

Using Sketch Engine with BAWE

**Hilary Nesi
&
Paul Thompson**

Version III 2014

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’

Open corpora

Corpus name	Language	Tokens	Words		
Brown	English	1,175,675	1,007,299		
ACL Anthology Reference Corpus (ARC)	English	49,348,397	38,792,655		
British Academic Written English Corpus (BAWE)	English	8,336,262	6,964,411		
British Academic Spoken English Corpus (BASE)	English	1,252,256	1,186,290		

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

Simple query:

[Query types](#) [Context](#) [Text types](#)

Query type simple lemma phrase word character CQL

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

Query **factor** 4,548 (545.6 per million)

Page 1 of 228 [Next](#) | [Last](#)

BAWE-1.txt... not the only reason for the continuation of racism, economic **factors** have also had an impact. The racism apparent at the time of

BAWE-1.txt... involved in the actual area (May, 1997). Human judgement is a key **factor** in this criticism, but as the decision process shows it is

BAWE-1.txt... pull in the decision, introduced by Morrell (1994), looking at **factors** identified as ‘pulling’ women into motherhood; psychological

BAWE-1.txt... pulling’ women into motherhood; psychological, social and economic **factors**, as well as looking at the influences of ‘push’ factors such

BAWE-1.txt... economic factors, as well as looking at the influences of ‘push’ **factors** such as national discourses on motherhood and social policy

BAWE-1.txt..., considering the influence of personal, social and economic **factors**, and changing opportunities for women. In considering all of

BAWE-1.txt... considered a reality (Gittins, 1993). However as well as personal **factors** attracting women to motherhood, economic and social factors

BAWE-1.txt... factors attracting women to motherhood, economic and social **factors** may also be influential. Lancaster has argued that children

BAWE-1.txt... in Andorka, 1978, p364). However it is now felt that economic **factors** do not influence the decision as much as in previous times,

BAWE-1.txt... are not needed to provide a source of income. Instead social **factors** have more of an input. Coleman claims that children are a form

BAWE-1.txt... often combine in a complex way, however they are not the only **factors** that affect the decision. National discourses on motherhood

BAWE-1.txt... 1996). These discourses in society influence the more personal **factors** that are involved on the decision to become a mother, as they

BAWE-1.txt... society influence and combine with personal, economic and social **factors** to create a complex combination of reasons as to why women

BAWE-1.txt... again (Gillespie, 2000). There is also the influence of economic **factors** to consider. Children are no longer economic assets, but are

BAWE-1.txt... to have (Andorka, 1978). This links quite closely to social **factors**, as a desired lifestyle can be quite important when considering

BAWE-1.txt... shows that as well as there being personal, economic and social **factors** influencing the decision not to have children, changes in employment

BAWE-1.txt... discourses emerging on motherhood and reproduction mean that these **factors** are possible in modern society. Therefore it is a complex combination

BAWE-1.txt... for women, and this impacts upon personal, economic and social **factors**, including security, happiness and social capital, that lead

BAWE-1.txt... do not have children is a complex process influenced by many **factors**, and based upon a variety of discourses and opportunities ingrained

BAWE-1.txt... outline why mass unemployment emerged at this time, including **factors** of global changes, technological advances and government policies

Page 1 of 228 [Next](#) | [Last](#)

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:

View options

Attributes	Structures	References
<input checked="" type="checkbox"/> word	<text>	Token number
<input type="checkbox"/> tag	<back>	text_discgroup
<input type="checkbox"/> lemma	<body>	text_discipline
<input type="checkbox"/> lempos	<div1>	text_educ
<input type="checkbox"/> sem	<div2>	text_genre
<input type="checkbox"/> textpart	<div3>	text_grade
	<div4>	text.l1
	<div5>	text.level
	<div6>	text.sex
	<docTitle>	text.studentage
	<epigraph>	div1.n
	<figure>	div1.type
	<formula>	div2.type

Page size (number of lines): 20
KWIC Context size (number of characters): 40

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

Icon for one-click sentence copying
 Allow multiple lines selection

XML template for one-click copying: [v]
Change View Options

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.

PS,Methodology recount	calculated by dividing the limit moment by the yield moment. Shape	Factor = 181.6 / 120.93 = 1.5, this is the typical value for a rectangular
PS,Exercise	is the identity matrix we get the Eigen Values as: Stiffness	Factor = Order of Stiffness Factor = We can write the system in the
LS,Methodology recount	Second dilution = Therefore, total dilution = And, total dilution	Factor = Therefore, it should be reported that there are 5.75 × 10
PS,Exercise	get the Eigen Values as: Stiffness Factor = Order of Stiffness	Factor = We can write the system in the form of: Where Taking the
AH,Essay	case studies, he noted that there was a combination of internal	factors - "urban disorder, popular heterodox religious movements led
AH,Essay	, T. Mackay, I, & Flege, J. 2001). They provided several	factors - age of L2 learning (AOL), length of residence (LOR), gender
LS,Case study	the times when he has become acutely ill and look for a common	factor - an exercise such as ranking of life events in order of perceived
LS,Essay	those without such experiences, and how psychological and social	factors - and in particular the family dynamic - have a large bearing
SS,Case study	flights to Buenos Aires from those areas may prove a negative	factor - Buenos Aires is cheap, but it remains fairly expensive to
SS,Essay	Cragg et al 2002, Chan et al 2006). The effect of the forth	factor - Business and IT executives' relationship management has been
AH,Essay	establishing that this factor is a subset of the general external	factor - challenges posed by the Europeans, acknowledging the centrality
PS,Critique + Critique	Conflicts could be broadly divided into two progress defining	factors - functional and dysfunctional. Traditionally conflicts were
AH,Essay	musical ability has not been identified. Strangely, a similar	factor - imitation ability in Flege's original report (Flegeal. 1995
PS,Methodology recount	tetrahedral of square planar geometry. Firstly, there are the steric	factors - it is easier to adopt a tetrahedral configuration if the
PS,Methodology recount	nom value to be used when calculating the stress concentration	factor - k t. Stress concentration and measuring stresses in materials
AH,Essay	suprasegmental qualities. However, it is a little confusing that the two	factors - L2 language use and L1 language use can be analyzed separately
SS,Methodology recount	balance in response to devaluation of real exchange rate to five	factors - recognition, decision, delivery, replacement and production
SS,Essay	instance, "economics institutions are defined as a system of social	factors - such as rules, beliefs, norms and organisations - that guide
SS,Essay	by the white to segregate the black society. Also religious	factors - the Christian church calling homosexuality a sin is another
AH,Explanation	help determine what code they use and states "Certain social	factors - who you are talking to, the social context of the talk, the
PS,Critique	the following properties; Collector area, = 8m 2Heat removal	factor , = 0.8Collector loss coefficient, = 3 W/m 2/KCollector slope
PS,Design specification	resistance caused by an induced strain is defined as the Gauge	Factor , 'G', which represents the strain sensitivity, where: Eqn.
SS,Critique	committing a POCA offence. Following guidelines is only a mitigating	factor , "it is not per se a defence ... to show that the guidance
SS,Literature survey	and suggests that print on demand technology has been a main	factor , "It's now so inexpensive to become a publisher" (Michael Healy
LS,Case study	bronchial hyperresponsivity. Remodelling is achieved by a number of	factors , (i) epithelium of conducting airways is damaged with loss

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.

