Using Sketch Engine with BAWE

Hilary Nesi

&

Paul Thompson

Version IV 2019
Revised by Andrew McIsaac

This manual will help you get started with Sketch Engine. Once you get proficient with this, you will be able to use the Sketch Engine manual available from the Sketch Engine website.

TABLE OF CONTENTS

Lesson 1      Making a simple concordance search
Lesson 2:     Examining collocations
Lesson 3      Attributes
Lesson 4      Corpus Query Types
Lesson 5      Examining Frequency
Lesson 6      Corpus Query Language
Lesson 7      Keywords
Lesson 1  Making a simple concordance search

1.1  Introduction to the interface

Click on the link to access the BAWE Corpus on Sketch Engine

Click ‘Concordance’ to start a concordance search. In the search box, write the word that you are interested in investigating. In this example, we have chosen the word ‘factor’.

![Concordance interface](image)
Click on ‘SEARCH’. You will get a page of results like this:

In the blue box in the top left corner, you can see how many instances of the word ‘factor’ (both singular and plural) occur in the whole corpus: 4548. If Sketch Engine finds that there are more than a few instances of a word, these will be displayed on a number of pages – in this case, there are 228 pages with the default setting of 20 lines on the page.

‘Factor’ and ‘factors’ appear on the screen in red because they are the search words. They are in the centre of the page. This kind of display is called a KWIC (Key Word In Context) concordance. It helps you to see what kinds of words surround the search term. For example, ‘factors’ are described as ‘key’, ‘social’ and ‘economic’. These words (‘key’, ‘social’ and ‘economic’) are collocates of ‘factors’.

We can obtain much more detail about collocations by using the ‘Collocations’ feature in Sketch Engine. We will look at how to do this in Lesson 2.

1.2 View options: length and number of concordance lines

At the moment you have 20 lines on the page. If you want to see more lines on the page, you can do this by scrolling to the bottom and changing the value of ‘Rows per page’.

If you select ‘Line detail view options’ next to any line number (the ‘i’ icon), you will see a page with details about the text the line came from. You might be interested in discipline and course title, but other pieces of information listed in this screenshot are not very informative. You may need to scroll down to find the information you need.
If you select ‘Discipline group’ (area), and then ‘save’, you will see letters standing for disciplinary groups in the left column. These are ‘AH’ for Arts and Humanities, ‘LS’ for Life Sciences, ‘PS’ for Physical Sciences and ‘SS’ for ‘Social Sciences’.

If you select ‘Discipline’ or scroll and select ‘Text genre’ you will see the discipline or genre of each concordance line.

You can select more than one of these by ticking multiple options. In this example, the disciplinary group and the genre were both selected.
The KWIC concordance line tells you which words come before and after your search word, but no more. You may want to see the search word in a larger context, and you may want to know more about the type of text it came from. If you click on the red search word in an individual concordance line, the wider context will be shown in a box at the bottom of the screen, as in this example:

```
that existed in the nineteenth century no longer continues the definition of 'Other' has changed (Miles, 1989), and it has left a legacy of racism in contemporary society. factors have also had an impact. The racism apparent at the time of colonialisation has been reinforced through the labour market. During this period the racial ideas were put into practice in the economy as those constructed as the 'Other' were exploited through the trade links established
```

You can increase the amount of context by clicking on either ellipsis above or below the text. (Note that the “</s><s>” is a tag to indicate the beginning and end of sentences. You can hide them if you wish by going to view options and deselecting the ‘s’ from the show structures tab, then pressing save.)

### 1.3 View options: Information about assignments

If you click on the ‘Line detail view options’ icon again, more details about the text will appear. Here is an example, with details of interest selected:

This tells us that the text was written by a female first year Sociology student aged 25 or older, whose first language is English, and who has received all her secondary
education in the UK. The assignment received a distinction grade and contained 1632 words.

Every assignment in the BAWE corpus has been coded for these categories of information, among others.

