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 1  
Making a simple concordance search

1.1  
Introduction to the interface

Click on the line which says ‘British Academic Written English Corpus’

In the ‘Query’ box, write the word that you are interested in investigating. In this example, we have chosen the word ‘factor’.

Click on ‘Make concordance’. You will get a page of results like this:

In the blue box in the top right 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.
‘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, or if you want longer lines, you can do this by clicking on ‘View options’. You will see this page:

If you enter ‘100’ in the ‘Page size’ box, and ‘80’in the ‘KWIC Context size’ box, and then click on the ‘Change View Options’ button, you will see much more text on the page.

If you select ‘text discgroup’ you will see blue 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 ‘text discipline’ or ‘text genre’ you will see the discipline or genre of each concordance line.
You can select more than one of these by holding down the Control key. In this example, the disciplinary group and the genre were both selected.

If you tick the ‘Shorten long references’ option in View Options the words in the blue column will be abbreviated. Don’t forget to click on ‘Change View Options’ too.

In the example below the words in the blue column have been abbreviated.
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:

You can increase the amount of context by clicking on ‘expand left’ or ‘expand right’.

### 1.3 View options: Information about assignments

If you click on the blue ‘references’ in the left hand column, more details about the text will appear in a box at the bottom of the screen. Here is an example:
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. You can see more information about each concordance line by going to the ‘View options’ menu:

Here we have selected three categories of information (grade, level and gender). To select more than one, you need to keep the Control key pressed down as you make each selection. The resulting display looks like this:

```
D.1.f  only reason for the continuation of racism, economic factors have also had an impact. <p> The racism apparent
M.1.f  actual area (May, 1997). <p> Human judgement is a key factor in this criterion, but as the decision process shows
D.2.f  decision, introduced by Morell (1994), looking at factors identified as 'pushing' women into motherhood, psychological, social and economic factors, as well as looking at the influences of 'push factors'
D.3.f  factors, as well as looking at the influences of 'push factors' such as national discourages on motherhood and social
D.4.f  considering the influence of personal, social and economic factors, and changing opportunities for women. In considering
D.5.f  (Ottino, 1993). <p> However as well as personal factors attracting women to motherhood, economic and social
D.6.f  attracting women to motherhood, economic and social factors may also be influential. Lancaster has argued that
D.7.f  1978, p364). However it is now felt that economic factors do not influence the decision as much as in previous
D.8.f  needed to provide a source of income. Instead social factors have more of an input. <p> Coleman claims that
D.9.f  combine in a complex way, however they are not the only factors that affect the decision. National discourages on
D.10.f  discourages in society influence the more personal factors that are involved on the decision to become a mother
D.11.f  influence and combine with personal, economic and social factors to create a complex combination of reasons as to
D.12.f  2000). <p> There is also the influence of economic factors to consider. Children are no longer economic assets
D.13.f  Andaoka, 1978). <p> This links quite closely to social factors, as a desired lifestyle can be quite important when
D.14.f  well as there being personal, economic and social factors influencing the decision not to have children, changes
```
In this example, all the lines were written by women (f) and all but one are level 2 distinction (D) grade. The other is a merit (M) grade.

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. (The ‘Node’ option will order the lines according to the form of the key word.)

In the screenshot below, you can see that the concordance lines are left sorted, and this page is the 19th of 76.

Here is page 1 of the concordance lines, right sorted. Notice the punctuation marks immediately after the search term – in Sketch Engine, punctuation is listed before letters of the alphabet.

A summary of page features:
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 ‘Make concordance’. You will get a page of results like this:

Click on ‘Collocations’ (circled in the screenshot above). The next screen will allow you to choose the range of words to consider, and the statistical measure of collocation that you want to use.
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. 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 the default choice of statistical measure, the T-scores and MI score. We recommend that you follow the default setting. 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 “Make candidate list”. The resulting list, shown in the screenshot below, is ordered by frequency and begins with some very common grammatical words. Notice that the T-score rankings mirror the frequency rankings.

