# Getting Started with the Sketch Engine

## 1. Background

The Sketch Engine is a web-based program which takes as its input a corpus of any language with an appropriate level of linguistic mark-up. The Sketch Engine has a number of language-analysis functions, the core ones being:

- **Concordances**
  
  Concordance is a view of all occurrences from the corpus for a given query including the textual context. The program is very powerful with a wide variety of query types and many different ways of displaying and organising the results.

- **Word Sketches**
  
  Word sketches are corpus-based summaries of a word's grammatical and collocational behaviour. They are described below in section 5.

How to read this guide:

- For the purposes of this guide, we use examples based on the Sketch Engine loaded with a sample corpus of English, the [British National Corpus (BNC)](http://www.sketchengine.co.uk/).
- For more information about the research around Sketch Engine, see the [publication list](http://www.sketchengine.co.uk/).
- Another good place to learn how to use the Sketch Engine in depth is [Lexicom](http://www.sketchengine.co.uk/): the annual, 5 day, workshop in Lexicography and Lexical Computing.
- Most terminology is defined as it is encountered below, however for a full glossary please see our [Jargon Buster](http://www.sketchengine.co.uk/).

## 2. Home page

The software is on the Sketch Engine website: [http://www.sketchengine.co.uk/](http://www.sketchengine.co.uk/). In what follows, we have added links to this website. To view these links you will need to login to the Sketch Engine website: [http://www.sketchengine.co.uk/](http://www.sketchengine.co.uk/). After following a link in this tutorial, you can click the back icon in your browser to get back to this tutorial (alternatively, if you **right click** you can open the link in a different window or tab). You can follow the instructions below in a separate window so that you can compare what you see in your working screen with the links and descriptions given in this tutorial.

Also, please note that if you are using a customer specific installation of Sketch Engine, rather than the [http://www.sketchengine.co.uk/](http://www.sketchengine.co.uk/) website, the appearance of your screen may be slightly different, for example with regard to the colour, logos or text formatting.

If you are not a registered user yet, we recommend that you set up a free Sketch Engine trial account before reading on, so that you can look at the examples on the [BNC](http://www.sketchengine.co.uk/) referenced below. Where possible we also provide alternative links to the same examples on the open [ACL Anthology Reference Corpus](http://www.sketchengine.co.uk/), which you can open without logging in. Note though that some of the text below relates specifically to the results on BNC and you will see different data and different numbers on ACL ARC.
Follow the links from http://www.sketchengine.co.uk/ page to either set up an account, or log in. The "home" screen looks like this: ?click here.

Wherever you are in Sketch Engine, the link back to this home page is always displayed at the top right hand corner. Likewise you can always see "Settings", which allows you to update personal information and your password, and the "Log out" link.

On the left hand side, you see options for creating corpora and a few other tools.

In the main panel you can select your corpus. Here we want to explore the British National Corpus, so we click on that.

If you prefer to work with an open corpus, you can go to the ?list of open corpora and click on the ACL Anthology Reference Corpus.

### 3. Generating a concordance

Your screen should then look like the link below:

?click here (BNC) or ?click here (ACL ARC; no login required)

In the left hand side panel the option:

- **Concordance** will always bring you back to this screen

while:

- **Word List**
- **Word Sketch**
- **Thesaurus**
- **Find X**
- **Sketch-Diff**

take you to other tools which will be described in the sections below.

To generate a concordance, you enter the main search term in the (simple) query box in the main panel of the screen.

If, like the BNC, the corpus is lemmatized, the terms will match the lemma (the stemmed form) as well as the word. If you enter `save`, the Sketch Engine will generate a concordance of all of the following:

i) `save`- `saved`- `saves`- `saving` (verb)

ii) `save`- `saves` (noun - what goalkeepers make)

iii) `save` (preposition: *everyone was killed save Franco himself*)

You can also enter phrases in the query box.

To make more specific searches, you can select from the dropdown "Query Type" menu. This allows you to make specific types of queries:
simple: is the standard query which will match the lemma as well as the word as described above

- lemma: will again match any lemma but here you can specify the part of speech (PoS i.e. the grammatical class e.g. noun, verb, adjective etc...). This option will not work for phrases. (Here and below we assume the corpus is, like the BNC, lemmatized and part-of-speech tagged. If it is not, not all of these query type options are available.)
- phrase: will match a phrase e.g. runs away, and any capitalised variant e.g. Runs away, but will not match the lemma, so in this example run away will not be found.
- Word form will match any word form exactly, you can select the PoS (e.g. noun or verb). You can also select whether you wish the system to match the exact capitisation you entered using "match case". For example, this will enable you to search for Bush rather than bush.
- character matches a character string. For example, ate will match words containing this character sequence. This might be particularly useful in languages where tokenisation is difficult.
- CQL: is for inputting complex queries using Corpus Query Language, described in Corpus Querying and Grammar Writing. "Default attribute" controls how CQL queries will be understood. The "tagset summary" box gives details of the part-of-speech tags used in the tagging.