Here we have selected three categories of information (grade, level and gender). The resulting display looks like this:

In this example, all the lines were written by women (f) and all but two are level 2 (2nd year undergraduate) distinction (D) grade. The others are merit (M) and distinction grades at level 1.

1.4 Sorting the concordance lines

You can sort your concordance lines according to the alphabetical order of the words that appear to the left or to the right of the key word. You can also sort by attribute, such as discipline or genre.
Here is page 1 of the concordance lines, right sorted by 1 token. Notice the punctuation marks immediately after the search term – in Sketch Engine, punctuation is listed before letters of the alphabet.
Lesson 2: Examining collocations

2.1 The collocation tool

As we have seen in Lesson 1, ‘key’, ‘social’ and ‘economic’ are collocates of the word ‘factors’. We can see whether words go together frequently by looking at KWIC concordance lines, but in Sketch Engine we can also use the collocation tool to discover statistical information about how strong the collocation is (i.e. whether it is not simply random chance that the words occur together within a given range of words).

Working with the word ‘factor’, as before, click on ‘Search’. You will get a page of results like this:

Click on ‘Collocations’ (top right and highlighted in the screenshot above). The screen that appears will allow you to choose the range of words to consider, and the statistical measure of collocation that you want to use, through the ‘Advanced’ tab.

In the screenshot above, the range has been set at -5 to 5, which means that five words to the left of the key word and five words to the right are considered. If you are reporting your findings it is important to state what range you have chosen – -5 to 5 is a common choice.
2.2 Measures of collocation: T-score and Mutual Information

The screenshot also shows choices of statistical measure, including T-score, MI score and LogDice. Collocates from a T-score calculation tend to be more frequent words, while collocates from an MI calculation tend to be less frequent words (Hunston 2002: 72-75 provides a good clear discussion of this).

Click on “GO”. The resulting list, shown in the screenshot below, is ordered by cooccurrences and begins with some very common grammatical words. Notice that the T-score rankings mirror the frequency rankings.

<table>
<thead>
<tr>
<th>Word</th>
<th>ψ Cooccurrences</th>
<th>Candidates</th>
<th>T-score</th>
<th>MI</th>
<th>LogDice</th>
</tr>
</thead>
<tbody>
<tr>
<td>the</td>
<td>2,122</td>
<td>429,343</td>
<td>40.98</td>
<td>3.18</td>
<td>7.32</td>
</tr>
<tr>
<td>.</td>
<td>1,701</td>
<td>313,580</td>
<td>37.10</td>
<td>3.31</td>
<td>7.45</td>
</tr>
<tr>
<td>of</td>
<td>1,466</td>
<td>270,136</td>
<td>34.44</td>
<td>3.31</td>
<td>7.45</td>
</tr>
<tr>
<td>,</td>
<td>1,402</td>
<td>391,643</td>
<td>31.74</td>
<td>2.71</td>
<td>6.80</td>
</tr>
<tr>
<td>and</td>
<td>1,061</td>
<td>207,623</td>
<td>29.10</td>
<td>3.23</td>
<td>7.38</td>
</tr>
<tr>
<td>to</td>
<td>988</td>
<td>186,666</td>
<td>28.16</td>
<td>3.26</td>
<td>7.39</td>
</tr>
<tr>
<td>a</td>
<td>925</td>
<td>125,736</td>
<td>28.16</td>
<td>3.75</td>
<td>7.80</td>
</tr>
<tr>
<td>in</td>
<td>791</td>
<td>137,911</td>
<td>25.45</td>
<td>3.39</td>
<td>7.51</td>
</tr>
<tr>
<td>is</td>
<td>726</td>
<td>110,721</td>
<td>24.70</td>
<td>3.59</td>
<td>7.69</td>
</tr>
</tbody>
</table>

If you click on “MI” you will get a differently ordered list, as in the next screenshot. This shows some very rare words which almost exclusively occur with “factor” in the BAWE corpus.
2.3 Defining the range of collocation

If you are interested in the word that immediately precedes “factor” or “factors”, you can change the range to -1, as in the screenshot below. Changing the range to 1 would show the words that immediately follow the key word you have chosen.
This is what you get if you choose a range of -1, sorted by T-score:

<table>
<thead>
<tr>
<th>Word</th>
<th>Cooccurrences</th>
<th>Candidates</th>
<th>+ T-score</th>
<th>MI</th>
<th>LogDice</th>
</tr>
</thead>
<tbody>
<tr>
<td>other</td>
<td>209</td>
<td>11,278</td>
<td>14.03</td>
<td>5.69</td>
<td>8.76</td>
</tr>
<tr>
<td>important</td>
<td>200</td>
<td>5,439</td>
<td>13.93</td>
<td>8.67</td>
<td>9.36</td>
</tr>
<tr>
<td>these</td>
<td>136</td>
<td>9,993</td>
<td>11.15</td>
<td>4.63</td>
<td>8.36</td>
</tr>
<tr>
<td>external</td>
<td>89</td>
<td>905</td>
<td>9.38</td>
<td>7.40</td>
<td>9.05</td>
</tr>
<tr>
<td>risk</td>
<td>88</td>
<td>1,723</td>
<td>9.28</td>
<td>6.55</td>
<td>8.84</td>
</tr>
<tr>
<td>key</td>
<td>77</td>
<td>1,774</td>
<td>8.03</td>
<td>3.31</td>
<td>8.84</td>
</tr>
<tr>
<td>friction</td>
<td>60</td>
<td>290</td>
<td>7.73</td>
<td>8.57</td>
<td>8.67</td>
</tr>
<tr>
<td>environmental</td>
<td>60</td>
<td>1,205</td>
<td>7.90</td>
<td>9.51</td>
<td>8.42</td>
</tr>
<tr>
<td>transcription</td>
<td>56</td>
<td>177</td>
<td>7.47</td>
<td>9.18</td>
<td>8.90</td>
</tr>
<tr>
<td>main</td>
<td>54</td>
<td>2,876</td>
<td>7.13</td>
<td>5.10</td>
<td>7.90</td>
</tr>
</tbody>
</table>

Notice the ellipses next to the far-right column in the previous screenshot.

If you click on them you can choose between ‘Only this’, where you will get the concordance lines for ‘factors’ and ‘factor’ where the key word is preceded by the collocate you have chosen, as in the screenshot below, where ‘important’ is the chosen collocate. If you click on ‘Whole concordance but this’, you will only get the concordance lines for ‘factors’ and ‘factor’ where the collocate you have chosen does not precede the key word.
Lesson 3  Attributes

3.1  View options: Information about word class

In the ‘View options’ menu (the eye symbol in the top right), you can also choose to see the word class for the search word or all the words in the concordance output. To do this, you need to tick the ‘tags’ box, under ‘Attributes’, and choose ‘For each token’, or ‘For KWIC only’, as below.

When you do this for ‘factor’, the concordance output is like this:

The code ‘NN1’ is used for common nouns in the singular, and ‘NN2’ for common nouns in the plural. You can see the complete set of codes for word class at:
http://ucrel.lancs.ac.uk/claws7tags.html
3.2 **View options: Information about other attributes**

The remaining ‘Attribute’ options are:

- lemmas
- lemposes
- sem
- textpart

If you choose ‘lemmas’ from the menu, next to the search word you will see the form of the word that you would find in a dictionary entry (i.e., the lemma):

If you choose ‘lemposes’, you will see the same information with the addition of the word class ('lemma' + 'Part-Of-Speech'):
If you choose ‘sem’, you will see a semantic code appear after the search word. These codes group words according to their meaning in the manner of a thesaurus. A full list of the semantic codes (tags) is provided on the first query page and at:

http://ucrel.lancs.ac.uk/usas/USASSemanticTagset.pdf

In the following example, the word ‘film(s)’ is coded as Q4.3 which stands for the category ‘The Media: TV, Radio and Cinema’.