Icon for one-click sentence copying
 Allow multiple lines selection
 Do not use Flash for copying to clipboard
 Checkbox for selecting lines
 Show line numbers
 Shorten long references

 XML template for one-click copying: ▼

Change View Options

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

PS,Methodo...	calculated by dividing the limit moment by the yield moment. Shape	Factor = 181.6 / 120.93 = 1.5, this is the typical value for a rectangular
PS,Exercise	is the identity matrix we get the Eigen Values as: Stiffness	Factor = Order of Stiffness Factor = We can write the system in the
LS,Methodo...	Second dilution = Therefore, total dilution = And, total dilution	Factor = Therefore, it should be reported that there are 5.75 × 10
PS,Exercise	get the Eigen Values as: Stiffness Factor = Order of Stiffness	Factor = We can write the system in the form of: Where Taking the
AH,Essay	case studies, he noted that there was a combination of internal	factors - "urban disorder, popular heterodox religious movements led
AH,Essay	, T. Mackay, I, & Flége, J. 2001). They provided several	factors - age of L2 learning (AOL), length of residence (LOR), gender
LS,Case st...	the times when he has become acutely ill and look for a common	factor - an exercise such as ranking of life events in order of perceived
LS,Essay	those without such experiences, and how psychological and social	factors - and in particular the family dynamic - have a large bearing
SS,Case st...	flights to Buenos Aires from those areas may prove a negative	factor - Buenos Aires is cheap, but it remains fairly expensive to
SS,Essay	Cragg et al 2002, Chan et al 2006). The effect of the forth	factor - Business and IT executives' relationship management has been
AH,Essay	establishing that this factor is a subset of the general external	factor - challenges posed by the Europeans, acknowledging the centrality
PS,Critiqu...	Conflicts could be broadly divided into two progress defining	factors - functional and dysfunctional. Traditionally conflicts were
AH,Essay	musical ability has not been identified. Strangely, a similar	factor - imitation ability in Flége's original report (Flégeal. 1995
PS,Methodo...	tetrahedral of square planar geometry. Firstly, there are the steric	factors - it is easier to adopt a tetrahedral configuration if the
PS,Methodo...	nom value to be used when calculating the stress concentration	factor - k t. Stress concentration and measuring stresses in materials
AH,Essay	suprasegmental qualities. However, it is a little confusing that the two	factors - L2 language use and L1 language use can be analyzed separately
SS,Methodo...	balance in response to devaluation of real exchange rate to five	factors - recognition, decision, delivery, replacement and production
SS,Essay	instance, "economics institutions are defined as a system of social	factors - such as rules, beliefs, norms and organisations - that guide
SS,Essay	by the white to segregate the black society. Also religious	factors - the Christian church calling homosexuality a sin is another
AH,Explana...	help determine what code they use and states "Certain social	factors - who you are talking to, the social context of the talk, the
PS,Critique	the following properties; Collector area, = 8m 2Heat removal	factor , = 0.8Collector loss coefficient, = 3 W/m 2/KCollector slope
PS,Design ...	resistance caused by an induced strain is defined as the Gauge	Factor , 'G', which represents the strain sensitivity, where: Eqn.
SS,Critique	committing a POCA offence. Following guidelines is only a mitigating	factor , "it is not per se a defence ... to show that the guidance
SS,Literat...	and suggests that print on demand technology has been a main	factor , "It's now so inexpensive to become a publisher" (Michael Healy

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:

SS,Critique interpretative stage of research may have been affected by this factor , but it could also be argued that this deductive approach may

SS,Essay suggested that workers were not merely driven by economical factor , but the informal side (i.e. informal norms, peer pressure)

LS,Exercise

Page 1 of 114

< expand left other, productivity rises as a result of group cohesiveness. Informal groups have great effect on the worker's temperament and well-being. The "bank wiring room experiment" suggested that workers were not merely driven by economical factor , but the informal side (i.e. informal norms, peer pressure) of the organisation was even more important than the formal rules and hierarchy side. One of the essential elements of human relations, identified by Schoen expand right >

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

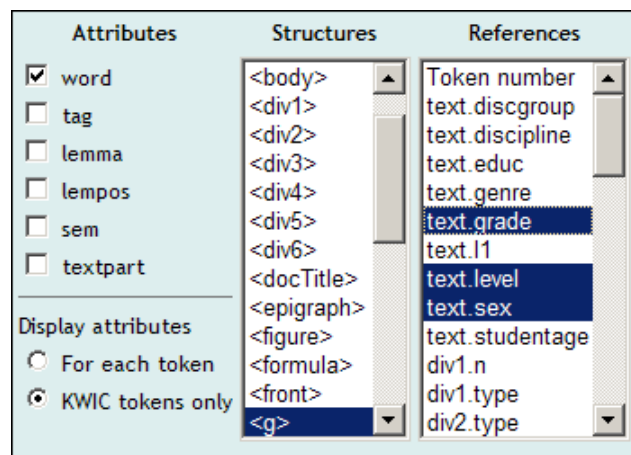
1.3 View options: Information about assignments

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:

text.discgroup SS
text.discipline Sociology
text.educ UKA
text.genre Essay
text.grade D
text.ll English
text.level 1
text.sex f
text.studentage 25-
Word Count 1632

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:

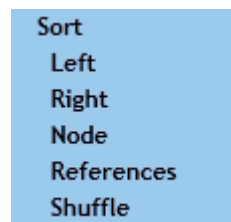


Here we have selected three categories of information (grade, level and gender). To select more than one, you need to keep the Control key pressed down as you make each selection. The resulting display looks like this:

D,1,f only reason for the continuation of racism, economic **factors** have also had an impact. <p><p> The racism apparent
 M,1,f actual area (May, 1997). <p><p> Human judgement is a key **factor** in this criticism, but as the decision process shows
 D,2,f decision, introduced by Morrell (1994), looking at **factors** identified as 'pulling' women into motherhood; psychological
 D,2,f into motherhood; psychological, social and economic **factors**, as well as looking at the influences of 'push' factors
 D,2,f factors, as well as looking at the influences of 'push' **factors** such as national discourses on motherhood and social
 D,2,f considering the influence of personal, social and economic **factors**, and changing opportunities for women. In considering
 D,2,f (Gitins, 1993). <p><p> However as well as personal **factors** attracting women to motherhood, economic and social
 D,2,f attracting women to motherhood, economic and social **factors** may also be influential. Lancaster has argued that
 D,2,f 1978, p364). However it is now felt that economic **factors** do not influence the decision as much as in previous
 D,2,f needed to provide a source of income. Instead social **factors** have more of an input. <p><p> Coleman claims that
 D,2,f combine in a complex way, however they are not the only **factors** that affect the decision. National discourses on
 D,2,f discourses in society influence the more personal **factors** that are involved on the decision to become a mother
 D,2,f influence and combine with personal, economic and social **factors** to create a complex combination of reasons as to
 D,2,f 2000). <p><p> There is also the influence of economic **factors** to consider. Children are no longer economic assets
 D,2,f Andorka, 1978). <p><p> This links quite closely to social **factors**, as a desired lifestyle can be quite important when
 D,2,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.

the same time, they may find the "critical success **factors** " (Boddy, 1998:165) as Boddy mentioned. Besides the full term, are at "increased risk of biological risk **factors** " (Menyuk, 1995: 3), especially those born below thirty to all natural-kind terms) is "decomposed into two **factors** " (Segal 2000, p.27). He accepts that there must be important factors in risk assessment than human capital **factors** ", which indirectly suggests that finance could be in order to know the performance of the "Forgetting **factor** ".
Part B Implementation and Test
The task of presented.
Fig.7: The performance of the "Forgetting **factor** ".
The figure 7 shows the result of the modified of the modified algorithm including the "Forgetting **factor** ". The convergent rate seems to be more efficient part, for the decline of the music industry. Demand **factors** "appear to play at least as great role in the current indicated it in horizontal cases were "countervailing **factors** " arising more than two years after entities have taken not to make price the overriding Competitive **Factor** "
Chisnall, Peter M, Strategic Business Marketing latest possible moment in hope that the "feel-good **factor** " following economic prosperity would turn around

A summary of page features:

user: anonymous corpus: British Academic Written English Corpus (BAWE)

Concordance Word List ?

Save

View options

KWIC

Sentence

Sort

Left

Right

Node

References

Shuffle

Sample

Filter

Frequency

Node

Node

Doc Id

Text Types

Collocations

ConcDesc

?

Menu position

Corpus: British Academic Written English Corpus (BAWE)

Hits: 298 (35.7 per million)

Agriculture,Methodology recount

Agriculture,Methodology recount

Archaeology,Methodology recount

Archaeology,Methodology recount

Biology,Methodology recount

Biology,Methodology recount

Biology,Methodology recount

Biological Sciences,Methodology recount

Biological Sciences,Methodology recount

Biological Sciences,Methodology recount

Biological Sciences,Methodology recount

Biological Sciences,Methodology recount

Biological Sciences,Methodology recount

Biological Sciences,Methodology recount

Biological Sciences,Methodology recount

Biological Sciences,Methodology recount

Biological Sciences,Methodology recount

Biological Sciences,Methodology recount

Biological Sciences,Methodology recount

Chemistry,Methodology recount

Chemistry,Methodology recount

Chemistry,Methodology recount

Chemistry,Methodology recount

Chemistry,Methodology recount

Chemistry,Methodology recount

Chemistry,Methodology recount

Chemistry,Methodology recount

Chemistry,Methodology recount

Chemistry,Methodology recount

The blue words show you the discipline and the genre family

Click on 'Left' or 'Right' to sort in alphabetical order the words to the left or right of the search

If you click on 'Node' the red words will be listed in alphabetical order

Click on 'References' to put the blue words in alphabetical order

Click on any of the references in blue to see more information about the assignment

Click on any of the red words to see more context

and the average composition	can be seen	in Table 2. Diets were made isonitrogenous by calculating
analysis for the compound feed	can be seen	in Table 3. </p> Table 2: Mean nutrient composition
ed to definite butchery. This	can be seen	when there is a clean cut, often with an area sticking
. The re-crystallised salts observed in the sample	can only be seen	in cross-polarised light and appear as coatings on
has on its rate of reproduction. </p><p> Some clones	can also be seen	to be more affected by the varying plant quality
moved P element carrying in the w+ gene. </p><p> As	can be seen	, there were approximately 10 times more white eyed
little p53 growth at all, but on what there is blue	can be seen	. There is much more lamin growth, showing more diploids
of oxygen evolution in isolated chloroplasts. <p> As	can be seen	from the results above, when DCMU, a herbicide, is
growth, showing more diploids were produced and blue	can be seen	here also. This implies that the reporter gene MEL-1
and e	can be seen	in figures 5 and 6. </p> Figure 5 Staining of 3rd instar
show	can be seen	in graph 1. below, when pH increases, the enzyme
Spar	can be seen	in the amount the shell has reduced in thickness.
eggs, as a slow recovery to the eggshells thickness	can be seen	in the results. Eventually both species eggshells
diagram below shows the generalised amino acid: </p><p> It	can be seen	that apart from having an amino group (A terminus
cells. </p><p> Referring to Appendix 1, Table 2, it	can be seen	that in samples two to four for the X antibody that
AH109-lamin = 1716 cfu/µg DNA, Y187-SV40 = 10 cfu/µg DNA. It	can be seen	that the most efficient was p53 and the least efficient
after DDT was introduced. Through this experiment it	can be seen	that the Peregrine Falcons eggs had suffered the
<p> Figure 2 shows a photograph of an X-a plate. It	can be seen	that there was very little p53 growth at all, but
with substrate. Although from the reaction: </p><p> it	can be seen	that water is also a reactant, the reaction is first
proportion of dropped aphids on Poor quality plants; clones	can be seen	to act in a number of ways and each to a different
blood cells (RBC) to form haemagglutination, which	can easily be seen	with naked eye. Another method of virus identification
reaction. </p><p> Effect of Changing the Monomer: This	can be seen	by comparison of the rate of the reaction for Polymer
Effect of time of addition: </p><p> The effect of this	can be seen	from the table derivativization of the starting material
d: - </p><p> As	can be seen	however the O-H stretch is lost. There were some
C-H stretches	can be seen	in results - "Plot of [F-] vs. Ionic radius", the
s. <p> This plot	can be seen	in solution C, KNO3, The Solubility of KIO4 is greatly
demonstrate the common-ion, "salting out" effect, this	can be seen	that a plot of logb against logA (see Figure 1) will
further simplified: </p><p> or </p><p> From equation 6 it	can be seen	that as the pressure increases; the temperature of
pressure against temperature was plotted (graph 1), it	can be seen	that only the three vibrations, A 1, B 1 and B 2
2ν point group. </p><p> From the character table, it	can be seen	that there are two occupied MO's and one vacant. </p>
molecular orbital energy diagram for the allyl anion, it	can be seen	when comparing their physical properties to their
pie bonds with oxygen. The effects of this bonding	can be seen	