If you do not want to specify context any more precisely, you are now ready to hit the "Make Concordance" button and see the concordance. You will find more information about manipulating the output in Section 4 below. Note that when you have obtained the concordance you can always get back to the query entry form described here by clicking on “Concordance” at the top of the left hand side panel. The next sections explain how to limit your search to a specific context or text type.

For the purposes of reading the following context and text type sections, make sure you are at the concordance entry form (by clicking concordance at the top of the left hand side menu) select "lemma" as the query type in the concordance entry form. For future reference note that all the options from this section are available with all the options described in the following sections on context and text type.

**The Context section**

Now open the Context section by clicking on the "Context" expert option in the left hand side panel.

With the Context option you can make various specifications on the lemmas and/or PoS in the words surrounding your query. For both the lemma and PoS constraints you can indicate whether the system should look for the lemmas (or PoS) to the left or right or at either side (both) of your query term. You also get a chance to specify how many tokens (words or punctuation), up to 15, of context to search for these constraints. You enter any number of lemmas or PoS and can specify if they should "all" apply, or whether "any" or "none" should be matched.

Here are some examples:

1. you want to search for the lemma shake (verb) followed by head (noun), to find instances such as she shook her head, if you agree shake your head, and shaking their heads in disbelief... You can do the following:
   - either type shake in the query box with PoS verb. Then type head in the Context lemma box PoS noun and specify Right and a window size (say 3 tokens)
   - or type head in the query box with PoS noun. Then type shake in the Context lemma box with PoS verb and specify Left and a window size (say 3 tokens)

   The results will be the same whichever route you take.

2. you want to search for the verb taste followed by any adjective; since a following adjective may appear either in position 1 (it tastes horrible), position 2 (it tastes really delicious), or even position 3 (it didn’t taste good at all) you can do the following:
   - either type taste in the query box with PoS verb. Then type any adjective in the Context lemma box PoS adjective and specify Right and a window size (say 3 tokens)
   - or type any adjective in the query box with PoS adjective. Then type taste in the Context lemma box with PoS verb and specify Left and a window size (say 3 tokens)

   The results will be the same whichever route you take.

3. Generating a concordance
taste quite so good). Type taste in the query box, with query type lemma and PoS "verb". Then - in the Context area - select "adjective" from the PoS list and specify Right and a window size of 3 tokens. This generates a concordance of 480 lines in the BNC. You can further refine your search by specifying two PoS in the Context section. In this case, if you select both "adjective" and "adverb" by holding the CTRL key to select more than one PoS you will get a smaller concordance of 125 lines, with examples such as it tastes bloody awful and it tastes surprisingly good.

You can clear any boxes with the "clear all" option at the bottom of the screen.

There are many more complex searches you can carry out using this feature - it is worth trying things out to see what is possible. For example, you could further refine the first search here (with head=Lemma and shake=Left Context) by also specifying a PoS in the Right Context. Thus specifying "adverb" in the Right Context will generate lines such as shook his head disapprovingly, whereas specifying "noun" will generate shook their heads in agreement. There are very many searches one might try, though in practice most searches are relatively simple.

Context searches can also be used to exclude unwanted items: thus you could input a query of weapons of using the phrase option for the Query type (described in the section above), then exclude "destruction" by typing it into the Context Lemma box, specifying Right and then selecting "None" from the drop-down list. This returns a concordance for any lines containing the string weapons of without the word destruction.

**The Text Type section**

Return to the concordance query form, if you are not already there, by clicking on "Concordance" at the top of the left hand side menu. Close the Context section by clicking on the expert option "Context" and select the option "Text Type", again, in the left hand side panel.

With the Text Types option you can limit your search to a part of the corpus. If you want to see how a word behaves in the spoken part of the corpus, enter the word in the search box (or combine with other search specifications as described above) and tick the boxes for "Spoken context governed" and "Spoken demographic". Your concordance will contain only spoken-language examples. The partitions available depend on the text types (also referred to as header information or metadata) provided in the corpus data.

4. **Manipulating your concordance output**

Once you have generated a concordance, there are several options for increasing its usefulness. Click on "Concordance", chose a query type simple search and enter the word haunt and click "Make Concordance".

The concordance screen looks like this:

?click here (BNC) or ?click here (ACL ARC; no login required)

As before, the options above the bar in the left hand side will take you to other parts of the program and are described below. The options below the horizontal bar in the left hand side menu allow you to work on this concordance.

The panel directly above the concordance tells you which corpus you are using, and how many hits match your search item. For haunt, there are 1098 concordance lines.
Moving around the concordance

You can move from one part of the concordance to another either by specifying a number in the Page box and selecting Go, or by clicking on Next, Last, First or Previous.

Finding out about a particular concordance line

If you click on one of the highlighted node words, more of its context appears in the panel at the bottom of the screen and you can further expand the context by clicking on expand left and/or expand right. To hide this extra context click on the "-" in the top left hand of the context window.

To get information about the source-text a particular concordance line comes from, click the document-id code at the left-hand end of the relevant line (assuming you have not changed the "View option" relating to "references", see below). This brings up "header" information in the bottom pane.

