Cambridge English Corpus

Cambridge Sketch Engine

Using the Learner Corpus (1.1)

Last updated: February 2012

This guide outlines the functions found in Sketch Engine that you can use to investigate the Learner Corpus. It builds on the functions outlined in the Getting Started guide and complements the Advanced Help guide.

This guide outlines only the procedures needed for running particular searches and queries; it does not give information on how you can use your corpus results for ELT, although suggestions about this can be found under *Using the Corpus in ELT* on the Cambridge Help page.

If you have any other queries, suggestions and/or feedback, please email corpus@cambridge.org

Contents

Outlined below are the areas covered in this guide, with a short note giving a further explanation. Click on the link to jump to that section.

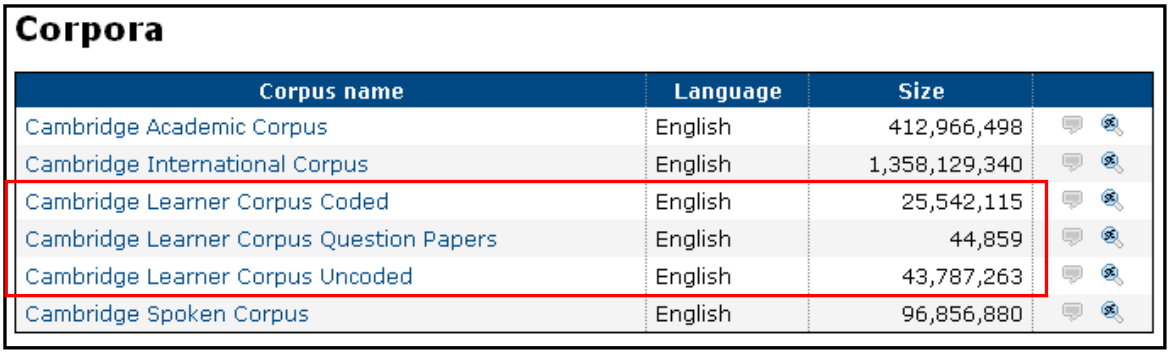
1. Overview

The Cambridge Learner Corpus (CLC) is a 45m word corpus of student responses to ESOL exams. In Sketch Engine, the CLC is broken down into two main parts: the error coded learner corpus (CLC coded) and the uncoded learner corpus (CLC uncoded) - both these are discussed in this guide.

In the CLC coded, the errors that are found in each exam script have been marked using a tag showing the type of error. We can then use these error codes to search for particular error types within the corpus. The error coding system is described in more detail in 4.1.

In addition to this the learner responses, a corpus of ESOL examination questions is also available to search (CLC question papers).

All corpus resources are displayed on the main menu screen when you sign into Sketch, as shown below. The CLC resources are shown here in red:



In this guide, the CLC coded in section 4, the CLC uncoded is discussed in section 5 and the CLC question papers in section 6.

2. Which corpus should I use?

Corpus	Features	Use
CLC coded	- Error information	searching for things that learners find difficult
CLC uncoded	- twice as much data as the CLC coded - scripts from a wider range of nationalities and first languages	searching for things that learners can do, e.g. "is this collocation used at B1 level"
CLC Question papers	- Text from ESOL question papers	Search to find the language used in questions, e.g by level, exam

3. Functions that are common to both the coded and uncoded Learner Corpora

3.1 Narrowing your search by nationality, L1, exam, level

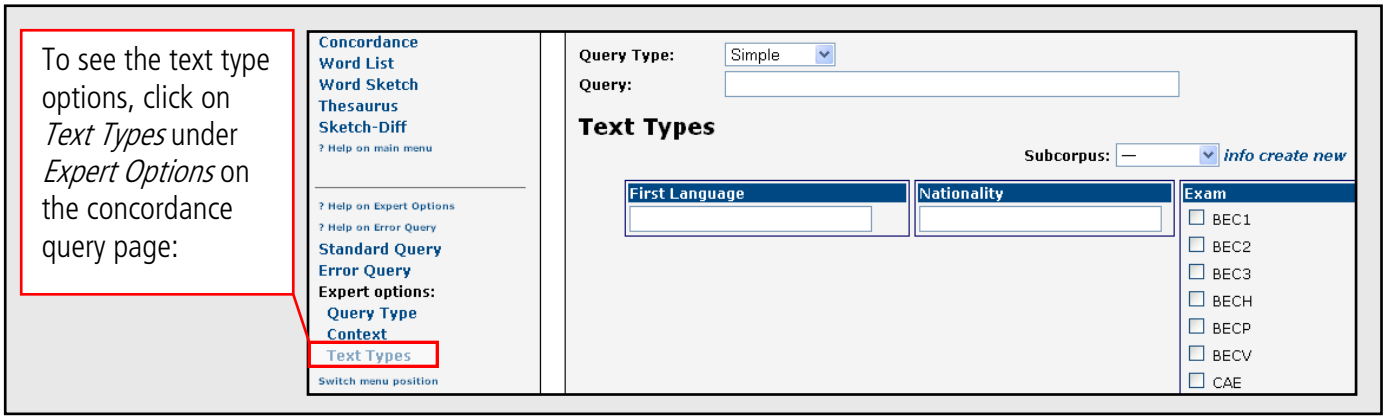
The *text types* function in the learner corpora work on the same principles as in the native speaker corpora, but includes different options.

Text types allows you to find results which match your target market.

For example, you can choose to look at only French speaking students from Switzerland, or only at results from the KET exam.

The extensive information we have about each candidate and each script means that you can look for any combination of a wide range of options.

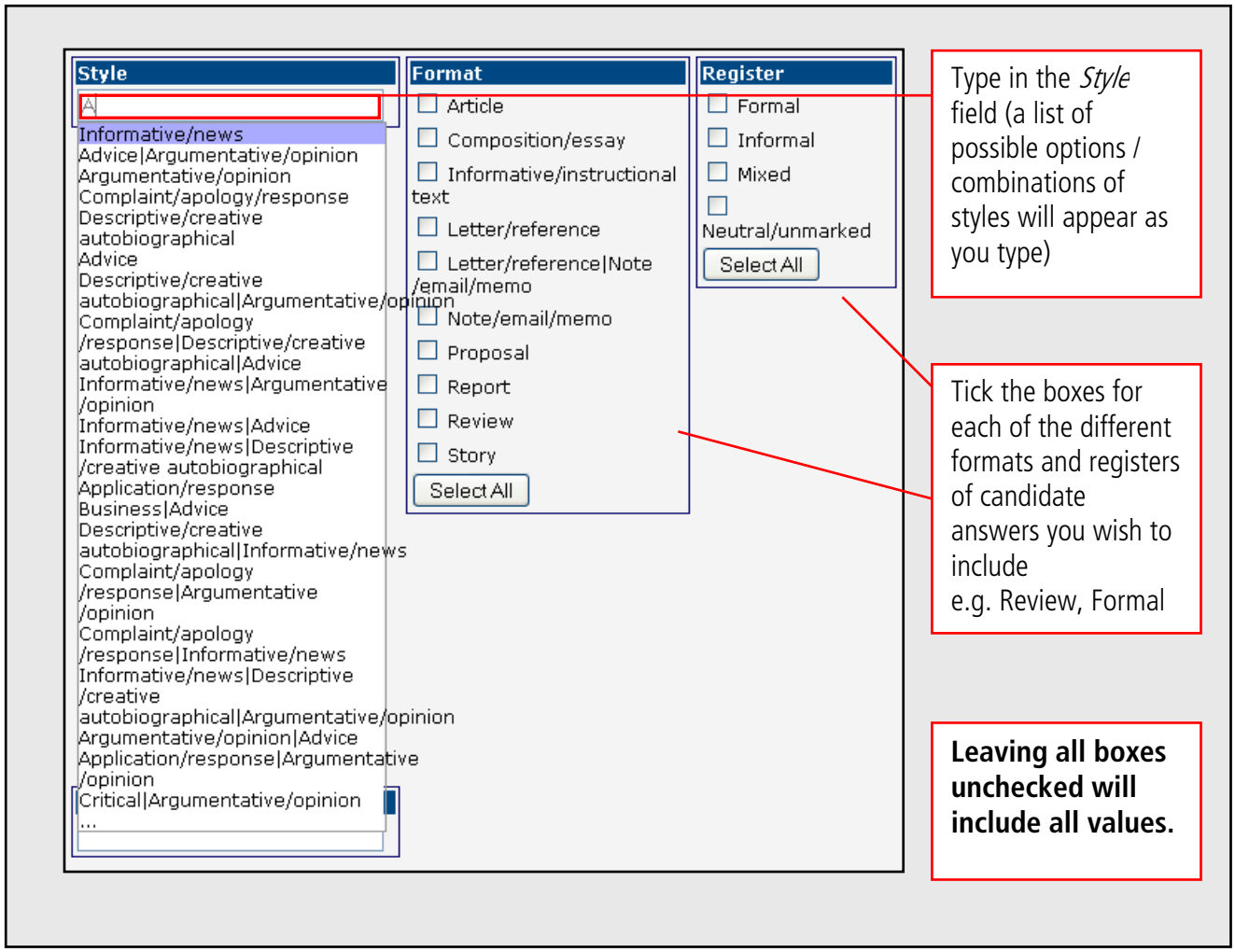
From the concordance screen, click on 'text types' to show the available options:



- To see results from a particular L1, type the name of the language into the *First Language* field
- To find results from a particular country, type the **name of the country** into the *Nationality* field (e.g. *Italy*, not *Italian*)
- To search for more than one option in either of these fields (e.g French **and** German speakers, or students from China **and** Taiwan), separate these values with vertical bar e.g. China|Taiwan.
- Select exams/level/age etc by checking the boxes in the various columns.
- Leaving all boxes unchecked will include all fields in that column by default.