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 and 0, as in the screenshot below. Typing in 0 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 to 0.
Notice the p/n links next to the collocates. If you click on ‘n’ you will only get the concordance lines for ‘factors’ and ‘factor’ where the collocate you have chosen does not precede the key word. (‘n’ stands for ‘negative’.) If you click on ‘p’ 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. (‘p’ stands for ‘positive’.)

The government reaction to the 1905 uprising was an **important factor** in why the Tsar was not overthrown. As Christopher Read points out, the Tsar's reaction was a significant factor in shaping the population's resistance to change. In the early modern period, the growth of the population was a result of a complex interplay between various factors, including demographic, economic, and environmental changes. The government reaction was a key factor in determining the course of events, and it is crucial to consider how these factors interacted to shape the outcome.

An extremely **important factor** in the development of capitalism and industrialization was the availability of cheap labor. The growth of the industrial revolution in the 19th century was driven by the expansion of the labor force and the availability of raw materials. The government's reaction to the 1905 uprising was an important factor in shaping the population's resistance to change. In the early modern period, the growth of the population was a result of a complex interplay between various factors, including demographic, economic, and environmental changes. The government reaction was a key factor in determining the course of events, and it is crucial to consider how these factors interacted to shape the outcome.

**Table:**

<table>
<thead>
<tr>
<th>Collocate</th>
<th>Freq</th>
<th>T-score</th>
<th>ML</th>
</tr>
</thead>
<tbody>
<tr>
<td>Other</td>
<td>209</td>
<td>14.03</td>
<td></td>
</tr>
<tr>
<td>important</td>
<td>200</td>
<td>13.93</td>
<td>6.075</td>
</tr>
<tr>
<td>these</td>
<td>135</td>
<td>11.150</td>
<td>4.630</td>
</tr>
<tr>
<td>external</td>
<td>89</td>
<td>9.378</td>
<td>7.401</td>
</tr>
<tr>
<td>risk</td>
<td>88</td>
<td>9.281</td>
<td>6.549</td>
</tr>
<tr>
<td>key</td>
<td>77</td>
<td>8.665</td>
<td>6.314</td>
</tr>
<tr>
<td>friction</td>
<td>60</td>
<td>7.726</td>
<td>8.567</td>
</tr>
<tr>
<td>environmental</td>
<td>60</td>
<td>7.661</td>
<td>6.512</td>
</tr>
<tr>
<td>main</td>
<td>54</td>
<td>7.135</td>
<td>5.105</td>
</tr>
<tr>
<td>many</td>
<td>56</td>
<td>7.030</td>
<td>4.045</td>
</tr>
<tr>
<td>These</td>
<td>49</td>
<td>6.742</td>
<td>4.760</td>
</tr>
<tr>
<td>economic</td>
<td>49</td>
<td>6.736</td>
<td>4.728</td>
</tr>
<tr>
<td>social</td>
<td>49</td>
<td>6.506</td>
<td>3.824</td>
</tr>
<tr>
<td>power</td>
<td>44</td>
<td>6.260</td>
<td>4.150</td>
</tr>
</tbody>
</table>

Lesson 3  Attributes

3.1  View options: Information about word class

In the ‘View options’ menu, 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 ‘tag’ box, under ‘Attributes’.

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:

- lemma
- lempos
- sem
- textpart

If you choose ‘lemma’ 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):

In order to further their argument, the authors justify their use of ethnicity (which, again, may be easier for the personnel to make sure no shelf or a while (pretending to look for books), and take a great and careless leap beyond their use of the cor as interchangeable in the book)? Only a paragraph’s place. Classical music sounds throughout the store, books from the shelf she was looking at, she did not
If you choose ‘lempos’, you will see the same information with the addition of the word class (‘lemma’ + ‘Part-Of-Speech’):

In order to further their argument, the authors identify their use of ethnicity (which, again, may be easier for the personnel to make sure no theft or a while (pretending to look for books), and take take-V

taken take-V

takes take-V

took take-V

a great and careless leap beyond their use of the con
as interchangeable in the book). Only a paragraph is
place. Classical music sounds throughout the store, a
books from the shelf she was looking at, she did not