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:

Sort
Left
Right
Node
References
Shuffle
Sample
Filter
Frequency
Node tags
Node forms
Doc IDs
Text Types
Collocations
ConcDesc
Visualize
?

Menu position

SS often combine in a complex way, however they are not the only **factors** that affect the decision. National discourses on motherhood
SS 1996). These discourses in society influence the more personal **factors** that are involved on the decision to become a mother, as they
SS society influence and combine with personal, economic and social **factors** to create a complex combination of reasons as to why women
SS again (Gillespie, 2000). There is also the influence of economic **factors** to consider. Children are no longer economic assets, but are
SS to have (Andorka, 1978). This links quite closely to social **factors**, as a desired lifestyle can be quite important when considering
SS shows that as well as there being personal, economic and social **factors** influencing the decision not to have children, changes in employment
SS discourses emerging on motherhood and reproduction mean that these **factors** are possible in modern society. Therefore it is a complex combination
SS for women, and this impacts upon personal, economic and social **factors**, including security, happiness and social capital, that lead
SS do not have children is a complex process influenced by many **factors**, and based upon a variety of discourses and opportunities ingrained
SS outline why mass unemployment emerged at this time, including **factors** of global changes, technological advances and government policies
SS Vickerstaff, 2003). However there were many more complicated **factors** involved in the steep decline during this period. One of the
SS causes of the illness rather than analysing social and cultural **factors** relevant to the women affected. In this view, female biological
SS perspective, London, Routledge, p.70Outline the social and political **factors** that led to the development of feminism in JapanThe development
SS Japan was also initiated by numerous other social and political **factors** that affected the political climate at that specific time and
SS forefront. This essay will deal with such social and political **factors** as these to account for the development of feminism in Japan
SS patriotic, nationalist fighters for the country. An additional **factor** affecting women was arguably their education. Debated by reformers
SS seen that a varied combination of changing social and political **factors** led to the development of feminism in Japan. Through increased
AH responsibility of the Americans may thus have been the deciding **factor** in the eventual communisation of the countries in Eastern Europe
SS proposed that [t]o separate, by scientific abstraction, these two **factors** of form and content which are in reality inseparably united
SS aspect of status or office and such is always influenced by **factors** other than the stipulations of the position itself' (David
AH the twentieth century marked the coming of 'modernity'. Other **factors** however, such as immigration and a rumbling social hierarchy

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.

Collocation candidates

Attribute:
word In the range from: -5 to 5

Minimum frequency in corpus: 5

Minimum frequency in given range: 3

Show functions: T-score MI MI3 log likelihood min. sensitivity logDice Sort by: T-score MI MI3 log likelihood min. sensitivity logDice

Make Candidate List

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.

Collocation candidates

Page [Next >](#)

	<u>Freq</u>	<u>T-score</u>	<u>MI</u>
p/n the	2574	46.118	3.458
p/n .	1744	37.665	3.350
p/n of	1646	36.938	3.481
p/n ,	1624	34.997	2.926
p/n and	1166	30.830	3.364
p/n to	1055	29.312	3.358
p/n a	967	28.891	3.817
p/n in	814	25.894	3.435
p/n is	761	25.397	3.655
p/n that	719	25.211	4.064
p/n as	613	23.390	4.177
p/n are	492	21.135	4.406
p/n be	484	20.560	3.934

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.

Collocation candidates

Page [Next >](#)

	<u>Freq</u>	<u>T-score</u>	<u>MI</u>
p/n Darcy	22	4.688	10.656
p/n fibroblast	6	2.448	10.618
p/n Fibroblast	4	1.999	10.518
p/n Va	4	1.999	10.518
p/n abiotic	6	2.447	10.255
p/n Forgetting	4	1.998	10.255
p/n T-box	4	1.998	10.255
p/n endowments	24	4.895	10.215
p/n eRF1	4	1.998	10.033
p/n gyromagnetic	5	2.234	9.992
p/n proximate	7	2.643	9.947
p/n aetiological	3	1.730	9.840

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.

Collocation candidates

Attribute: In the range from: to:

Minimum frequency in corpus:

Minimum frequency in given range:

Show functions: Sort by:

This is what you get if you choose a range of -1 to 0.

	<u>Freq</u>	<u>T-score</u>	<u>MI</u>
p/n other	209	14.031	5.086
p/n important	200	13.932	6.075
p/n these	135	11.150	4.630
p/n external	89	9.378	7.401
p/n risk	88	9.281	6.549
p/n key	77	8.665	6.314
p/n friction	60	7.726	8.567
p/n environmental	60	7.661	6.512
p/n transcription	56	7.470	9.180
p/n main	54	7.135	5.105
p/n many	56	7.030	4.045
p/n These	49	6.742	4.760
p/n economic	49	6.736	4.728
p/n social	49	6.506	3.824
p/n power	44	6.260	4.150