3.2 Style, format and register

Three of the options available on the text types allow you to search for answers of a particular style, format and formality. You can use these to look for specific types of tasks that candidates have been set, such as informal letters, business reports, creative stories and informative articles. Use these in combinations to narrow your search: e.g. Business + letter / reference + formal



3.3 Creating a subcorpus

If you are frequently working with the same part of the Learner Corpus, you may wish to create a subcorpus of these scripts so that you don't have to select all your required options from the text types each time.

Creating a subcorpus also allows you to use other functions, such as **Word List** to find keywords from your subcorpus compared to the whole of the CLC or to another group of students. (For more about these functions, see section 6.) A subcorpus will also give you frequency information only for your results, and not for the whole corpus.

To create a subcorpus:

1. Go to the concordance query page and make sure the text type options are visible (click on *Text Types* under *Expert Options* on the left hand menu if they are not displayed)

2. Next to the text types heading there is a drop down box labelled subcorpus. This is where you will be able to select your subcorpus from when you have created it. Next to this box is the link *create new*. Click on the link.

3. Sketch Engine will now bring up a list of all the text types along with their word or token counts. This page may take a moment to load. You can then select the parameters you want to include.

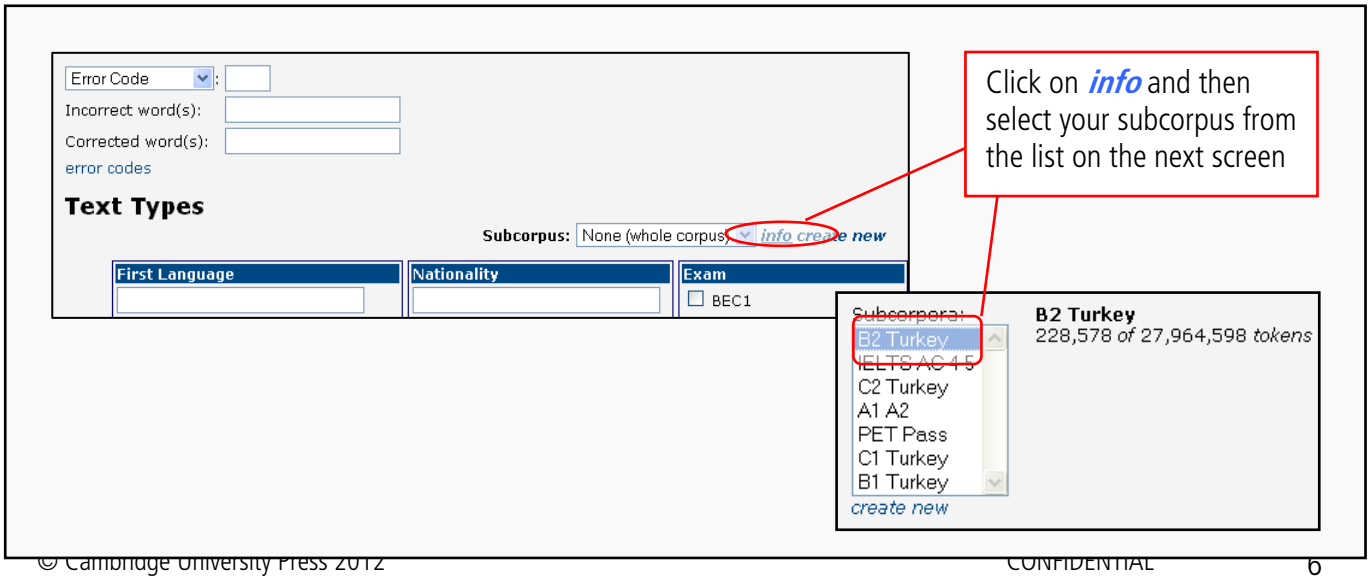
4. Once you are happy with your selection, click *Create Subcorpus*

You can now access this subcorpus every time you enter the CLC by choosing this subcorpus from the drop down box on the text types for concordance queries, **Word Lists** and **Word Sketch** (Advanced options).

Tips:

- To see **the breakdown of the CLC coded or uncoded** for each text type in words, tokens or documents, go to the subcorpus creation page where a full breakdown is given

- To find out how many words your subcorpus contains, select the *Text Types* option on the concordance query page, and click on *info* next to the subcorpus box. Choose your subcorpus from the list that appears

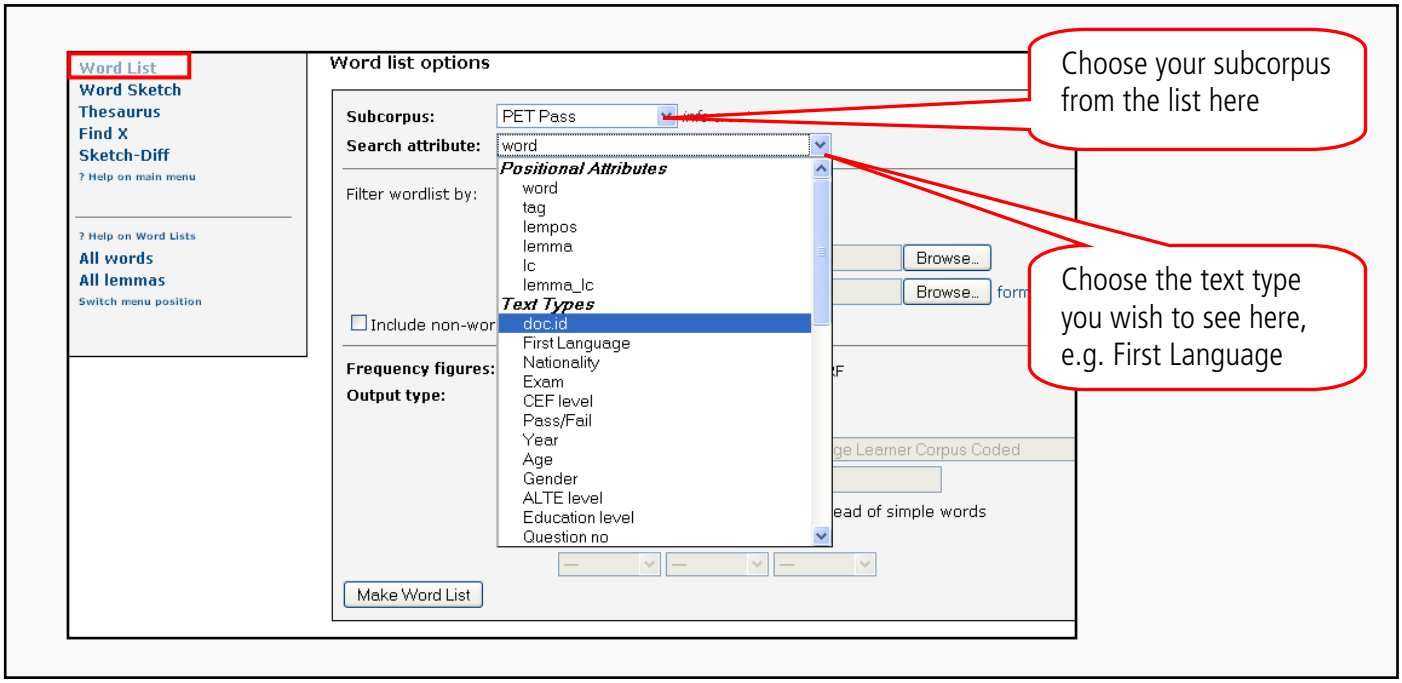


- To see the **breakdown of your subcorpus** by a particular text type:

  1) Go to **Word List** on the left hand menu, then go to the *Text Types* box:

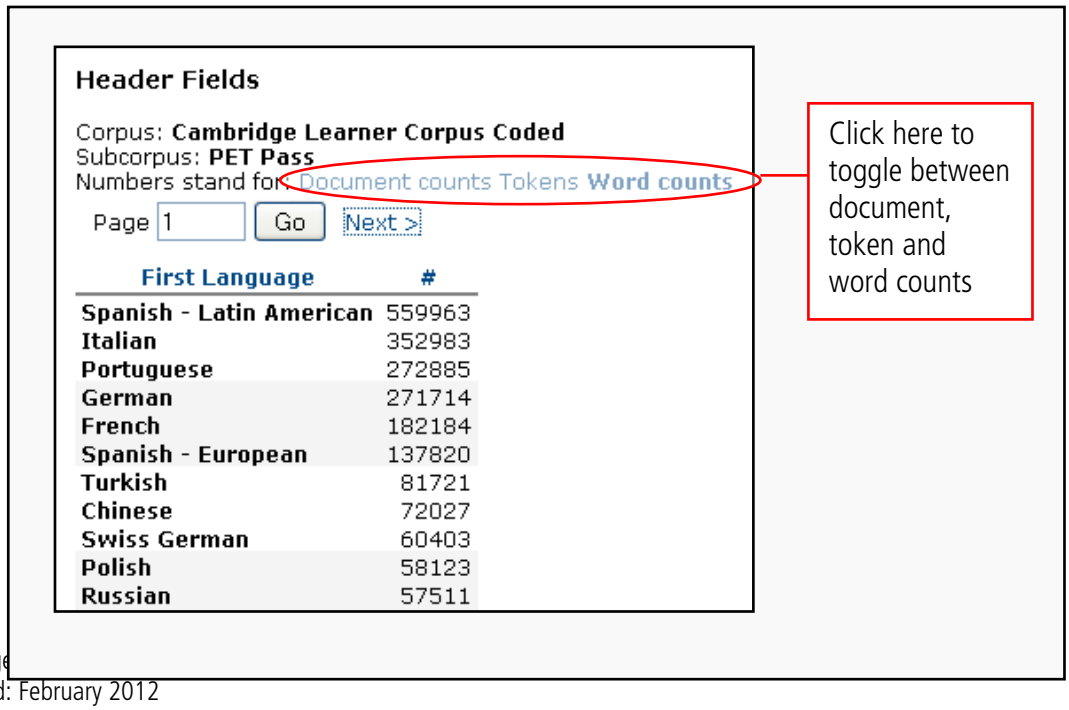
  2) Select your subcorpus from the drop down list