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’.

pluralism. By the late thirties, fifty such films/Q4.3 were produced a year, also serving as a charac
ter of the Brazilian national identity. The fi
nal symbol. The production of Brazilian films/Q4.3 brought a major contribution in the transfor

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.

using a look up table that uses empirical define any parameters you use in your and benefits of using these tools. Within
he rear wing is a crucial component on a
ship must comply to. </p><p> The
acial regulations which are supplied to the rear wing is a crucial component on a
gs that were used during the 2004-2005>
> (as accessed 17/01/2006) </p><p>

formulae running-text

formula heading

Formula abstract

Formula abstract

Formula running-text

Formula running-text

Formula running-text

Formula bibliography

and calibrates the required values. </p>

EXPE

and in the construction of your model. C is the

One CAD and Computational Fluid Dynamics:

One car and its design can greatly affect the pe

One World Championship is one such series th

One teams showing the FIA rules regarding rea

One car and its design can greatly affect the pe

One season. These illustrate clearly the various

One 2006 Technical Regulations </p><p> 2
Lesson 4    Corpus Query Types

So far, you have made queries by typing words into the ‘Query’ box. In this lesson, you will find out about other ways to make queries: ‘lemma’, ‘phrase’ and ‘word form’.

4.1  Basic query types
Click on ‘Query Type’ on the first query page:

You can choose between various types of query:

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 tick the ‘match case’ box, 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 ‘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 below, 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:

```
as John Hatfield in 1803 took minutes to
position. In addition to take the ortho-
logical positivists being forced to take the
before each reading was taken the necessary
organizations, and for managers to take a
government, which failed to take a decisive
incompetent, how did Andersen stuff move to take
into what things are without taking any
began to dramatically increase and to take a
intensity was being recorded. Care was taken to
further education (FE) must take a prominent
compny’s foundation. He takes a prominent
 technique forces the reader to take the
general, Taylor and Francis take a good
both governments will soon take a stronger
programme was finished, I have taken up a
that: <p>Regression involves taking the
and would therefore have to take his class
that Catusus considers taking a feminine
in emotions of the listeners. He took the
Brooke - Bohm Theory has not taken pole
hierarchy but later on in her career took up the
Four readings shall be taken at different
therefore when the readings were taken, the
```

If you want to see how two or three words co-occur within a short span of text, you can use the ‘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 below, 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:

```
as John Hatfield in 1803 took minutes to
position. In addition to take the ortho-
logical positivists being forced to take the
before each reading was taken the necessary
organizations, and for managers to take a
government, which failed to take a decisive
incompetent, how did Andersen stuff move to take
into what things are without taking any
began to dramatically increase and to take a
intensity was being recorded. Care was taken to
further education (FE) must take a prominent
compny’s foundation. He takes a prominent
 technique forces the reader to take the
general, Taylor and Francis take a good
both governments will soon take a stronger
programme was finished, I have taken up a
that: <p>Regression involves taking the
and would therefore have to take his class
that Catusus considers taking a feminine
in emotions of the listeners. He took the
Brooke - Bohm Theory has not taken pole
hierarchy but later on in her career took up the
Four readings shall be taken at different
therefore when the readings were taken, the
```
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) at level 4 (Masters Level of study).

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. 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 texts produced only by speakers of English as a first language.

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.

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 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.
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).