The ‘textpart’ code shows whether the word occurs in any of the following parts of the assignment:

- Abstract
- Bibliography
- Epigraph
- Figure
- Front
- Heading
- List
- Note
- Quote
- Running text
- Table
- Title

The majority of concordance lines will come from the ‘running text’ part which is the main body of the assignment.
Lesson 4  Corpus Query Types

So far, you have made queries by doing a Basic search. In this lesson, you will find out about other ways to make queries: ‘lemma’, ‘phrase’ and ‘word form’ through the Advanced search tab.

4.1  Basic query types

You can choose between various types of query on the Advanced tab of the Concordance page:

If you choose the ‘lemma’ option, and type in an uninflected form of a word, all the inflected forms of the word will appear in the concordance lines.

The ‘phrase’ option enables you to search for a sequence of words:
The ‘word’ option enables you to limit the search to a particular sequence of letters, for example if you want ‘take’ but not ‘takes’, ‘taking’, ‘taken’ or ‘took’. If you untick the box below, this will restrict the search to words which use upper and lower case in exactly the same way as in the search word. For example, ‘Lines’ will find ‘Lines’ but not ‘lines’, or ‘LINES’.

Below the ‘Word Form’ box, there is another called ‘CQL’. This stands for Corpus Query Language and you will learn about this in Lesson 6.

4.2 Using ‘Context’

If you want to see how two or three words co-occur within a short span of text, you can use the ‘Filter context’ option. If you type a word or a phrase in the box, you can then specify a lemma that must appear before or after this word or phrase, using the lemma filter. In the example to the left, we have chosen ‘position’, preceded by the lemma ‘take’, which must occur one, two or three words (tokens) before ‘position’. The results of this search look like this:
You can do the same kind of search specifying a lemma to occur to the right of the search word. A third possibility is to specify lemmas to the left and to the right. This might enable you to find phrases which have some variability.

4.3 Using ‘Text Types’

If you want to limit your query to a subsection of the corpus, you can use ‘Text types’.
For example, in the screenshot below, we have chosen to look for ‘factor’ within Physical Sciences (PS) with a Distinction grade (D) at level 4 study (taught masters).
You can also search for a specific discipline, genre and/or contributor first language. If you type the first few letters of the category, the available options will appear. For example, you can select multiple languages for the “First language of author” category, such as the three Chinese options, as shown below. However, most contributors used English as their first language, and so if you select a first language that was used by few contributors your search results will be limited.

You can see the full range of genre categories on the “Corpus Holdings page” at www.coventry.ac.uk/bawe.

In the open version of Sketch Engine, you have the option of using a subcorpus of, for example, texts produced only by speakers of English as a first language. This saves having to specify your choice of author first language for each search when making multiple searches.

In the subscription version you can create your own subcorpora from any of the text type options.
Lesson 5  Examining Frequency

5.1  Comparing frequency

Sketch Engine provides ways to find information about the relative frequency of lexical items. We can compare frequencies of words across disciplinary groupings, disciplines, genres or levels, for example. In the example below, we can see that the word ‘entropy’ only occurs in the Life Sciences and the Physical Sciences.

If we choose ‘Text Types’ from the ‘Frequency’ menu, we can see that ‘entropy’ is only used in four disciplines: Physics, Chemistry, Biological Sciences and Computer Science.
It is most frequent in Physics.

<table>
<thead>
<tr>
<th>Discipline</th>
<th>↓ Frequency</th>
<th>Relative [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1  Physics</td>
<td>5</td>
<td>1,434.5</td>
</tr>
<tr>
<td>2  Chemistry</td>
<td>4</td>
<td>1,702.8</td>
</tr>
<tr>
<td>3  Biological Sciences</td>
<td>3</td>
<td>419</td>
</tr>
<tr>
<td>4  Computer Science</td>
<td>1</td>
<td>265</td>
</tr>
</tbody>
</table>

The figures in the ‘Rel’ column indicate the relative frequency of the word. Relative frequency takes into account the number of texts in each category, so that if there are more texts in one category than in another this difference doesn’t distort the relative frequency ranking.

In the frequency below, we can see the relative frequency of the word ‘liable’ across disciplines and across genres. ‘Liable’ is overwhelmingly more frequent in Law and in Problem Questions.