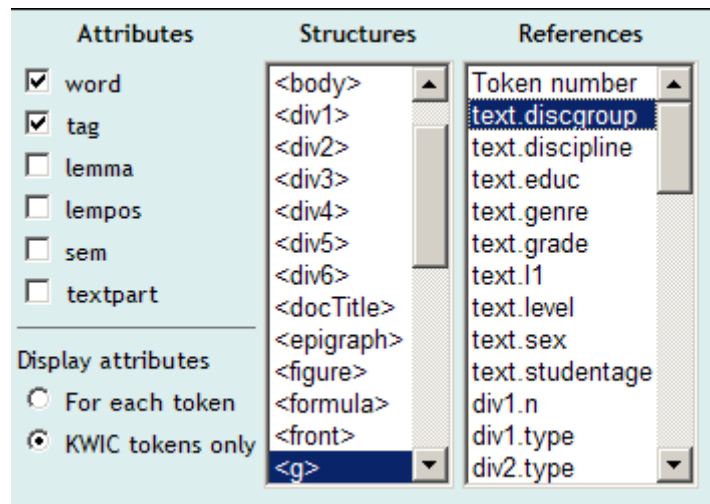
Notice the [p/n](#) links next to the collocates. If you click on 'n' you will only get the concordance lines for 'factors' and 'factor' where the collocate you have chosen does not precede the key word. ('n' stands for 'negative'.) If you click on 'p' you will get the concordance lines for 'factors' 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 had been no perceived need for revolution. Other **important factors** include that this was the first time that confidence in and deterministic fashion that mortality was the most **important factor** in shaping the population 'replacing one bald stereotype with phenomenon whilst simultaneously relegating other **important factors** to mere conditions or circumstances rather than causes of population is accepted that fertility and mortality were both **important factors** within the early modern period '...and a host of social, cultural even increasing metabolism. </p><p> Metabolism is an **important factor** in losing weight. If the basal metabolism increases, more energy not corrode such as metal ones would, which is an **important factor** to consider when designing an electric drill as they are likely support. However, there were also a number of other **important factors** which meant the Bolsheviks could carry out a coup - including degree of popular support but there were also other **important factors** such as Lenin's leadership and the unpopularity and mistakes Revolution (London, 1985), p. 41 </p><p> An extremely **important factor** in independence becoming more likely was the slide to war which and the Aborigines was undeniably the single most **important factor** . Civilization and Christianity were considered by missionaries was primarily attributable to a number of highly **important factors** such as the spread of Calvinist writings due to a vigorous The structure of the movement itself was a highly **important factor** in ensuring the success of Calvinism in the second half of Therefore, in conclusion, there were a number of highly **important factors** which led to Calvinism becoming the most successful brand of , American Colonies, p. 133 </p><p> A second highly **important factor** in determining the ability to create a stable society was the as Britain's first American colony. </p><p> A second **important factor** which was hugely influential in the failure of Roanoke was local Native American tribe was probably the most **important factor** , because they had initially provided food and provisions for relation to weapons, has long been considered an **important factor** in contributing to the demise of Native Americans. Axtell even with Natives. Trade was undoubtedly the single most **important factor** . To be precise, the turning point at which the Natives became transcription. </p><p> Transposition is thought to be an **important factor** in evolution, due to the mutations it causes, and the selective and any associated nipple or skin problems are all **important factors** to determine. Any pain should be investigated e.g. whether

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:

patriotic, nationalist fighters for the country. </p><p> An additional **factor** /NN1 affecting women was arguably their education. Debated by reformers seen that a varied combination of changing social and political **factors** /NN2 led to the development of feminism in Japan. Through increased responsibility of the Americans may thus have been the deciding **factor** /NN1 in the eventual communisation of the countries in Eastern Europe proposed that [t]o separate, by scientific abstraction, these two **factors** /NN2 of form and content which are in reality inseparably united

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 (ie, the lemma):

In order to further their argument, the authors **take** /take a great and careless leap beyond their use of the cor
ify their use of ethnicity (which, again, may be **taken** /take as interchangeable in the book)? Only a paragraph s
easier for the personnel to make sure no theft **takes** /take place. Classical music sounds throughout the store; t
or a while (pretending to look for books), and **took** /take 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 **take** /take-V a great and careless leap beyond their use of the con-
 tify their use of ethnicity (which, again, may be **taken** /take-V as interchangeable in the book)? Only a paragraph s
 easier for the personnel to make sure no theft **takes** /take-V place. Classical music sounds throughout the store; n
 or a while (pretending to look for books), and **took** /take-V 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’.

politanism. By the late thirties, fifty such **films** /Q4.3 were produced a year, also serving as a char
 cts of the Brazilian national identity. The **film** /Q4.3 genre of the chanchada came hereby to play a
 nal symbol. The production of Brazilian **films** /Q4.3 brought a major contribution in the transforma

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.

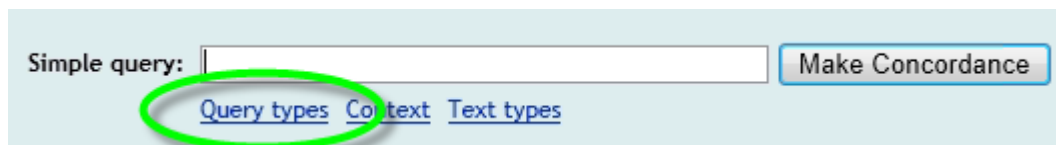
using a look up table that uses empirical **formulae** /running-text and calibrates the required values. </p> EXPE
 id define any parameters you use in your **formula** /heading and in the construction of your model. C is the c
 and benefits of using these tools. Within **Formula** /abstract One CAD and Computational Fluid Dynamics ;
 he rear wing is a crucial component on a **Formula** /abstract One car and its' design can greatly affect the pe
 ersionship must comply to. </p><p> The **Formula** /running-text One World Championship is one such series th
 chnical regulations which are supplied to **Formula** /running-text One teams showing the FIA rules regarding rea
 he rear wing is a crucial component on a **Formula** /running-text One car and its' design can greatly affect the pe
 gs that were used during the 2004-2005 **Formula** /running-text One season. These illustrate clearly the various
 > (as accessed 17/01/2006) </p><p> **Formula** /bibliography 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

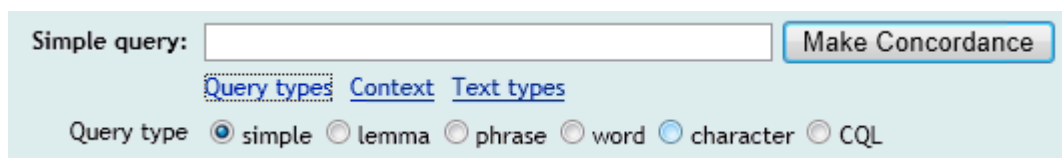
Click on 'Query Type' on the first query page:



Simple query:

[Query types](#) [Context](#) [Text types](#)

You can choose between various types of query:



Simple query:

[Query types](#) [Context](#) [Text types](#)

