598 | 23,751,249 |
| Cambridge Learner Corpus Uncoded | English | 48,050,495 | 41,472,367 |
| Cambridge Spoken Corpus | English | 86,372,443 | 72,844,759 |
| English Profile Project Corpus | English | 874,245 | 770,242 |

Show 1 more corpora

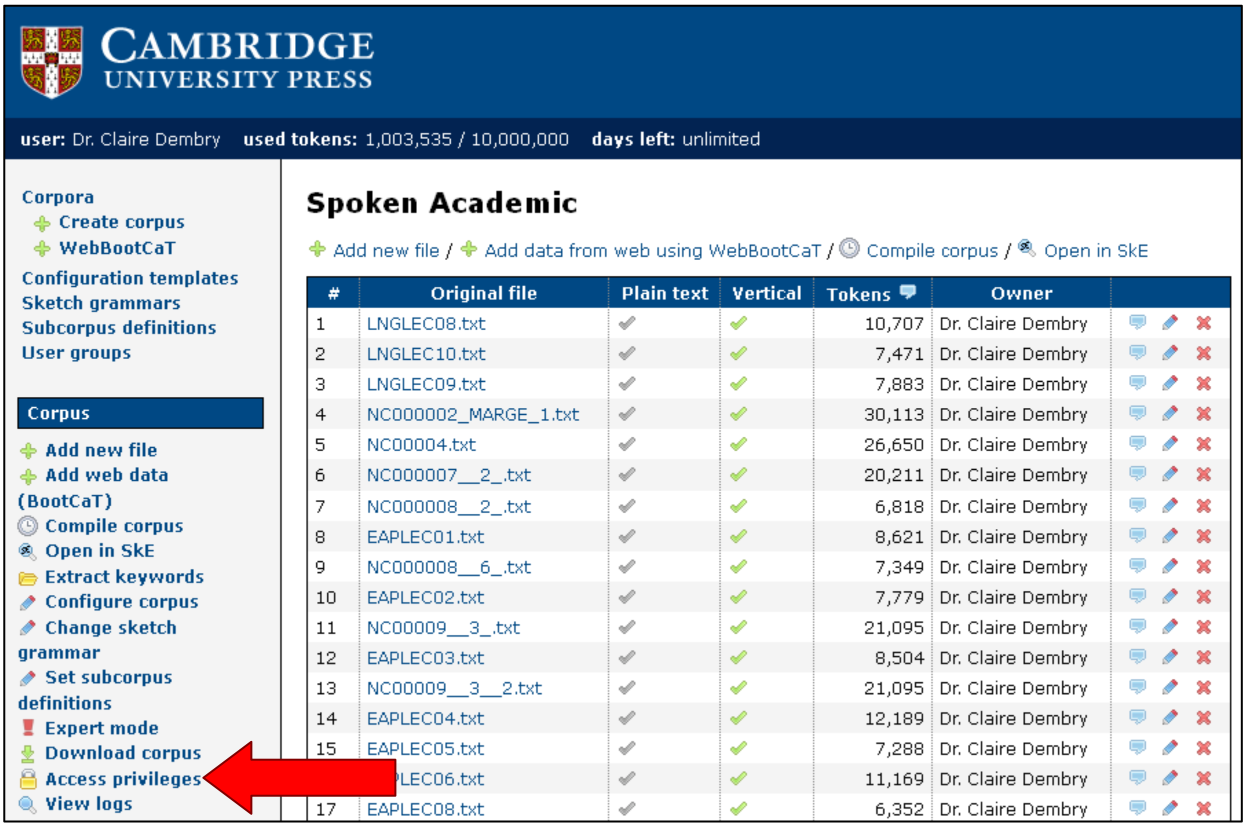
My corpora

| Corpus ID | Corpus name | Language | Tokens |
|---|---|---|---|
| Spoken_Academic | Spoken Academic | English | 1,003,544 |