<table>
<thead>
<tr>
<th>text discipline</th>
<th>Freq</th>
<th>Rel [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Law</td>
<td>94</td>
<td>805.1</td>
</tr>
<tr>
<td>Business</td>
<td>19</td>
<td>149.4</td>
</tr>
<tr>
<td>Engineering</td>
<td>10</td>
<td>48.2</td>
</tr>
<tr>
<td>Philosophy</td>
<td>6</td>
<td>65.0</td>
</tr>
<tr>
<td>History</td>
<td>4</td>
<td>47.8</td>
</tr>
<tr>
<td>Psychology</td>
<td>2</td>
<td>24.2</td>
</tr>
<tr>
<td>English</td>
<td>2</td>
<td>21.7</td>
</tr>
<tr>
<td>Computer Science</td>
<td>2</td>
<td>25.4</td>
</tr>
<tr>
<td>Comparative American Studies</td>
<td>2</td>
<td>31.0</td>
</tr>
<tr>
<td>Agriculture</td>
<td>2</td>
<td>17.1</td>
</tr>
<tr>
<td>Other</td>
<td>1</td>
<td>19.1</td>
</tr>
<tr>
<td>Politics</td>
<td>1</td>
<td>18.4</td>
</tr>
<tr>
<td>Physio</td>
<td>1</td>
<td>16.9</td>
</tr>
<tr>
<td>Hospitality, Leisure &amp; Tourism Management</td>
<td>1</td>
<td>12.3</td>
</tr>
<tr>
<td>Health</td>
<td>1</td>
<td>14.2</td>
</tr>
<tr>
<td>Classics</td>
<td>1</td>
<td>14.0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>text genre</th>
<th>Freq</th>
<th>Rel [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Problem question</td>
<td>59</td>
<td>303.1</td>
</tr>
<tr>
<td>Essay</td>
<td>48</td>
<td>64.0</td>
</tr>
<tr>
<td>Critique</td>
<td>22</td>
<td>114.7</td>
</tr>
<tr>
<td>Explanation</td>
<td>4</td>
<td>34.8</td>
</tr>
<tr>
<td>Case study</td>
<td>2</td>
<td>17.1</td>
</tr>
<tr>
<td>Research report</td>
<td>1</td>
<td>26.4</td>
</tr>
<tr>
<td>Narrative recount</td>
<td>1</td>
<td>26.4</td>
</tr>
<tr>
<td>Methodology recount</td>
<td>1</td>
<td>5.0</td>
</tr>
<tr>
<td>Empathy writing</td>
<td>1</td>
<td>57.6</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Collocation candidates</th>
<th>Freq</th>
<th>T-score</th>
<th>Ml</th>
<th>logDice</th>
</tr>
</thead>
<tbody>
<tr>
<td>for</td>
<td>57</td>
<td>7.420</td>
<td>5.862</td>
<td>5.086</td>
</tr>
<tr>
<td>to</td>
<td>42</td>
<td>5.960</td>
<td>3.639</td>
<td>2.866</td>
</tr>
<tr>
<td>.</td>
<td>14</td>
<td>2.244</td>
<td>1.321</td>
<td>0.548</td>
</tr>
<tr>
<td>if</td>
<td>8</td>
<td>2.778</td>
<td>5.807</td>
<td>5.009</td>
</tr>
<tr>
<td>in</td>
<td>5</td>
<td>1.134</td>
<td>1.020</td>
<td>0.247</td>
</tr>
<tr>
<td>on</td>
<td>3</td>
<td>1.332</td>
<td>2.115</td>
<td>1.337</td>
</tr>
<tr>
<td>under</td>
<td>3</td>
<td>1.700</td>
<td>5.764</td>
<td>4.925</td>
</tr>
</tbody>
</table>

This shows us that ‘for’ and ‘to’ are the most frequent right collocates of ‘liable’.
5.2 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 ‘word’, ‘tag’, ‘lemma’, ‘lempos’, ‘semantic tag’ or ‘textpart’ (for explanations of these, see 3.2).

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’.
You can also search for n-grams or clusters of words. This search will create a list of 4-grams, clusters of four words which occur together.

The Word Sketch option will provide a full picture of the collocations

Click on the question mark for more information.