<table>
<thead>
<tr>
<th>Discipline</th>
<th>↓ Frequency</th>
<th>Relative [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1  Law</td>
<td>94</td>
<td>1,012.2</td>
</tr>
<tr>
<td>2  Business</td>
<td>19</td>
<td>260.7</td>
</tr>
<tr>
<td>3  Engineering</td>
<td>10</td>
<td>78.4</td>
</tr>
<tr>
<td>4  Philosophy</td>
<td>6</td>
<td>106.1</td>
</tr>
<tr>
<td>5  History</td>
<td>4</td>
<td>62.8</td>
</tr>
<tr>
<td>6  Psychology</td>
<td>2</td>
<td>46.2</td>
</tr>
<tr>
<td>7  English</td>
<td>2</td>
<td>34.5</td>
</tr>
<tr>
<td>8  Computer Science</td>
<td>2</td>
<td>46.2</td>
</tr>
<tr>
<td>9  Comparative American Studies</td>
<td>2</td>
<td>42.8</td>
</tr>
<tr>
<td>10 Agriculture</td>
<td>2</td>
<td>32.1</td>
</tr>
</tbody>
</table>

We can get more frequency information by doing a concordance search for ‘liable’, and then getting collocation information in the range 0 to 1 (see Lesson 2, section 2.1).
### 5.2 N-Grams

You can search for n-grams or clusters of words from the menu on the left side of the page. The search in the screenshot below will create a list of 4-grams, clusters of four words which occur together. 4-grams are the most common length, as they tend to be long enough to be meaningful, but common enough that the frequency is not too low.

<table>
<thead>
<tr>
<th>Word</th>
<th>Cooccurrences</th>
<th>Candidates</th>
<th>T-score</th>
<th>MI</th>
<th>LogDice</th>
</tr>
</thead>
<tbody>
<tr>
<td>for</td>
<td>57</td>
<td>54,827</td>
<td>7.42</td>
<td>5.96</td>
<td>5.09</td>
</tr>
<tr>
<td>to</td>
<td>42</td>
<td>188,666</td>
<td>5.96</td>
<td>3.64</td>
<td>2.87</td>
</tr>
<tr>
<td>.</td>
<td>14</td>
<td>313,580</td>
<td>2.24</td>
<td>1.32</td>
<td>0.55</td>
</tr>
<tr>
<td>if</td>
<td>8</td>
<td>7,993</td>
<td>2.78</td>
<td>5.81</td>
<td>5.01</td>
</tr>
<tr>
<td>in</td>
<td>5</td>
<td>137,911</td>
<td>1.13</td>
<td>1.92</td>
<td>0.25</td>
</tr>
<tr>
<td>under</td>
<td>3</td>
<td>3,088</td>
<td>1.78</td>
<td>5.78</td>
<td>4.62</td>
</tr>
<tr>
<td>on</td>
<td>3</td>
<td>38,754</td>
<td>1.33</td>
<td>2.11</td>
<td>1.34</td>
</tr>
</tbody>
</table>

You can also refine your search further by going to the ‘Advanced’ tab. Here you can change settings such as the minimum frequency of each n-gram shown, select n-grams containing specific words, or find key n-grams in relation to a different reference corpus.

Below is an example search for n-grams of length 4 with a minimum frequency of 5, where we find key n-grams in the BAWE corpus with the Brown corpus as reference. Brown is chosen because it is a general-purpose corpus available with the open version of Sketch Engine. We do not include non-words (e.g. punctuation).
The results are shown in the next screenshot. These n-grams are those most frequent in the BAWE corpus relative to the frequency of each cluster of words in the Brown corpus (if you have an account you may use a larger corpus that is only available to subscribers, such as English Web).

You can find n-grams containing specific words by choosing the “containing word”. Here is an n-gram search for 3- and 4-grams containing the word “factor”.

There is also an option to find n-grams matching a regular expression. You will learn more about regular expressions and how to use them in lessons 5.4 and 6.
5.3 Word Sketch

The Word Sketch option will provide a full picture of the collocations. It is an easy way of seeing how a word behaves in terms of its collocations.