Query type simple lemma phrase word character CQL

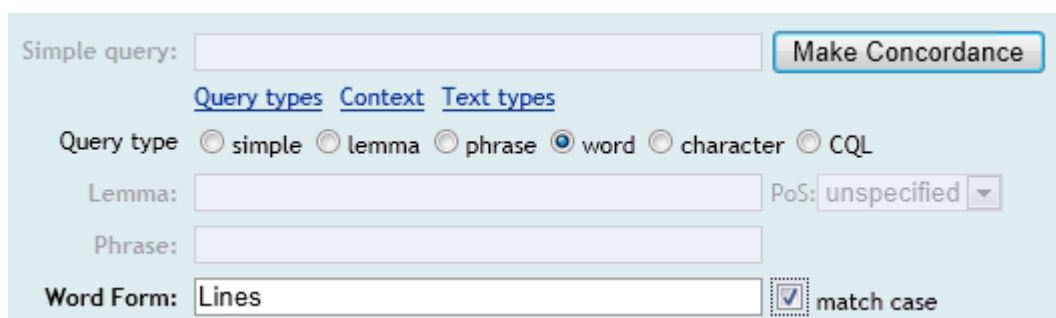
If you choose the 'lemma' option, and type in an uninflected form of a word, all the inflected forms of the word will appear in the concordance lines.

continuation of the ideology then racist actions would not **take** place. *</p><p>* I will now look at why racism still persists in this aspect, and as Hindess points out "they cannot be **taken** as a...reliable account" (Hindess, 1973). The term reliability in the end. Cicourel also supports this idea, however he **takes** it one step further by saying that these decisions are influenced ample the University of Edinburgh, and prevented from **taking** exams that would gain them a place on the Medical Register

The 'phrase' option enables you to search for a sequence of words:

(London, 1984), p. 106. *</p><p>* In this essay **I am going to** argue from a standpoint assuming that a gap between the theory of W. G. Sebald's 'Love's Last Shift'. Firstly however, **I am going to** discuss each of the styles and establish the clear difference between the beginning with an analysis of the unity of mind **I am going to** proceed by attempting to apply the ordinary, unanalyzed conception of the beginning of the 20th century. After that, **I am going to** question these academic concepts most agreed in the past by the need to endure in the future contexts. In this paper, **I am going to** discuss the crucial elements of literacy development and explore the implications of literacy teaching. In the following part of my paper, **I am going to** link up drama with literacy learning on the basis of the conclusions that the outputs are other states of the organism. *</p><p>* **I am going to** examine a functional analysis of pain. I shall assume that

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'.



Simple query:

[Query types](#) [Context](#) [Text types](#)

Query type simple lemma phrase word character CQL

Lemma: PoS:

Phrase:

Word Form: match case

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'

Simple query:

[Query types](#) [Context](#) [Text types](#)

Query type simple lemma phrase word character

Lemma:

Phrase:

Word Form:

Character:

CQL:

Context

Lemma filter

Window: tokens.

Lemma(s): of these items.

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** the rope around his neck, believing that
 position. In addition to *take* the ortho **position** the electrophile would encounter significant
 logical positivists being forced to *take* the **position** that philosophical work surrounding ethics
 before each reading was *taken* the necessary **positions** of the driver and detector were calculated
 organizations, and for managers to *take* a **position** as strategists, as argued by Drucker and
 government, which failed to *take* a decisive **position** on the issue. Le Pen openly criticized
 incompetent, how did Andersen staff move to *take* **positions** in Freddie's Investment Divisions, especially
 into what things are without *taking* any **position** that things are. The psychological reduction
 began to dramatically increase and to *take* a **position** as a prevalent type of inter-state economic
 intensity was being recorded. Care was *taken* to **position** the detector in the place which yielded
 further education (FE) must *take* a prominent **position** in peoples' lives to assist in the process
 company's foundation. He *takes* a prominent **position** at the head of the company making the kind
 technique forces the reader to *take* the **position** of the Anishinabe people, and consequently
 general, Taylor and Francis *takes* a good **position** in this field. *</p> Section Three: Identification*
 both governments will soon *take* a stronger **position** on climate change (The White House, 2007
 programme was finished, I have *taken* up a **position** of an architectural assistant at White
 that: *</p> "Regression involves taking the position* of a child in some problematic situation
 and would therefore have to *take* his class **position** . *</p><p> Overall it could be argued that*
 that Catullus considers *taking* a feminine **position** in man to man relationship to be as serious
 in emotions of the listeners. He *took* the **position** of defending Helen, simply because he wanted
 Broglie - Bohm Theory has not *taken* pole **position** . According to James T. Cushing, historical
 hetaira but later on in her career *took* up the **position** of a pallake. Some may have been free alien
 Four readings shall be *taken* at different **positions** along the beam and averages calculated.
 therefore when the readings were *taken* , the **position** of the IRT was constantly changing and

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).

Query Type: Simple
Query: factor

Text Types

Subcorpus:

text.discgroup	text.grade	text.level	text.sex	text.studentage
<input type="checkbox"/> AH	<input type="checkbox"/> D	<input type="checkbox"/> 1	<input type="checkbox"/> f	<input type="checkbox"/> 25-
<input type="checkbox"/> LS	<input type="checkbox"/> M	<input type="checkbox"/> 2	<input type="checkbox"/> m	<input type="checkbox"/> 26+
<input checked="" type="checkbox"/> PS	<input type="checkbox"/> unknown	<input type="checkbox"/> 3	<input type="button" value="Select All"/>	<input type="checkbox"/> unknown
<input type="checkbox"/> SS	<input type="button" value="Select All"/>	<input checked="" type="checkbox"/> 4		<input type="button" value="Select All"/>
<input type="button" value="Select All"/>		<input type="checkbox"/> unknown		
		<input type="button" value="Select All"/>		

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.

text.discipline

text.genre
Methodology recount

text.l1
sp

type in the first few letters

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.

Subcorpus: English info

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

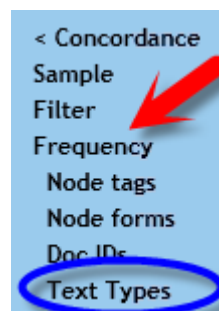
Lesson 5 Examining Frequency

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.

Corpus: **British Academic Written English Corpus (BAWE)**
 Hits: 13

LS 4,1868 J/cal = -18,34 J K⁻¹ mol⁻¹. </p><p> So, . </p> 7. **Entropy** determination <p> According to the previous steps,
 LS approximately, or 340,15 K in absolute temperature. </p><p> The **entropy** change of the transition can be determined by using
 LS the pressure is held constant. </p><p> The change of **entropy** is -1183,2 J K⁻¹ mol⁻¹. This is a measure of the
 PS are for Gibbs Free Energy (which also includes an **entropy** term). However they provide a qualitative comparison
 PS nitrogen and carbon dioxide was used to calculate the **entropy** of vapourisation of both these substances. </p><p>
 PS for expt. 3 (Carbon Dioxide) <p/> Conclusions <p> The **entropy** of vapourisation of carbon dioxide was found to be
 PS are in agreement with Trouton's rule that states the **entropy** of vapourisation of most substances is approximately
 PS radiation that strikes it's surface) </p><p> That is; the **entropy** or the disorder of a closed system must always increase
 PS it enables us to work out things such as how much **entropy** there is within black holes. It has also brought
 PS thermodynamics of black holes, specifically black hole **entropy** . </p><p> It is theorised that at extremely high energies
 PS it provides only a very limited explanation of the **entropy** of black holes; there are currently five different
 PS the same idea of a system that lies between a state **entropy** and order. This system can subsequently be used to
 PS PRNG. The use of PRNG and ASCII mapping reduces the **entropy** and therefore instead of 2⁴⁰ we have got only 2



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.