**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

**Spoken Academic**

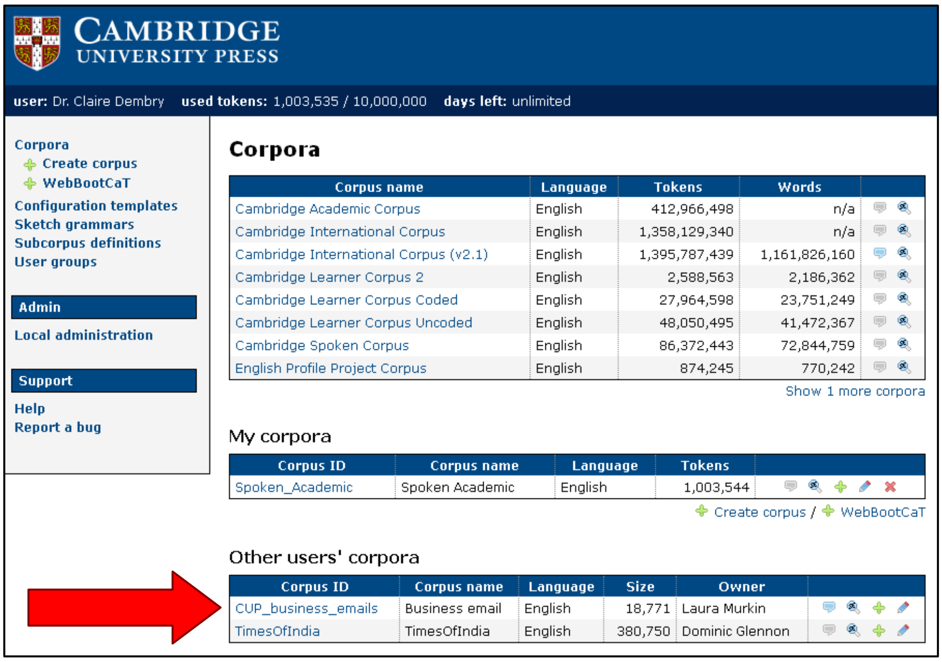
Add new file / Add data from web using WebBootCaT / Compile corpus / Open in SkE

| # | Original file | Plain text | Vertical | Tokens | Owner |
|---|---|---|---|---|---|
| 1 | LNGLEC08.txt | ✓ | ✓ | 10,707 | Dr. Claire Dembry |
| 2 | LNGLEC10.txt | ✓ | ✓ | 7,471 | Dr. Claire Dembry |
| 3 | LNGLEC09.txt | ✓ | ✓ | 7,883 | Dr. Claire Dembry |
| 4 | NC000002_MARGE_1.txt | ✓ | ✓ | 30,113 | Dr. Claire Dembry |
| 5 | NC00004.txt | ✓ | ✓ | 26,650 | Dr. Claire Dembry |
| 6 | NC000007__2_.txt | ✓ | ✓ | 20,211 | Dr. Claire Dembry |
| 7 | NC000008__2_.txt | ✓ | ✓ | 6,818 | Dr. Claire Dembry |
| 8 | EAPLEC01.txt | ✓ | ✓ | 8,621 | Dr. Claire Dembry |
| 9 | NC000008__6_.txt | ✓ | ✓ | 7,349 | Dr. Claire Dembry |
| 10 | EAPLEC02.txt | ✓ | ✓ | 7,779 | Dr. Claire Dembry |
| 11 | NC00009__3_.txt | ✓ | ✓ | 21,095 | Dr. Claire Dembry |
| 12 | EAPLEC03.txt | ✓ | ✓ | 8,504 | Dr. Claire Dembry |
| 13 | NC00009__3__2.txt | ✓ | ✓ | 21,095 | Dr. Claire Dembry |
| 14 | EAPLEC04.txt | ✓ | ✓ | 12,189 | Dr. Claire Dembry |
| 15 | EAPLEC05.txt | ✓ | ✓ | 7,288 | Dr. Claire Dembry |
| | LEC06.txt | ✓ | ✓ | 11,169 | Dr. Claire Dembry |
| 17 | EAPLEC08.txt | ✓ | ✓ | 6,352 | Dr. Claire Dembry |

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:



CAMBRIDGE UNIVERSITY PRESS

**user:** Dr. Claire Dembry **used tokens:** 1,003,535 / 10,000,000 **days left:** unlimited

**Corpora**

| Corpus name | Language | Tokens | Words |
|---|---|---|---|
| Cambridge Academic Corpus | English | 412,966,498 | n/a |
| Cambridge International Corpus | English | 1,358,129,340 | n/a |
| Cambridge International Corpus (v2.1) | English | 1,395,787,439 | 1,161,826,160 |
| Cambridge Learner Corpus 2 | English | 2,588,563 | 2,186,362 |
| Cambridge Learner Corpus Coded | English | 27,964,598 | 23,751,249 |
| Cambridge Learner Corpus Uncoded | English | 48,050,495 | 41,472,367 |
| Cambridge Spoken Corpus | English | 86,372,443 | 72,844,759 |
| English Profile Project Corpus | English | 874,245 | 770,242 |

My corpora

| Corpus ID | Corpus name | Language | Tokens |
|---|---|---|---|
| Spoken_Academic | Spoken Academic | English | 1,003,544 |

Other users' corpora

| Corpus ID | Corpus name | Language | Size | Owner |
|---|---|---|---|---|
| CUP_business_emails | Business email | English | 18,771 | Laura Murkin |
| TimesOfIndia | TimesOfIndia | English | 380,750 | Dominic Glennon |