The results of this query are shown below.
<table>
<thead>
<tr>
<th>modifier</th>
<th>2288</th>
<th>1.9</th>
<th>pp_of</th>
<th>804</th>
<th>2.9</th>
<th>object_of</th>
<th>659</th>
<th>1.5</th>
<th>and/or</th>
<th>574</th>
<th>0.9</th>
<th>subject_of</th>
<th>350</th>
<th>1.6</th>
</tr>
</thead>
<tbody>
<tr>
<td>detailed</td>
<td>62</td>
<td>9.52</td>
<td>variance</td>
<td>15</td>
<td>8.73</td>
<td>perform</td>
<td>53</td>
<td>9.42</td>
<td>Interpretation</td>
<td>16</td>
<td>7.69</td>
<td>reveal</td>
<td>17</td>
<td>7.62</td>
</tr>
<tr>
<td>statistical</td>
<td>55</td>
<td>9.44</td>
<td>data</td>
<td>34</td>
<td>7.1</td>
<td>conduct</td>
<td>26</td>
<td>8.64</td>
<td>visualisation</td>
<td>7</td>
<td>7.44</td>
<td>confirm</td>
<td>5</td>
<td>6.93</td>
</tr>
<tr>
<td>swot</td>
<td>49</td>
<td>9.43</td>
<td>text</td>
<td>11</td>
<td>6.89</td>
<td>undertake</td>
<td>15</td>
<td>8.34</td>
<td>testing</td>
<td>7</td>
<td>7.44</td>
<td>show</td>
<td>40</td>
<td>6.52</td>
</tr>
<tr>
<td>fe-</td>
<td>27</td>
<td>6.96</td>
<td>shaft</td>
<td>4</td>
<td>6.69</td>
<td>carry</td>
<td>22</td>
<td>8.1</td>
<td>mmr-</td>
<td>3</td>
<td>7.31</td>
<td>enable</td>
<td>7</td>
<td>6.36</td>
</tr>
<tr>
<td>regression</td>
<td>37</td>
<td>8.64</td>
<td>category</td>
<td>5</td>
<td>6.19</td>
<td>complete</td>
<td>12</td>
<td>8.04</td>
<td>summary</td>
<td>3</td>
<td>7.07</td>
<td>suggest</td>
<td>15</td>
<td>6.03</td>
</tr>
<tr>
<td>comparative</td>
<td>30</td>
<td>8.57</td>
<td>plate</td>
<td>4</td>
<td>6.11</td>
<td>run</td>
<td>8</td>
<td>6.89</td>
<td>collection</td>
<td>3</td>
<td>6.9</td>
<td>assume</td>
<td>5</td>
<td>6.0</td>
</tr>
<tr>
<td>discourse</td>
<td>28</td>
<td>8.53</td>
<td>sample</td>
<td>10</td>
<td>6.01</td>
<td>base</td>
<td>26</td>
<td>6.77</td>
<td>conclusion</td>
<td>3</td>
<td>6.89</td>
<td>demonstrate</td>
<td>5</td>
<td>5.86</td>
</tr>
<tr>
<td>critical</td>
<td>34</td>
<td>8.31</td>
<td>nature</td>
<td>11</td>
<td>5.87</td>
<td>undergo</td>
<td>4</td>
<td>6.72</td>
<td>calculation</td>
<td>3</td>
<td>6.82</td>
<td>prove</td>
<td>5</td>
<td>6.63</td>
</tr>
<tr>
<td>finite</td>
<td>22</td>
<td>8.18</td>
<td>result</td>
<td>18</td>
<td>5.86</td>
<td>incorporate</td>
<td>5</td>
<td>6.68</td>
<td>design</td>
<td>13</td>
<td>6.71</td>
<td>focus</td>
<td>4</td>
<td>5.58</td>
</tr>
<tr>
<td>far</td>
<td>26</td>
<td>8.17</td>
<td>ratio</td>
<td>5</td>
<td>5.63</td>
<td>provide</td>
<td>25</td>
<td>6.55</td>
<td>description</td>
<td>6</td>
<td>6.59</td>
<td>look</td>
<td>5</td>
<td>5.58</td>
</tr>
<tr>
<td>genetic</td>
<td>25</td>
<td>8.11</td>
<td>section</td>
<td>6</td>
<td>5.62</td>
<td>apply</td>
<td>10</td>
<td>6.4</td>
<td>discussion</td>
<td>6</td>
<td>6.54</td>
<td>follow</td>
<td>7</td>
<td>5.55</td>
</tr>
<tr>
<td>isolate</td>
<td>20</td>
<td>8.07</td>
<td>mechanism</td>
<td>4</td>
<td>5.6</td>
<td>lack</td>
<td>4</td>
<td>6.36</td>
<td>assessment</td>
<td>3</td>
<td>6.5</td>
<td>indicate</td>
<td>5</td>