  3) Select the text type you wish to see the word count for from the drop down list under *Search attribute*



  4) Click *Make Word List*

This will then show you the breakdown of this text type in your subcorpus. You can change the count from words to tokens or documents:

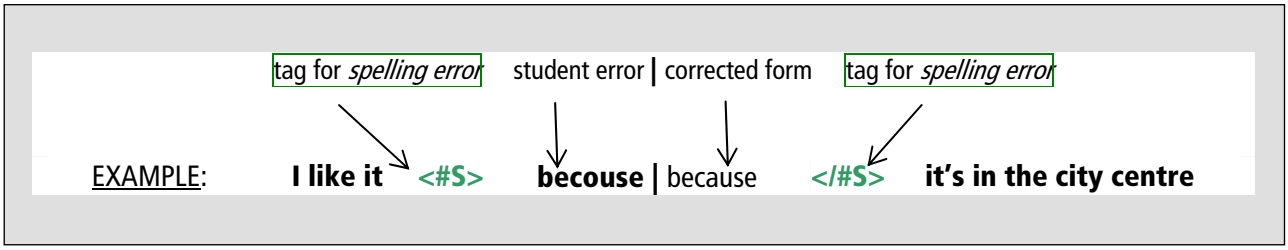


4. Using the Cambridge Learner Corpus (coded)

4.1 Error coding

The CLC coded contains exam scripts from ESOL exams that have been error coded to indicate learner errors.

These errors have a particular format. Error codes are shown at the start and end of each error type. The incorrect word or phrase (written by the student) is separated from the correct one (added by our error coders) with a vertical bar: | . An example of this can be seen below. In this case, a student has made a mistake with spelling the word *because*:



Where a word is missed out by the student, a blank space is shown before the vertical line that separates the student error and the corrected form

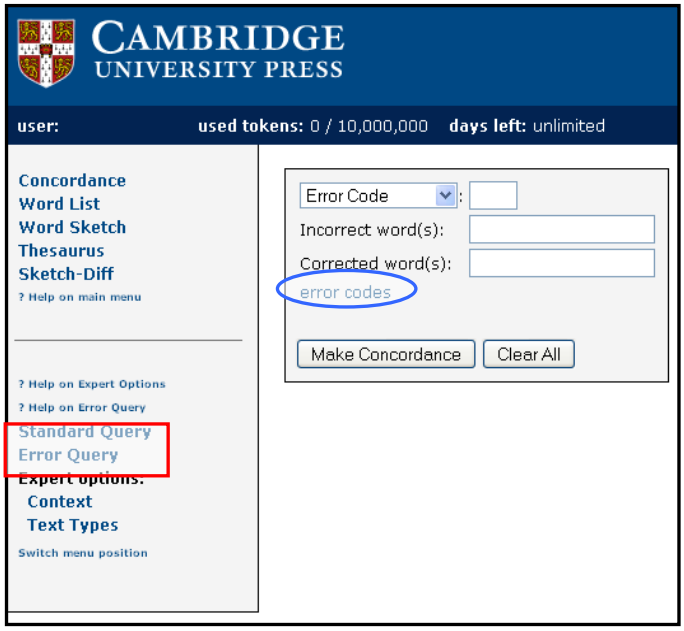
In Sketch Engine, the error display looks like this:

That's <#DY> realy | really </#DY> important <#S>	becaus | because	</#S> I just have one <#S> swimsuite | swimsuit
soon Love. Emily. I went to San Diego <#S>	becouse | because	</#S> I read a leaflet that said that it
rained. So I went to the beach every day <#S>	becouse | because	</#S> my hotel was near there and I enjoyed
was near there and I enjoyed that time <#S>	becouse | Because	</#S> I love the sun. I also visited the
</#RV> that they save animals in danger, <#S>	becouse | because	</#S> they <#RV> make | do </#RV><#UD> a |

A full list of the error codes used in the CLC can be found within the Cambridge Help available on Sketch Engine (http://www.cambridge.org/sketch/error_codes.html). More information on how these error codes are assigned can also be found in this location: http://www.cambridge.org/sketch/error_system.html

4.2 Working with error codes

The default search option for the CLC coded is to run an *error query* (as opposed to a *standard query*, which is the default option for all other corpora). The error query screen looks like this (the options to switch between standard and error queries are shown in red):



As shown in the screenshot, there are three boxes into which you can type. These are explained in more detail under *Making an Error Query*.

To use an error code or correction code, choose *Error Code* (or *Correction Code*) from the drop down menu and then enter your error code. You can find a list of error / correction codes by clicking on the *error codes* link (highlighted in blue on the screenshot).

For example, here is an excerpt from the Error code list:

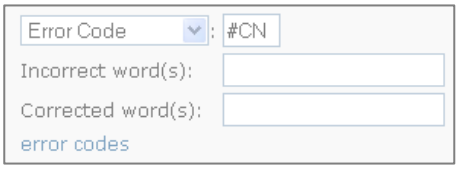
#AS	argument structure error
#CD	wrong determiner because of noun countability
#CE	complex error
#CL	collocation or tautology error
#CN	countability of noun error
#CQ	wrong quantifier because of noun countability

N.B. The # sign is a necessary part of the code. Type all codes in capital letters.

Making an Error Query:

The *Error Query* form has three boxes which allow you to search for errors by type, correction and for particular words. Their use, and possible combinations, are outlined here:

The **Error Code** box: use this to search for all errors of a particular type, e.g. all noun countability errors

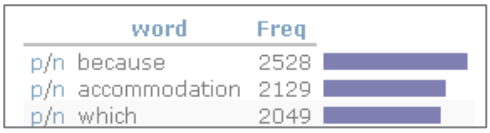


You can use this, for example, to find the top 10 mistakes students make with countable nouns: *informations, advices* etc

The **Correction Code** box: use this to search for all corrections of a particular error type, e.g. all words that are commonly misspelled.

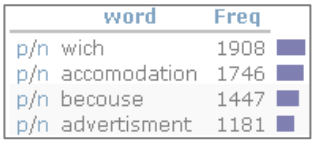
You can use this to find the top 10 words misspelled by students by using the Node form function, e.g. "because" (whereas a similar search using the Error Code option will show the most frequent actual misspellings, listing all variants separately, e.g. "becaus", "becouse")

Correction code search node forms:

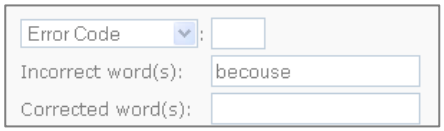


vs

Error code search forms:



Incorrect word(s): use this to search for a word that has been wrongly used, to see the contexts in which students make mistakes



For instance,
- commonly misused words like "funny"
- or for examples of a particular mistake, like "becouse"

Use the vertical bar | to separate words when searching for multiple words at once

Corrected word(s): use this to search for the mistakes students make when they should have written a particular word

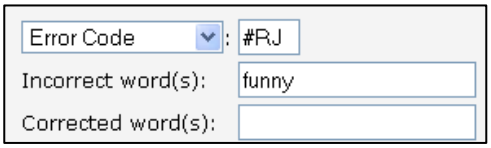
E.g. search for "information" to see all spelling and grammar mistakes made with "information"

You can also use a combination of these search boxes to find more specific examples:

Error code plus corrected word

Error code <#S> + corrected word "beautiful" : finds all the incorrect ways students spell beautiful

Error code plus incorrect word



Error code <#RJ> + incorrect word "funny": finds all instances where "funny" should be replaced with another adjective

It is worth exploring the various possible combinations to familiarise yourself with this function, as entering the right type of query can save you time when it comes to manipulating your concordance to find the results you need.