N.B. The search term must be a lexical word and not a functional, or grammatical, word. Click on the question mark for more about acceptable search terms.

The results of this query are shown below.

You can click on any of the headings to find usages of the word in context. The screenshot on the next page shows the concordances of ‘analysis’ when a modifier follows (click modifier to see a similar page).
You can find out more about how the Word Sketch works, and more you can do with it, by clicking on the “ABOUT” tab on the Word Sketch page.

Here is another example word sketch for the search term ‘apply’. This is a verb, and so its usage patterns and column headings are different to the noun ‘analysis’.
5.4 Word lists

We can create a word list by using the ‘Word list’ tool. Click on the ‘Word list’ link in the main menu. You can choose to create frequency lists by ‘words’, ‘tag’, ‘lemmas’, ‘lemposes’, ‘sem’ or ‘textpart’ (for explanations of these, see 3.2), as well as nouns, verbs, adjectives, adverbs, pronouns, conjunctions, and prepositions.

In the following screenshot, we have created a simple word list of the most frequent words in the corpus. You can see that these are all punctuation marks or function words; content words will come lower down the list.

You can use a form of wild card to identify words with a particular morphology. For this, we use regular expressions, which you will learn more about in Lesson 6. We are going to look for words which end with ‘-ation’ in this example. We use a full stop to indicate ‘any character’, followed by an asterisk which means ‘1 or more of
the previous’ which in this case means ‘one or more characters’ and then ‘ation’, which gives the following: .\*ation

This will result in the following list of words, ending in ‘-ation’.
Lesson 6    Corpus Query Language

In the last example in Lesson 5, you used a regular expression to find words ending ‘-ation’. Regular expressions are a part of Corpus Query Language, which is used for the following purposes:

- Specifying word class
- Looking for grammatical patterns, or lexicogrammatical patterns
- Restricting searches to specified sections or categories of text

CQL has its own syntax and you need to make sure that you get the form of a CQL query exactly right. If you make a mistake with this, you will get an error message like the following:

In a simple CQL query for a single word, the square brackets enclose the query, and the first part of the query identifies what category you are looking for (the technical term is ‘attribute’), followed by an equals sign, and then a code of some sort (the technical term is ‘value’) and this has to appear within a set of double quotation marks. For example, if you want to look for all singular proper nouns, you select the CQL query type, and then you write:

    [tag="NP1"]

In the example shown in the screenshot, the problem was that the second double quotation mark was missing.

In this example, tag is the attribute and NP1 is the value. The attribute tag is used when you want to specify a part of speech.
6.1 Using CQL to specify word class

In the example above, CQL was used to find all examples in the corpus of all items belonging to the word class, ‘singular proper noun’. The code for this is \textbf{NP1}. A full list of word class codes used in the BAWE corpus can be found at:

\url{http://ucrel.lancs.ac.uk/claws7tags.html}

The codes to identify nouns all begin with N, adjectives with JJ, and verbs with V. Verbs, for example, are further divided as follows:

- The verb ‘be’: VB
- The verb ‘do’: VD
- The verb ‘have’: VH
- VM: Modal verbs
- VV: Lexical verbs

A third letter is placed at the end of any verb code to show:

- 0: Base form of the verb
- D: past tense
- G: -ing’ ending
- I: bare infinitive
- N: past participle
- Z: ‘-s’ ending

When you look for a verb, you have to have three characters in the value, but you can substitute the second and/or third character for a full stop if you do not want to restrict your search so precisely. Here are some examples:

- V.G
- VB.
- V..

To find all the instances of a specified part of speech, use \texttt{[tag = "X"]}. For example \texttt{[tag="V.G"]} finds all the –ing participles in the corpus, and \texttt{[tag="NP.."]} finds all the proper nouns.