<td>5.32</td>
</tr>
<tr>
<td>in-depth</td>
<td>18</td>
<td>7.97</td>
<td>demand</td>
<td>6</td>
<td>5.58</td>
<td>employ</td>
<td>5</td>
<td>6.28</td>
<td>comparison</td>
<td>4</td>
<td>5.94</td>
<td>help</td>
<td>4</td>
<td>5.27</td>
</tr>
<tr>
<td>post</td>
<td>17</td>
<td>7.92</td>
<td>evidence</td>
<td>8</td>
<td>5.4</td>
<td>present</td>
<td>9</td>
<td>6.26</td>
<td>result</td>
<td>16</td>
<td>5.72</td>
<td>seem</td>
<td>7</td>
<td>5.27</td>
</tr>
<tr>
<td>network</td>
<td>20</td>
<td>7.87</td>
<td>reason</td>
<td>6</td>
<td>5.35</td>
<td>combine</td>
<td>4</td>
<td>6.21</td>
<td>reviewer</td>
<td>4</td>
<td>5.64</td>
<td>require</td>
<td>6</td>
<td>5.13</td>
</tr>
<tr>
<td>thorough</td>
<td>17</td>
<td>7.85</td>
<td>organisation</td>
<td>5</td>
<td>5.34</td>
<td>require</td>
<td>16</td>
<td>6.07</td>
<td>report</td>
<td>5</td>
<td>5.51</td>
<td>identify</td>
<td>4</td>
<td>4.98</td>
</tr>
<tr>
<td>above</td>
<td>18</td>
<td>7.71</td>
<td>structure</td>
<td>10</td>
<td>5.33</td>
<td>focus</td>
<td>6</td>
<td>6.01</td>
<td>test</td>
<td>5</td>
<td>5.51</td>
<td>come</td>
<td>4</td>
<td>4.94</td>
</tr>
<tr>
<td>financial</td>
<td>25</td>
<td>7.58</td>
<td>design</td>
<td>5</td>
<td>5.26</td>
<td>enable</td>
<td>6</td>
<td>5.99</td>
<td>action</td>
<td>4</td>
<td>4.72</td>
<td>determine</td>
<td>4</td>
<td>4.88</td>
</tr>
<tr>
<td>empirical</td>
<td>16</td>
<td>7.55</td>
<td>situation</td>
<td>5</td>
<td>5.14</td>
<td>propose</td>
<td>4</td>
<td>5.94</td>
<td>data</td>
<td>6</td>
<td>4.64</td>
<td>provide</td>
<td>8</td>
<td>4.46</td>
</tr>
<tr>
<td>cost-benefit</td>
<td>23</td>
<td>7.54</td>
<td>issue</td>
<td>7</td>
<td>5.12</td>
<td>use</td>
<td>51</td>
<td>5.91</td>
<td>fact</td>
<td>4</td>
<td>4.53</td>
<td>allow</td>
<td>2</td>
<td>4.41</td>
</tr>
<tr>
<td>swot-</td>
<td>13</td>
<td>7.54</td>
<td>relation</td>
<td>5</td>
<td>5.1</td>
<td>limit</td>
<td>4</td>
<td>5.67</td>
<td>method</td>
<td>6</td>
<td>4.47</td>
<td>use</td>
<td>17</td>
<td>4.34</td>
</tr>
<tr>
<td>data</td>
<td>54</td>
<td>7.53</td>
<td>pattern</td>
<td>4</td>
<td>5.09</td>
<td>involve</td>
<td>7</td>
<td>5.52</td>
<td>evidence</td>
<td>4</td>
<td>4.45</td>
<td>find</td>
<td>7</td>
<td>4.31</td>
</tr>
<tr>
<td>dimensional</td>
<td>12</td>
<td>7.37</td>
<td>performance</td>
<td>5</td>
<td>5.06</td>
<td>offer</td>
<td>5</td>
<td>5.44</td>
<td>aspproach</td>
<td>4</td>
<td>4.29</td>
<td>take</td>
<td>2</td>
<td>3.99</td>
</tr>
<tr>
<td>historical</td>
<td>16</td>
<td>7.32</td>
<td>project</td>
<td>5</td>
<td>4.92</td>
<td>include</td>
<td>7</td>
<td>5.28</td>
<td>study</td>
<td>3</td>
<td>3.98</td>
<td>give</td>
<td>5</td>
<td>3.56</td>
</tr>
<tr>
<td>ratio</td>
<td>22</td>
<td>7.24</td>
<td>case</td>
<td>7</td>
<td>4.92</td>
<td>allow</td>
<td>9</td>
<td>5.2</td>
<td>factor</td>
<td>4</td>
<td>3.85</td>
<td>make</td>
<td>4</td>
<td>2.63</td>
</tr>
</tbody>
</table>
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:

```
[tag="NP1"]
```