Manipulating your concordance

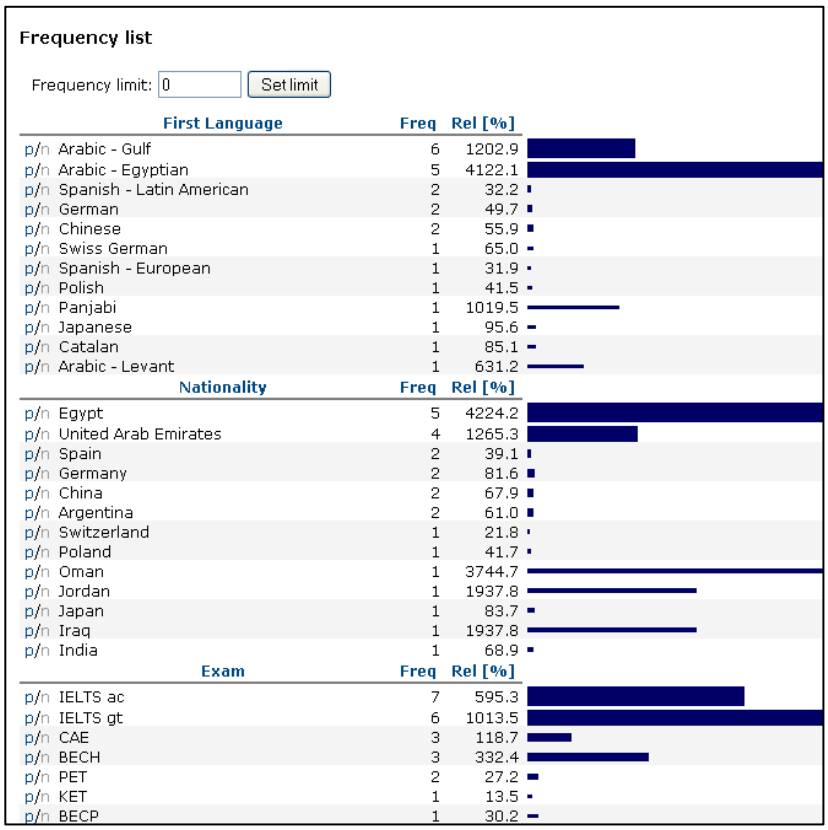
Once you've created your concordance, you can use the standard functions to manipulate your results, but with a few differences:

When you use *multilevel frequency*, remember that the corrected word will be counted as a token in the line. If you have searched for an error code, or an incorrect word, the node will be the incorrect word. The word one token to the right will be the correction.

- For some error codes, such as missing preposition, there will be a blank for the node (the preposition has been omitted).
- For error codes such as unnecessary preposition, the corrected word will be a blank.

If you have searched for a corrected word, this will be the node along with the incorrect word.

- *Text types* will show you the breakdown of your results by first language, nationality, age, exam level, etc. You can therefore see which groups of students most commonly make a certain mistake.



Frequency list

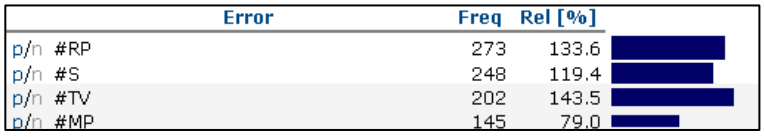
Frequency limit: 0 Set limit

First Language	Freq	Rel [%]
p/n Arabic - Gulf	6	1202.9
p/n Arabic - Egyptian	5	4122.1
p/n Spanish - Latin American	2	32.2
p/n German	2	49.7
p/n Chinese	2	55.9
p/n Swiss German	1	65.0
p/n Spanish - European	1	31.9
p/n Polish	1	41.5
p/n Panjabi	1	1019.5
p/n Japanese	1	95.6
p/n Catalan	1	85.1
p/n Arabic - Levant	1	631.2

Nationality	Freq	Rel [%]
p/n Egypt	5	4224.2
p/n United Arab Emirates	4	1265.3
p/n Spain	2	39.1
p/n Germany	2	81.6
p/n China	2	67.9
p/n Argentina	2	61.0
p/n Switzerland	1	21.8
p/n Poland	1	41.7
p/n Oman	1	3744.7
p/n Jordan	1	1937.8
p/n Japan	1	83.7
p/n Iraq	1	1937.8
p/n India	1	68.9

Exam	Freq	Rel [%]
p/n IELTS ac	7	595.3
p/n IELTS gt	6	1013.5
p/n CAE	3	118.7
p/n BECH	3	332.4
p/n PET	2	27.2
p/n KET	1	13.5
p/n BECP	1	30.2

- You can search for **all error types** by entering **#.*** into the *Error Code* box. When you have made your concordance, go to *Text Types* to see the types of error made in order of frequency:

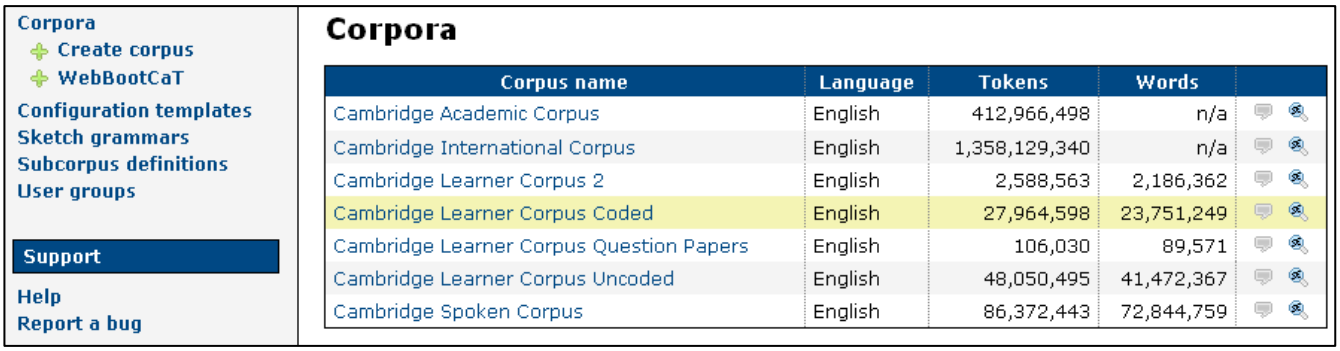


Error	Freq	Rel [%]
p/n #RP	273	133.6
p/n #S	248	119.4
p/n #TV	202	143.5
p/n #MP	145	79.0

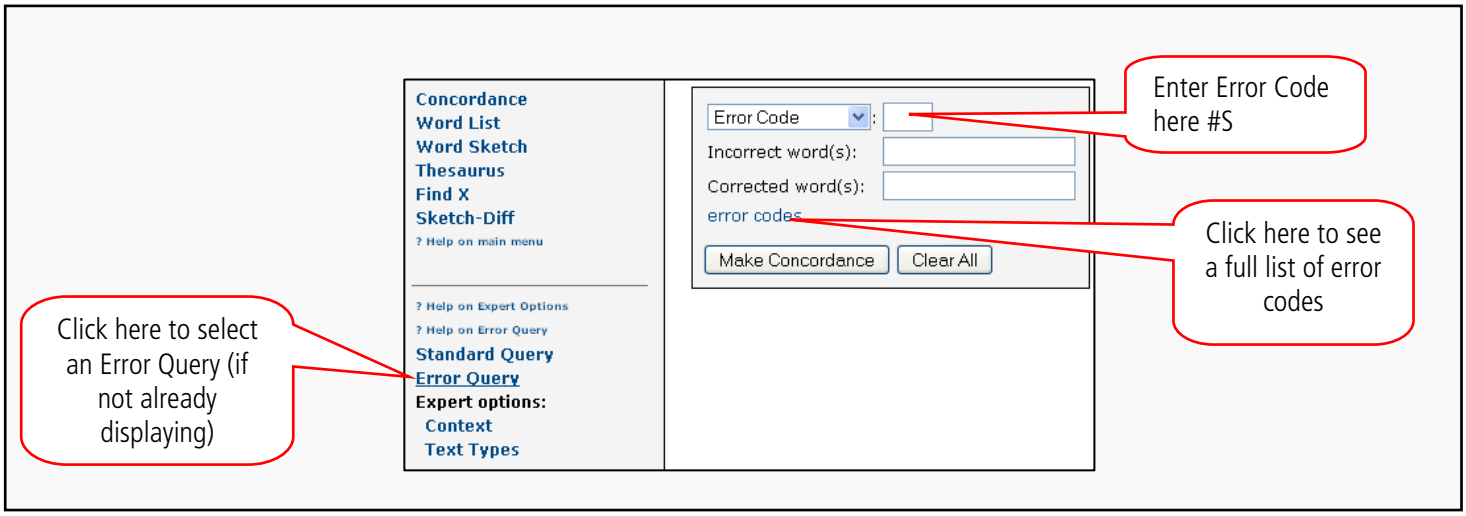
4.3 A worked example

How can we find the spelling errors made by Italian-speaking students taking the KET exam?

1) To access the CLC coded, go back to the main screen (by clicking on the logo) and then click on Cambridge Learner Corpus Coded (you don't need to do this step if you're already working with the coded Learner Corpus.)



2) Click on *Error Query* from the options in the menu on the bottom left of the screen. This then displays the error query screen in the main pane.