<u>text.discipline</u>	<u>Freq</u>	<u>Rel</u> [%]
p/n Physics	5	233.6
p/n Chemistry	4	142.8
p/n Biological Sciences	3	56.4
p/n Computer Science	1	36.5

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.

<u>text.discipline</u>	<u>Freq</u>	<u>Rel [%]</u>	
p/n Law	94	805.1	
p/n Business	19	149.4	
p/n Engineering	10	48.2	
p/n Philosophy	6	65.0	
p/n History	4	47.8	
p/n Psychology	2	24.2	
p/n English	2	21.7	
p/n Computer Science	2	26.4	
p/n Comparative American Studies	2	31.0	
p/n Agriculture	2	17.1	
p/n other	1	19.1	
p/n Politics	1	10.4	
p/n Physics	1	16.9	
p/n Hospitality, Leisure & Tourism Management	1	12.3	
p/n Health	1	14.2	
p/n Classics	1	14.0	

<u>text.genre</u>	<u>Freq</u>	<u>Rel [%]</u>	
p/n Problem question	69	3005.1	
p/n Essay	48	64.0	
p/n Critique	22	114.7	
p/n Explanation	4	34.8	
p/n Case study	2	17.1	
p/n Research report	1	26.4	
p/n Narrative recount	1	26.4	
p/n Methodology recount	1	5.0	
p/n Empathy writing	1	57.6	

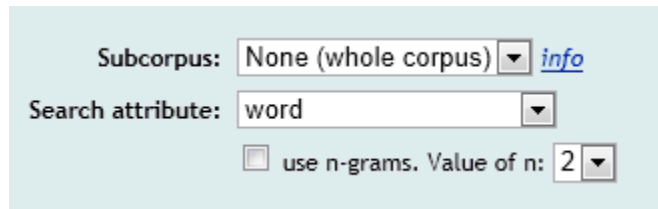
We can get more frequency information by doing a concordance search for ‘liable’, and then getting collocation information in the range 0 to 1 (see Lesson 2, section 2.1).

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

This shows us that ‘for’ and ‘to’ are the most frequent right collocates of ‘liable’.

5.2 Word lists

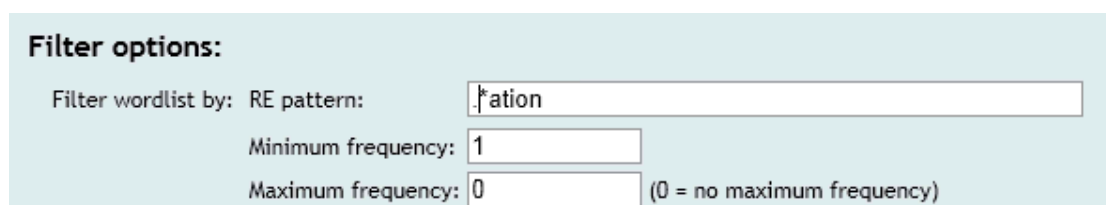
We can create a word list by using the ‘Word list’ tool. Click on the ‘Word list’ link in the main menu. You can choose to create frequency lists by ‘word’, ‘tag’, ‘lemma’, ‘lempos’, ‘semantic tag’ or ‘textpart’ (for explanations of these, see 3.2).



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.

<u>word</u>	<u>Freq</u>
the	429343
,	391643
.	313580
of	270136
and	207623
to	188666
in	137911
a	125736
is	110721
)	91843
(90538
that	78781
'	72584
The	62638
as	62128
be	58053
for	54827

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’.

<u>word</u>	<u>Freq</u>
information	1466.9
situation	792.9
population	650.3
relation	505.2
communication	382.5
application	367.5
education	358.6
consideration	341.2
explanation	324.5
interpretation	313.8
organisation	313.6
combination	292.7
formation	278.6
location	267.0
investigation	253.8
creation	251.1
representation	241.0

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.

Subcorpus: [info](#)
 Search attribute:
 use n-grams. Value of n:

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

Word Sketch Entry Form [?](#) 

Lemma:
 Part of speech:
[Advanced options](#)

Click on the question mark for more information.

The results of this query are shown below.

analysis (noun) British Academic Written English Corpus (BAWE) freq = **4608** (552.8 per million) Click on collocates in boldface to get multi word sketches.