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 NP1. A full list of word class codes used in the BAWE corpus can be found at:

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 you are using a regular expression, as you did in 5.2). Here are some examples:

V.G
VB.
V..

To find all the instances of a specified part of speech, use [tag = “X”]. For example [tag="V.G"] finds all the –ing participles in the corpus, and [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 [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 \([\text{lemma} = \text{"impact"} \& \text{tag} = \text{"V.."}]\) will find only the verb forms of the specified lemma (in this case impact).

\[
\begin{array}{|c|c|}
\hline
\text{Query Type:} & \text{CQL} \\
\hline
\text{CQL:} & \left[\text{lemma} = \text{"impact"} \& \text{tag} = \text{"V.."}\right] \\
\hline
\end{array}
\]

The query \([\text{lemma} = \text{"different"}] \& \text{tag} = \text{"I.|R."}]\) finds the prepositions and adverbs following different.

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

\[
\begin{array}{|c|c|}
\hline
\text{Query Type:} & \text{CQL} \\
\hline
\text{CQL:} & \left[\text{lemma}=\text{"fast"} \& \text{tag} != \text{"J.."]}\right] \\
\hline
\end{array}
\]

The query \([\text{lemma}="\text{talk}"] \& \text{tag} = \text{"V.."}] \& \text{word} != \text{"about"]}\) finds the verb talk followed by anything but about.

The query \([\text{tag} = \text{"VB(D|M|R|Z)"] \& \text{tag} = \text{"VVN"]}\) 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 \([\text{tag} = \text{"VB(D|M|R|Z)"]} [\text{tag} = \text{"VVN"]}\) finds am, are, is, were or was followed by the past participle of a lexical verb, with at most four words in between.

\[
\begin{array}{|c|c|}
\hline
\text{Query Type:} & \text{CQL} \\
\hline
\text{CQL:} & \left[\text{tag} = \text{"VB(D|M|R|Z)"]} [\text{tag} = \text{"VVN"]}\right] \\
\hline
\end{array}
\]

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 \([\text{word}=\text{"he"]}\) within <text l1="English”/> looks for he only within those files produced by writers whose first language is English.

\[
\begin{array}{|c|c|}
\hline
\text{Query Type:} & \text{CQL} \\
\hline
\text{CQL:} & \text{[word="he"] within <text l1="English”/>} \\
\hline
\end{array}
\]
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="w4+" /> looks for think within quotations:

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.

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.

Similarly [lemma="government" & !textpart="running-text"] finds all uses of ‘government’ outside the running text.