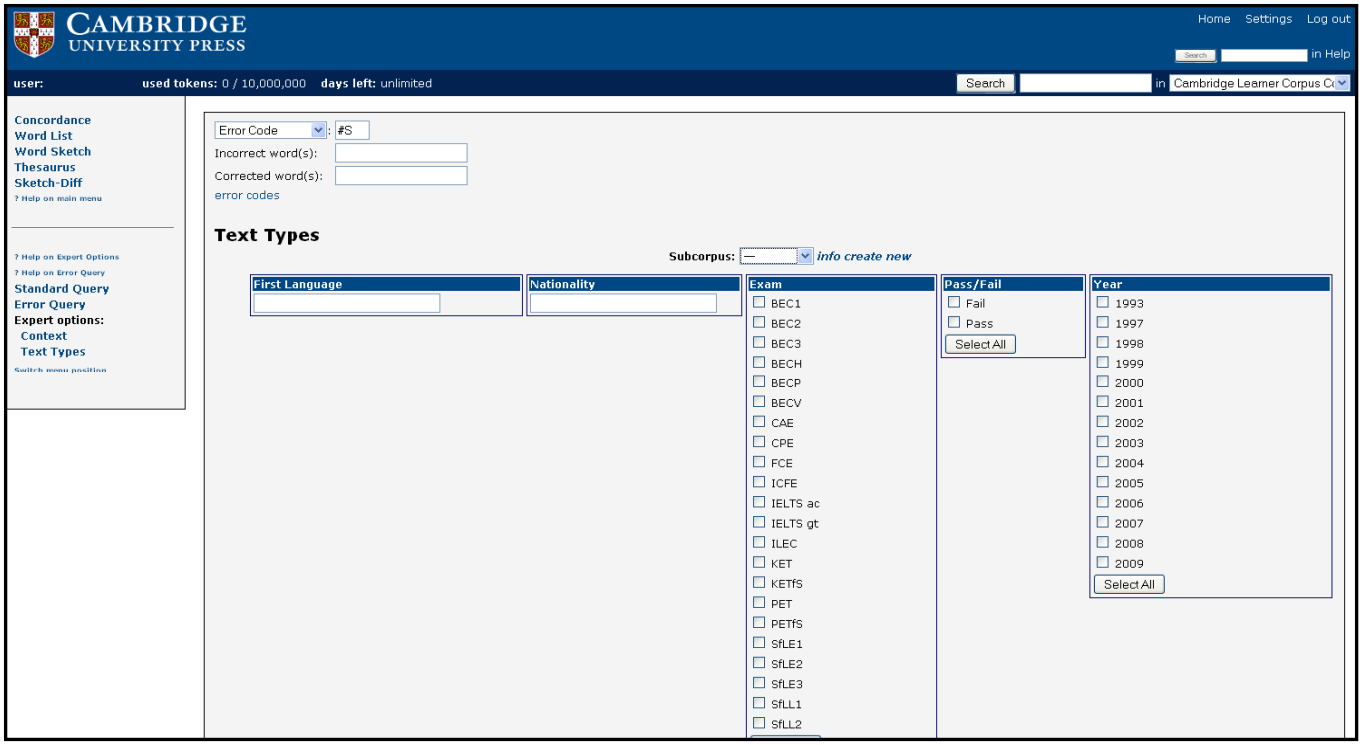


3) Click on ***error codes*** to open the list of error codes. Search for the code for spelling errors. (The code is #S)

4) Type this code into the box next to *Error Code*, in the main pane.

5) Now we need to narrow our search to those spelling errors made by Italian KET students only. To do this, click on *Text type* from the *Expert Options* menu on the bottom left of the screen.

6) The possible text types are then displayed in the main pane below the *Error Query* options. These are shown in the screen shot opposite:



We can use this screen to narrow down our search:

7) In the *First Language* box, type Italian

8) In the exam column, check the box for KET

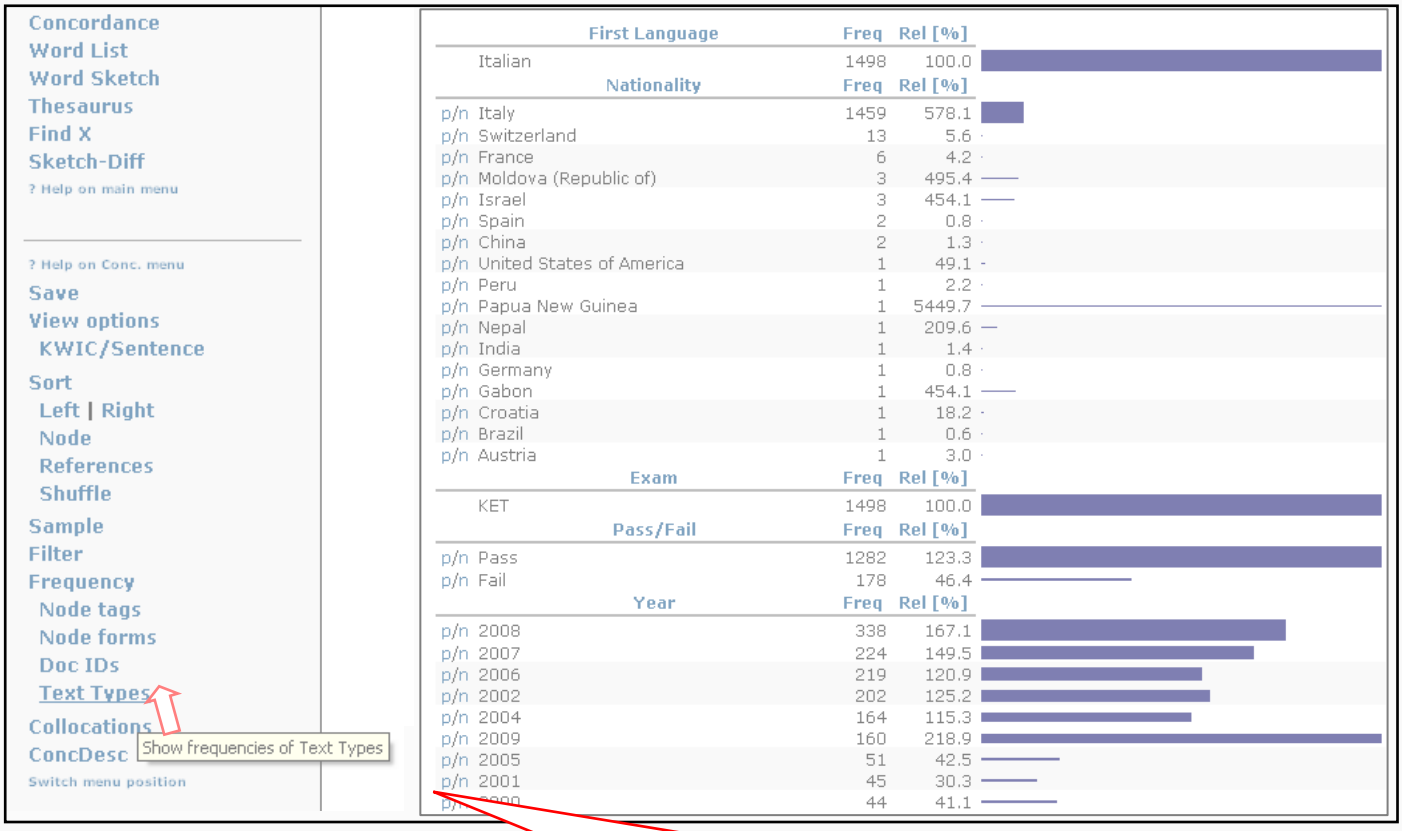
9) Scroll down to the bottom of the screen and click *Make concordance*

work. yours sincerely, <#S>	**Hy**	| Hi </#S> July! I think I left my phone there
carry | bring </#RV> anything. It doesn't <#S>	**metter**	| matter </#S> . <#UT> For <#FV><#RV> arrive at
<#MP> | . </#MP> See you <#S>	**tomorow**	| tomorrow </#S> <#MP> | . </#MP>
<#RT> on | at </#RT><#R> 19 | 7 </#R><#S>	**a'clock**	| o'clock </#S> , and you can bring a <#RP> videogame
Remember to bring <#RD> the | a </#RD><#S>	**pensil**	| pencil </#S> and <#RD> the | a </#RD> rubber.
June and it finishes on 14 July. Don't <#S>	**forghet**	| forget </#S> your art book. If we look
</#R> . This bed will cost three <#S>	**tousand**	| thousand </#S> and fifteen pounds. Thank
See you soon My new house is <#S>	**wonderfool**	| wonderful </#S> ! It is on the beach in California
California. Here the weather is always <#S>	**beatiful**	| beautiful </#S> ! It is very big <#RP> , | ; </#RP>
has two <#AGN> floor | floors </#AGN> : <#S>	**downstair**	| downstairs </#S> there <#AGV> are | is </#AGV>

You can further refine your concordance results to find out whether students of a particular age, school level, or taking an exam paper in a certain year made these mistakes.