If you want to specify two or more alternatives for a given slot in the value, you can put the choices inside normal brackets and divide with a pipe character. For example \texttt{[tag = "VB(D|N)"]} captures all instances of the past participle and the past tense of the verb BE.
6.2 Using CQL to find grammatical patterns

We can combine a number of searches of the type we have described, using the lemma, tag and/or lempos attributes: Here are some examples:

A search such as [lemma = "impact" & tag = "V.."] will find only the verb forms of the specified lemma (in this case impact).

The query [lemma = "different"] [tag = "I|R."] finds the prepositions and adverbs following different.

The exclamation mark preceding the equals sign means does not equal. For example, the query [lemma="fast" & tag != "J.."] will find fast as a noun, verb and adverb, but not as an adjective:

The query [lemma="talk" & tag = "V.."] [word !="about"] finds the verb talk followed by anything but about.

The query [tag = "VB(D|M|R|Z)"] [tag = "VVN"] finds am, are, is, were or was followed by the past participle of a lexical verb, and so will identify passive constructions.

Empty brackets [] allow any one word to come between the two attributes. Adding numbers between curled brackets, e.g. {1,3} specifies the range. For example [tag = "VB(D|M|R|Z)"] [{0,4} [tag = "VVN"] finds am, are, is, were or was followed by the past participle of a lexical verb, with at most four words in between.
6.3 Using CQL to restrict searches to specified sections or categories of text

You can use ‘within’ followed by an equation within angle brackets `<XX="XX"/>` to look for items within specified files. For example, the query `[word="he"] within <text l1="English"/>` looks for *he* only within those files produced by writers whose first language is English.

```
CQL
[word = "he"] within <text l1 = "English"/>
```

You can also use ‘within’ to limit your search to items which occur in sections of text which have been annotated as quotations. For example `[lemma="think"] within <quote lang="w+"/>` looks for *think* within quotations:

```
CQL
[lemma = "think"] within <quote lang = "\w+"/>
```

A query with ‘textpart’ will search for items which only occur in a specified part of the text: the main body (‘running-text’), the bibliography or the abstract. For example, the query `[lemma="government" & textpart="running-text"]` finds all instances of *government* that only occur in running text.

```
CQL
[lemma = "government" & textpart = "running-text"]
```

The following queries will search for *government* in bibliographies and abstracts:

- `[lemma="government" & textpart="bibliography"]`
- `[lemma="government" & textpart="abstract"]`

The exclamation mark ! preceding the equals sign can be used to exclude specified files or text parts. For example, the query `[word="he"] within <text l1!="English"/>` looks for *he* only within those files produced by writers whose first language is not English.

```
CQL
[word = "he"] within <text l1 != "English"/>
```

Similarly `[lemma="government" & !textpart="bibliography"]` finds all uses of ‘government’ outside the bibliography.
Lesson 7  Keywords*

*Keywords are only available with the subscription version of Sketch Engine. Many institutions provide access with an institutional login, or you can register for a 30-day trial to test out the premium features.

Keyword extraction can be used to determine words used uniquely or relatively more frequently in the BAWE corpus compared to a (usually large and general) reference corpus.

7.1 Extracting Keywords

Click on Keywords in the main menu, as seen below, or find it near the bottom on the sidebar menu.

![Corpus Menu Screenshot]

We can choose the reference corpus here, along with other settings. The default reference corpus (English Web 2013) is the largest available general corpus and may therefore be the best one to suit your needs.
Another useful setting is the ‘Focus on’ setting, that allows you to change a slider to change the focus of the search, based on the relative rarities of words in the corpora.

Using the default settings, the resulting search should show a page like the one to the right. “PUNC” is the BAWE corpus lemma for all punctuation marks, which English Web 2013 does not use, so this first lemma can be ignored.

Here you can see all the words which occur relatively more often in BAWE compared to the English Web 2013 corpus.
On its own, this isn’t very enlightening. Keyword extraction can also be conducted using the whole BAWE corpus as the reference corpus, compared with a subcorpus of BAWE representing a discipline, genre or level of study.

For example, the below screenshots show a search for keywords in the whole BAWE corpus against just those in the Biological Sciences discipline.

The results are all nearly all technical terms relating to that discipline.