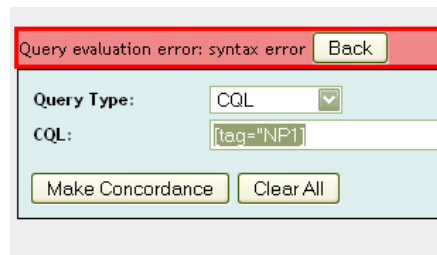
modifier	2268	1.9	pp_of-I	804	2.9	object_of	659	1.5	and/or	574	0.9	subject_of	350	1.6
detailed	62	9.52	variance	15	8.73	perform	53	9.42	interpretation	16	7.69	reveal	12	7.62
statistical	55	9.44	data	34	7.1	conduct	26	8.64	visualisation	4	7.66	confirm	5	6.93
swot	49	9.43	text	11	6.89	undertake	15	8.34	testing	7	7.44	show	45	6.52
fe-	37	8.98	shaft	4	6.69	carry	42	8.1	nmr-	4	7.31	enable	7	6.36
regression	37	8.64	category	5	6.19	complete	12	8.04	summary	4	7.07	suggest	15	6.03
comparative	30	8.57	plate	4	6.11	run	8	6.89	collection	5	6.9	assume	5	6.0
discourse	38	8.53	sample	10	6.01	base	26	6.77	conclusion	9	6.89	demonstrate	5	5.86
critical	34	8.31	nature	11	5.87	undergo	4	6.72	calculation	6	6.82	prove	4	5.63
finite	22	8.18	result	18	5.86	incorporate	5	6.68	design	13	6.71	focus	4	5.58
far	36	8.17	ratio	5	5.63	provide	35	6.55	discussion	6	6.59	look	5	5.58
genetic	25	8.11	section	6	5.62	apply	10	6.4	description	6	6.54	follow	7	5.55
isotope	20	8.07	mechanism	4	5.6	lack	4	6.36	assessment	5	6.5	indicate	5	5.32
in-depth	18	7.97	demand	6	5.58	employ	5	6.28	comparison	4	5.94	help	4	5.27
peste	17	7.92	evidence	8	5.4	present	9	6.26	result	16	5.72	seem	7	5.27
network	30	7.87	reason	6	5.35	combine	4	6.21	review	4	5.64	require	8	5.13
thorough	17	7.85	organisation	5	5.34	require	16	6.07	report	5	5.51	identify	4	4.98
above	18	7.71	structure	10	5.33	focus	6	6.01	test	6	5.51	come	4	4.94
financial	25	7.58	design	5	5.26	enable	6	5.99	action	4	4.72	determine	4	4.88
empirical	16	7.55	situation	5	5.14	propose	4	5.94	data	6	4.64	provide	8	4.46
cost-benefit	13	7.54	issue	7	5.12	use	51	5.91	fact	4	4.53	allow	5	4.41
swot-	13	7.54	relation	5	5.1	limit	4	5.67	method	6	4.47	use	17	4.34
data	54	7.53	pattern	4	5.09	involve	7	5.52	evidence	4	4.45	find	7	4.31
dimensional	12	7.37	performance	5	5.06	offer	5	5.44	approach	4	4.29	take	7	3.99
historical	16	7.32	project	5	4.92	include	7	5.28	study	4	3.98	give	5	3.56
ratio	22	7.24	case	7	4.92	allow	9	5.2	factor	4	3.85	make	4	2.63

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

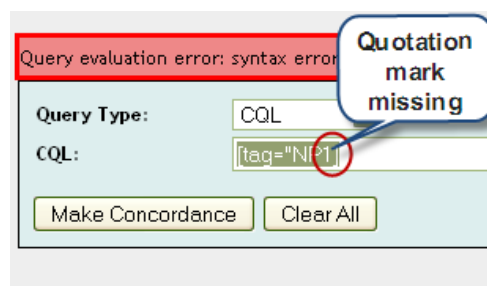
CQL has its own syntax and you need to make sure that you get the form of a CQL query exactly right. If you make a mistake with this, you will get an error message like the following:



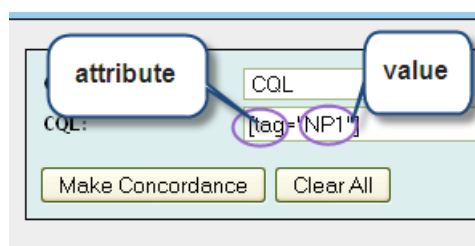
In a simple CQL query for a single word, the square brackets enclose the query, and the first part of the query identifies what category you are looking for (the technical term is ‘attribute’), followed by an equals sign, and then a code of some sort (the technical term is ‘value’) and this has to appear 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 **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.

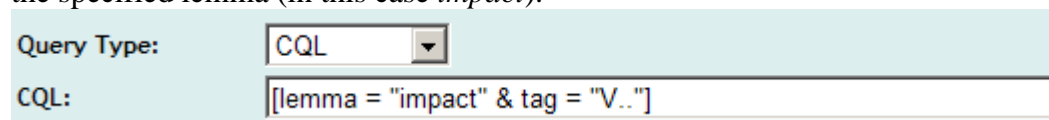
If you want to specify two or more alternatives for a given slot in the value, you can put the choices inside normal brackets and divide with a pipe character. For example [tag = "VB(D|N)"] captures all instances of the past participle and the past tense of the verb BE.

Query Type:	<input type="text" value="CQL"/>
CQL:	<input type="text" value="[tag = \" vb(d n)\"]"=""/>

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 lemos attributes: Here are some examples:

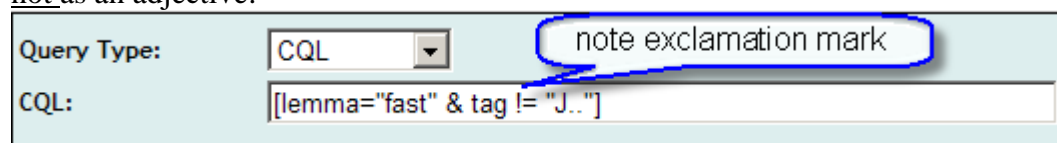
A search such as [lemma = "impact" & tag = "V.."] will find only the verb forms of the specified lemma (in this case *impact*).



Query Type: CQL
CQL: [lemma = "impact" & tag = "V.."]

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:



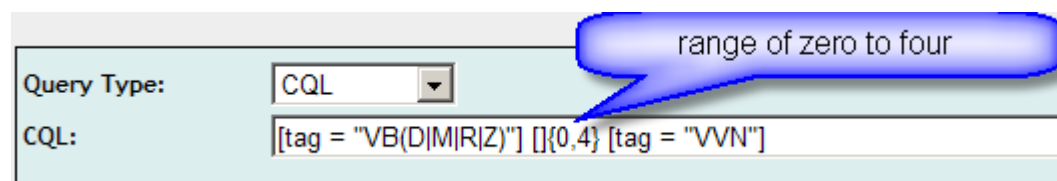
Query Type: CQL
CQL: [lemma="fast" & tag != "J.."]

note exclamation mark

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.

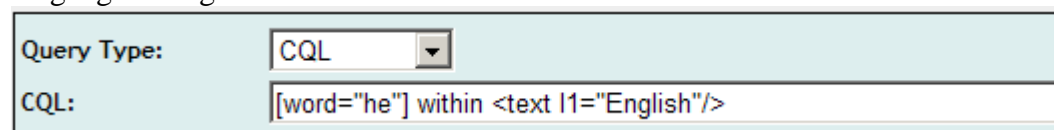


Query Type: CQL
CQL: [tag = "VB(D|M|R|Z)"] []{0,4} [tag = "VVN"]

range of zero to four

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.



Query Type: CQL
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:

Query Type:	CQL
CQL:	[lemma="think"] within <quote lang="\w+" />
Make Concordance	Clear All

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.

Query Type:	CQL
CQL:	[lemma="government" & textpart="running-text"]
Make Concordance	Clear All

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.

Query Type:	CQL	note exclamation mark
CQL:	[word="he"] within <text l1!="English"/>	
Make Concordance	Clear All	

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