By clicking *Text Types* on the left menu next to the concordance, you can see the frequency of errors made by Italian KET students:

- By age, school level, pass/fail, etc
- You can click on a year or an exam question to see only errors made in that particular paper or part of the paper



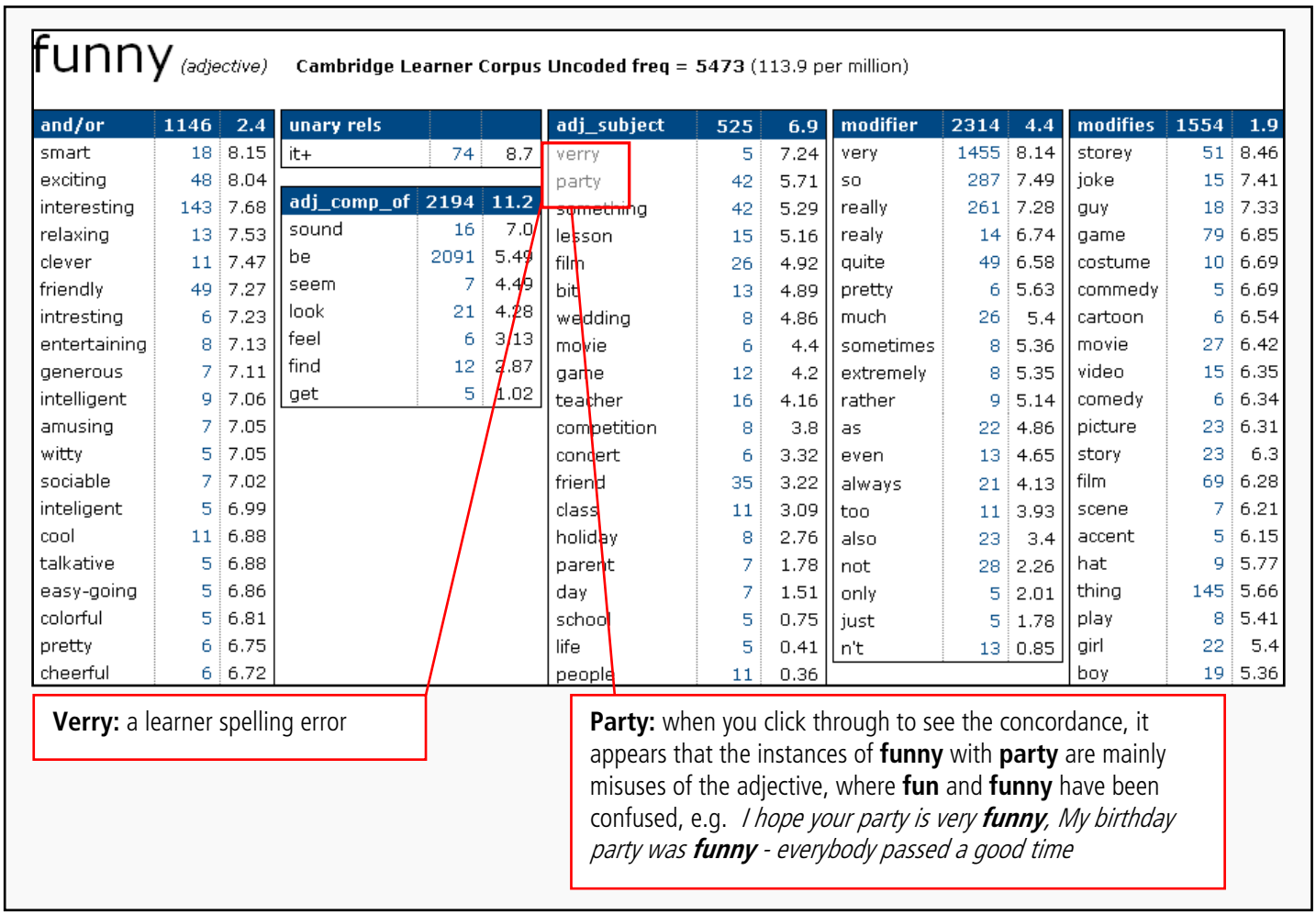
Clicking on *p* of p/n will take you to a concordance of Italian KET spelling errors from that particular text type only e.g. 2007 answers only

5. Using the Cambridge Learner Corpus (uncoded)

The CLC uncoded works in the same way as the native speaker corpora found in Sketch (namely, the Academic, International and Spoken corpora). This means that all functions outlined in the Getting Started and Advanced Help (e.g. searching, sorting, Word Sketches) also apply here.

While the data is of course different (and so has different metadata information, and therefore different options e.g. for making subcorpora of scripts by candidates of a particular first language or nationality), the types of queries available are the same as in the native speaker corpora. You can run a simple, lemma, phrase, word form or CQL search. For more information on these query types, see Getting Started.

One difference is that **Word Sketch** in the CLC may return unusual results. This is because student errors are in the text. For example, a Word Sketch for the adjective *funny* will return the following results:



funny *(adjective)* **Cambridge Learner Corpus Uncoded freq = 5473** (113.9 per million)

and/or	1146	2.4
smart	18	8.15
exciting	48	8.04
interesting	143	7.68
relaxing	13	7.53
clever	11	7.47
friendly	49	7.27
intresting	6	7.23
entertaining	8	7.13
generous	7	7.11
intelligent	9	7.06
amusing	7	7.05
witty	5	7.05
sociable	7	7.02
inteligent	5	6.99
cool	11	6.88
talkative	5	6.88
easy-going	5	6.86
colorful	5	6.81
pretty	6	6.75
cheerful	6	6.72

unary rels		
it+	74	8.7

adj_comp_of	2194	11.2
sound	16	7.0
be	2091	5.49
seem	7	4.49
look	21	4.28
feel	6	3.13
find	12	2.87
get	5	1.02

adj_subject	525	6.9
verry	5	7.24
party	42	5.71
something	42	5.29
lesson	15	5.16
film	26	4.92
bit	13	4.89
wedding	8	4.86
movie	6	4.4
game	12	4.2
teacher	16	4.16
competition	8	3.8
concert	6	3.32
friend	35	3.22
class	11	3.09
holiday	8	2.76
parent	7	1.78
day	7	1.51
school	5	0.75
life	5	0.41
people	11	0.36

modifier	2314	4.4
very	1455	8.14
so	287	7.49
really	261	7.28
realy	14	6.74
quite	49	6.58
pretty	6	5.63
much	26	5.4
sometimes	8	5.36
extremely	8	5.35
rather	9	5.14
as	22	4.86
even	13	4.65
always	21	4.13
too	11	3.93
also	23	3.4
not	28	2.26
only	5	2.01
just	5	1.78
n't	13	0.85

modifies	1554	1.9
storey	51	8.46
joke	15	7.41
guy	18	7.33
game	79	6.85
costume	10	6.69
commedy	5	6.69
cartoon	6	6.54
movie	27	6.42
video	15	6.35
comedy	6	6.34
picture	23	6.31
story	23	6.3
film	69	6.28
scene	7	6.21
accent	5	6.15
hat	9	5.77
thing	145	5.66
play	8	5.41
girl	22	5.4
boy	19	5.36

Verry: a learner spelling error

Party: when you click through to see the concordance, it appears that the instances of **funny** with **party** are mainly misuses of the adjective, where **fun** and **funny** have been confused, e.g. *I hope your party is very **funny**, My birthday party was **funny** - everybody passed a good time*

6. Using Cambridge Learner Corpus Question Papers

The Cambridge English Corpus also contains a corpus of the exam question papers that accompany the scripts found in the CLC coded and uncoded. These question papers are available to search as with any of the other corpora. There is also a link to the PDF versions of the exam question papers, for those who need to see the original. The PDFs are password protected – please contact a member of the Corpus team if you need access to them.

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 (ver2.1)	English	1,936,950,149	1,618,726,462
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 Question Papers	English	106,030	89,571
Cambridge Learner Corpus Uncoded	English	48,050,495	41,472,367
Cambridge Spoken Corpus	English	86,372,443	72,844,759

The text types in the CLC question paper corpus allow you to search for questions by exam, CEF level, year, question number and also the register, style and format of the tasks set.

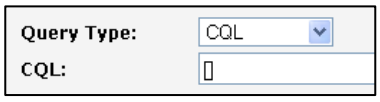
This means, for instance, that you can search for:

- A specific exam paper to accompany an answer you have found in the CLC coded / uncoded
- Any exam questions from a particular year
- All papers from a particular exam
- All part 2 questions from KET
- All business report writing tasks
- A certain word or phrase in all exam questions

To see all the papers in a certain group without using a specific search term:

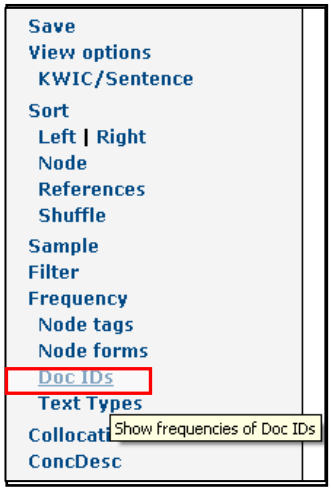
1) Select your text type options

2) Run a **CQL** search for [] (wildcard search)

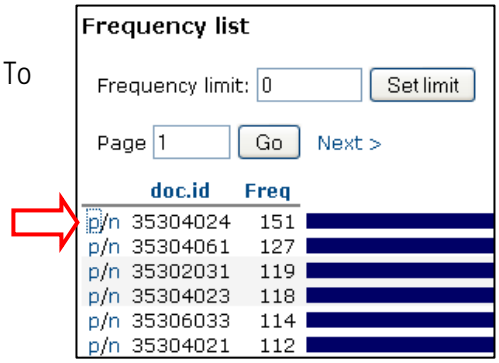


3) Make your concordance

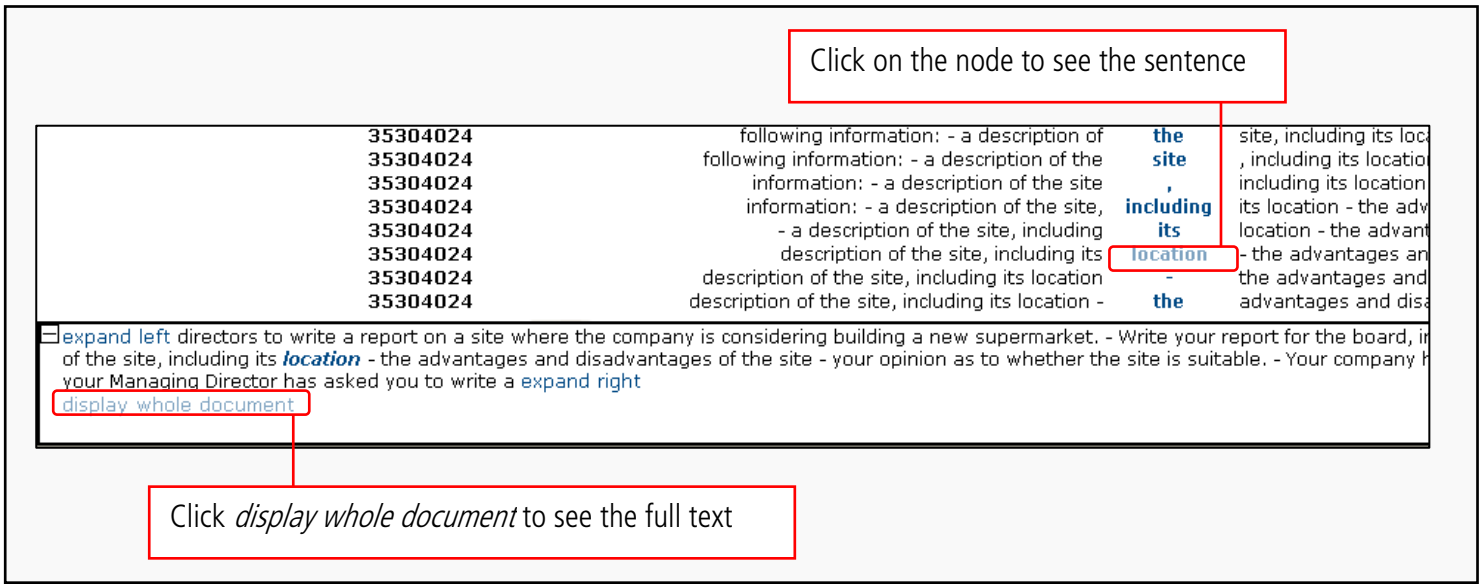
4) Go to *Doc IDs* under *Frequency* on the left hand menu



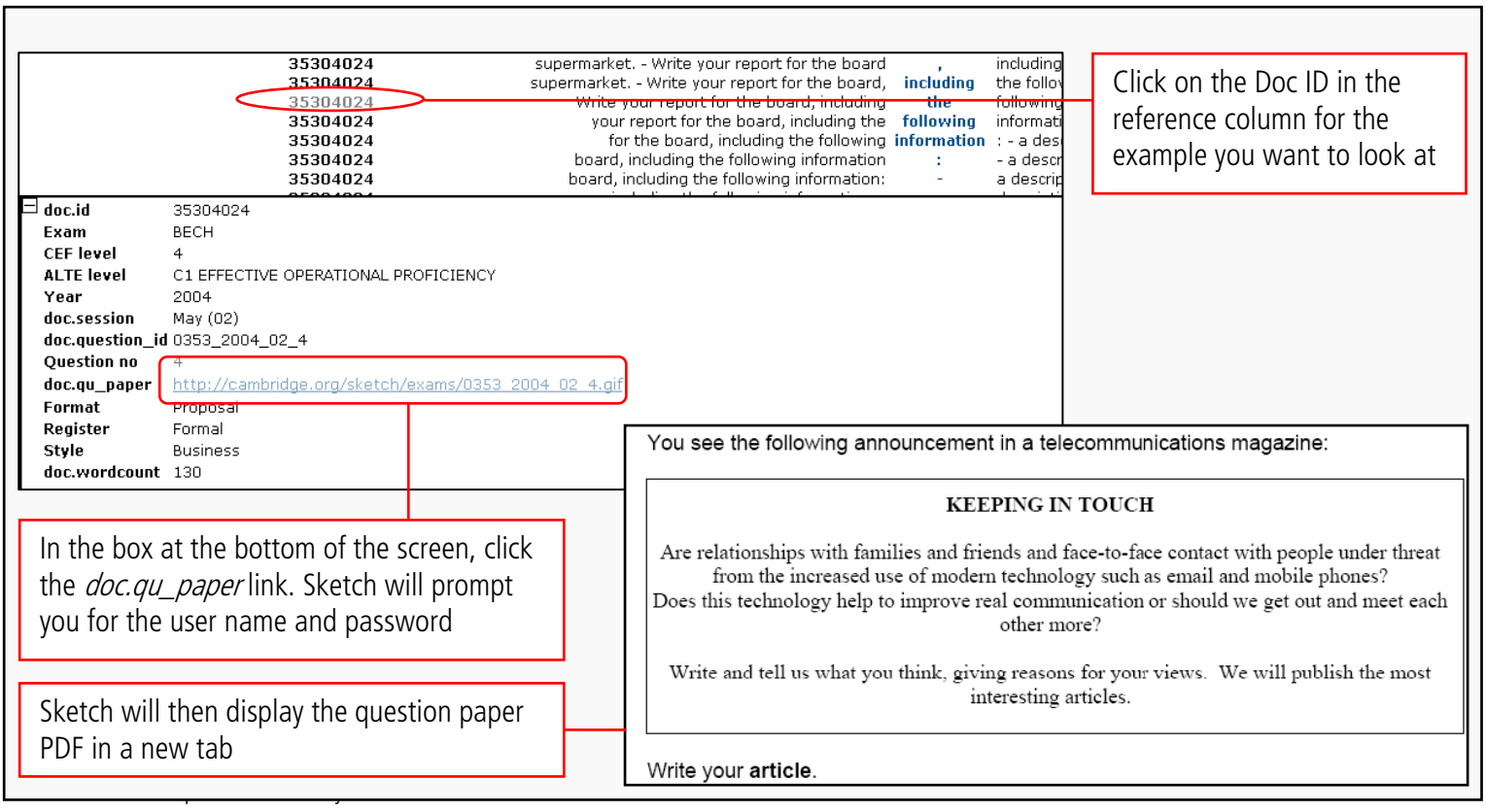
This will show you a list of all the different exam papers in that group. To see the actual question, click on *p* of p/n and then click on one of the nodes in the concordance line.



You can then click to expand the displayed text to see the whole document.



If you have been given the password, you can also click on the document ID in the reference column, and click the link in the document information box



7. More Advanced Functions

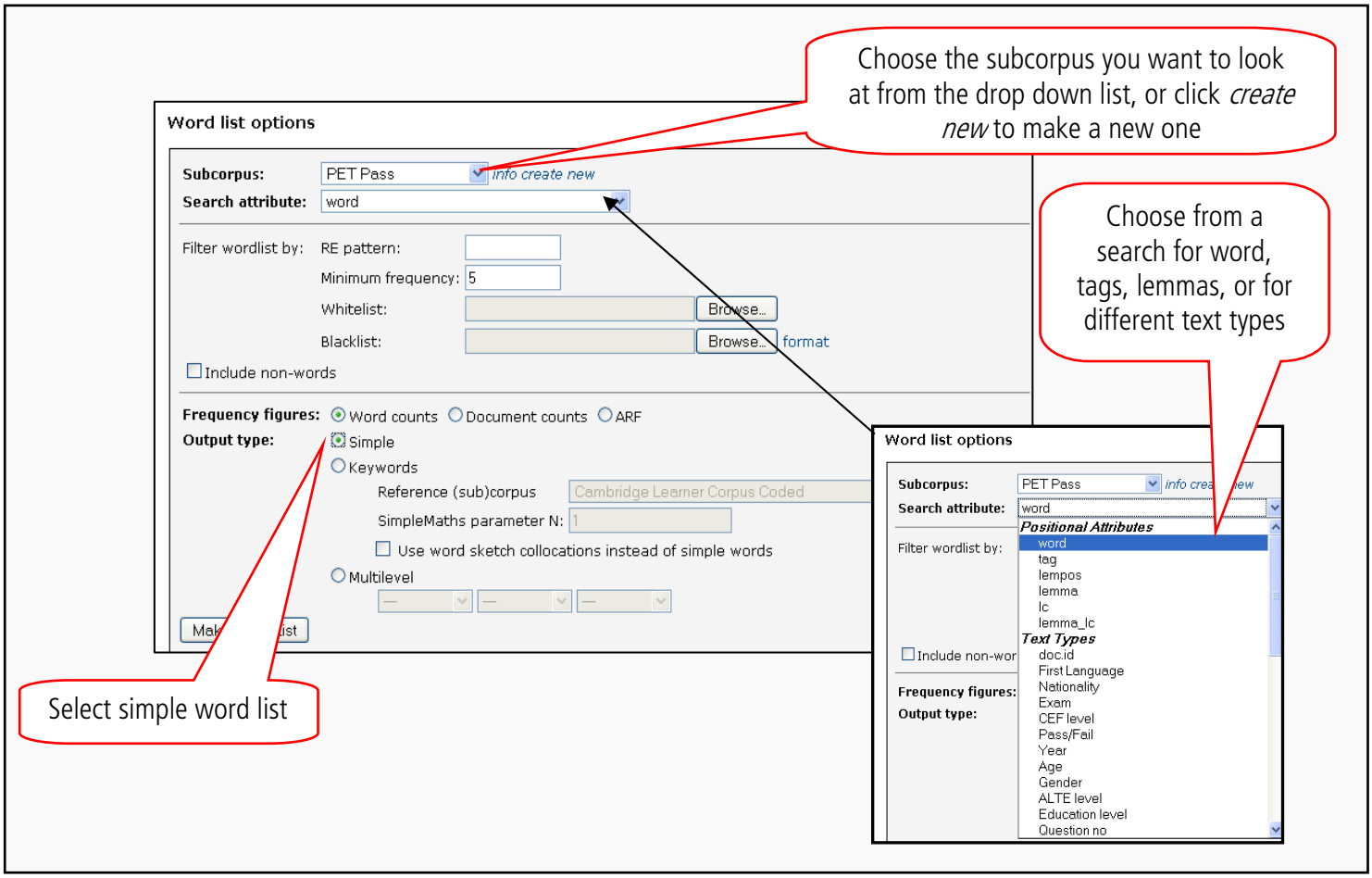
7.1 Creating word lists

You can find the most frequent words in your subcorpus by using **Word List** to make a simple word list:

1) Selecting the corpus your subcorpus is located in from the homepage (either CLC coded or uncoded), go to Word List on the top left hand menu

2) Choose your subcorpus from the drop-down list on the Word List entry form

3) You can search for a word, a tag, a lemma, or a text type



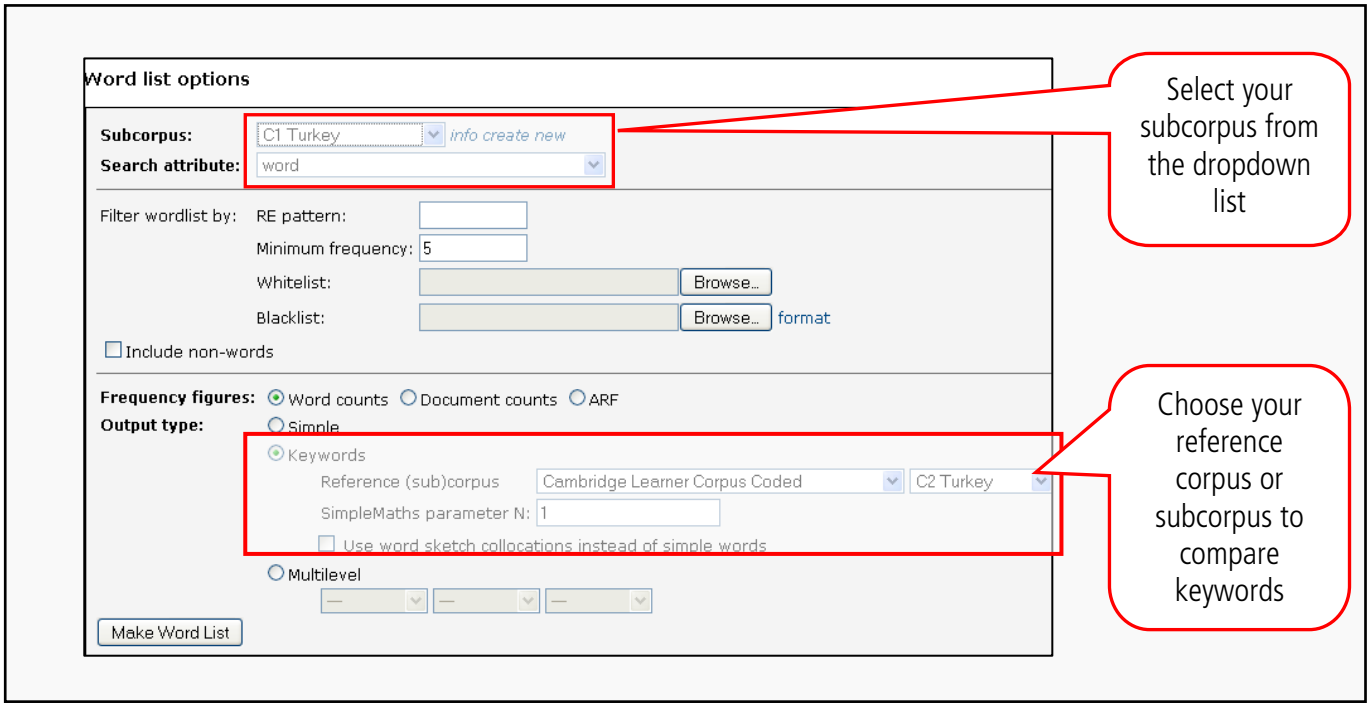
4) Click *Make Word List* to see a list of the most frequent words / tags / lemmas etc. The raw frequency is given alongside the frequency per million for each word

word	Freq
I	134421
to	93100
the	89635
you	70143
a	58155
and	57740
in	36392
my	36205
it	31812
is	29811
of	28626
was	26076
that	24386

Comparing across corpora and subcorpora

If you want to compare the key words in one subcorpus to another, or to the whole of the CLC coded or uncoded, you can do this using the *Keywords* function on **Word List**. For example, you can compare a subcorpus of Spanish speaking FCE students to the whole CLC uncoded, or a group of C2 students to C1 students.

- To do this, go to Word List on the left hand menu
- In the Keywords box, select the subcorpora / corpora you wish to compare e.g. C1 Turkish students and C2 Turkish students in the CLC coded



- Press *Make Word List* to see which words are strong keywords in your first corpus / subcorpus in comparison to your reference corpus.

The higher the score in the right hand column (here shown in red), the more frequently this word occurs in your subcorpus compared to your reference corpus.

	preloaded/cupclcc2:C1 Turkey			*preloaded/cupclcc2:C2 Turkey*			
word	**Freq**	**ARF**	**ARF/mill**	**Freq**	**ARF**	**ARF/mill**	**Score**
%	76	22.0	725.0	0	0.0	0.0	8.2
report	95	38.5	1267.1	2	1.6	67.9	8.1
Students	39	20.5	674.7	0	0.0	0.0	7.7
stalls	46	19.5	642.0	0	0.0	0.0	7.4
45	24	15.5	509.2	0	0.0	0.0	6.1
Bennington	27	15.4	507.5	0	0.0	0.0	6.1
Catering	35	15.4	507.4	0	0.0	0.0	6.1
Kavanagh	34	15.3	504.3	0	0.0	0.0	6.0
Student	29	14.9	489.8	0	0.0	0.0	5.9
Express	31	14.9	488.8	0	0.0	0.0	5.9
complaints	28	14.8	487.2	0	0.0	0.0	5.9
Services	28	14.7	483.5	0	0.0	0.0	5.8

7.2 Using error tags with CQL

This section explores using CQL in the CLC coded. For a full introduction to using CQL, please see the CQL Help guide.

For a CQL query within the CLC coded, choose *Standard Query* (not *Error Query*) from the left-hand menu, and select CQL from the query box options, as in the native speaker corpora. You can perform a CQL search as normal in the CLC coded if you just want to find instances of a particular word, phrase or grammatical structure. However, if you want to use CQL to search for these examples where they occur within a particular error type, you will need to use a different syntax in your query.

In Sketch Engine terms, errors are structures rather than tags, as they can cover a number of tokens. The syntax for specifying a structure in CQL is *within <str/>*, where str is the name of the structure. Errors are <err/ structures, corrections are <corr/>, and the code is given by the type attribute, so *<err type="#TV"/>* is a #TV error. So the CQL query you need to find continuous tense errors would be:

[lemma="be"][tag="VVG"] within <err type="#TV"/>

Or to get absolutely all verbs:

[lemma="be"][tag="V.*G"] within <err type="#TV"/>

Sorting by error tags:

You can sort by error tags – from the concordance, click on *View Options*, then set References to *Error* – this is the column on the left-hand side that displays the document ID by default. Click *Change View Options* button and you get the concordance window, but now the error codes are displayed in the left-hand column. Then click *References* under *Sort* in the left-hand menu and it will sort by the error code, as that is what you have set the references to display. By default, all the non-error lines will sort first, but you can jump to the error that you want by selecting it from the *Jump to* box.

To find frequencies of error tags or filter by error tag click *Text Types* under *Frequency* on the left-hand side, then right at the bottom of the page are frequency lists for error and correction codes. Clicking the *p* next to one of these will open a concordance containing just those errors.

Searching for more than one tag:

You can use regular expression syntax to specify more than one tag - #.*T will get all preposition errors, #[RM]T will get #RT and #MT.

Searching for more than one word (with an error tag):

You can search for more than one word with a particular error tag - for example #RA with *which*, *who* and *that* to look for relative pronoun errors. You need to use *within <err/>* to specify the error code. Bars within values mean 'or', so to search for multiple words you can just use *[word="which|who|that"]*, so for the example you give:

[word="which|who|that"] within <err type="#RA"/>

[word="was|were"] [tag="V.*G"] within (<s/> containing [word="\?"])

This means that they all have to be in the same sentence, and the 'was/were being' can occur anywhere within the sentence, rather than just towards then end. It also means that the 'was/were being' part is highlighted as the node rather than everything up to the question mark.