The concordance menu

In the lower section of the left hand side panel there are various options for refining your concordance.

- **View Options**: takes you to a new screen in the main panel that allows you to change the concordance view in various ways. To summarise the functions available when you select View Options (NB if you do click view options then you can select view concordance to get back):
  - the Attributes column allows you to change from the default display (in which only the text is visible in the concordance line) to a number of alternative views in which you can see PoS-tags, lemmatized forms, and any other fields of information, either for the node word only ("KWIC tokens only") or for every word in the concordance line ("For each token"). The function can be useful for finding out why an unexpected corpus line has matched a query, as the cause is sometimes an incorrect PoS-tag or lemmatization
  - the Structures column allows you to change from the default display to show the beginning and end tags for structures such as sentences, paragraphs and documents.
  - the References column dictates the type of information regarding the source texts which appears (in blue) at the left-hand end of the concordance line. The default is an identifier for the document that the concordance line is taken from. Any other fields of information about corpus documents can be selected and the value that the concordance line has for that field will then be seen. For example, if the corpus encodes whether a document is imaginative writing or not, and the appropriate feature (e.g. in the BNC this is "Domain for written corpus texts") is selected in the References column and change view options is clicked, then the domain of the concordance lines will be displayed in the left hand column and we can see those that come from an "imaginative" text.
  - the Page Size box (bottom left) allows you to specify a longer page length for the display: the default is that each page of concordances contains 20 lines. (Increasing the Page Size will slow down initial retrieval of the concordance.)
  - KWIC Context size allows you to specify the size of the context window in number of characters
  - Sort good dictionary examples allows you to specify how many lines of ‘good’ examples that the system should automatically rank at the top of the concordance according to the GDEX program (see http://www.kilgarriff.co.uk/Publications/2008-KilgEtAl-euralex-gdex.doc)
  - Icon for one-click sentence copying: You can add an icon for copying lines from the concordance
  - Allow multiple lines selection - Allow user to select and/or copy more than one line at once.
  - XML template for one-click copying (A feature used for specific projects only)
• **KWIC/Sentence** lets you toggle between standard KWIC concordance view (which appears by default) and full sentence view.

• **Save** gives you options for sorting the concordance. You can specify whether the output is text or xml, how many pages, whether a heading is included, whether the lines are numbered, whether the KWIC are aligned in the output and the maximum number of lines.

• **Sort**: Sorting is often a quick way of revealing patterns. If you select this option in the left hand side panel you obtain a screen in the main panel with various complex options for sorting (see the page specific help on Sort) you can alternatively use the other options below sort to simply sort by:
  - **Left**: one token (word or punctuation) to the left
  - **Right**: one token to the right
  - **Node**: the KWIC (also referred to as the node word)
  - **References**: sorting according to whichever references you display to the left of the concordance lines (as described in view options above).
  - **Shuffle**: this shuffles the concordance so that the lines are arbitrarily ordered. Since the sample option described below always provides the same ordering for a give sized sample, this allows you to jumble the concordance so you can view only a portion of the concordance or your sample, without bias from the ordering.

• **Sample**: This allows you to create a random sample of the corpus lines. You can specify the size of the sample (i.e. the number of lines) or use the default of 250. For example, if you search for *play* (verb) and decide that you do not want to analyse 37,632 lines, use this option to reduce this to a manageable number. (see also specific help on the random sample page)

• **Filter**: This allows you to specify constraints on the context of your KWIC to retrieve a subset of your concordance. See the filter page specific help

• **Frequency** allows you to produce two types of frequency information regarding your search term:
  1. **Multilevel frequency distribution** shows the frequency of each form of a given lemma. To see how this works, make a concordance for *forge* (verb): when the concordance displays, select Frequency and use the Multilevel frequency distribution section. The (default) "first level" shows you the frequencies of the forms *forge*, *forged*, *forging* and *forges*. The second and third levels allow more complex searches of this type: for example if you check "second level" and select "1R" (=word one position to right of node word) you will see which words appear in this position and how frequent each of these words is.
  2. **Text type frequency distribution** shows how your search term is distributed through the texts in the corpus. You may find, for example, that a word like *police* appears significantly more often in newspaper texts than in other text types. This is a potentially useful tool which could show you - for example - that a particular medical term is not restricted to specialised medical discourse. As with the "references" column in the "View Options" screen, the actual values you can select depend on the corpus you are using, and how it has been set up in the Sketch Engine.

• The frequency option is also described in the page specific help on frequency. You can alternatively use the simpler frequency options below **Frequency** to simply sort by:
  - **Node tags**: the PoS tags for all the KWIC word forms (node word types)
  - **Node forms**: the word forms for all the KWIC word forms
  - **Doc IDs**: frequency distribution over the document ids
  - **Text Types**: frequency distribution over all the text types specified for the corpus
• **Collocations** allows you to generate lists of words that co-occur frequently with your node word (its "collocates"). Where word sketches (see the next section) are available, they give a more sophisticated account of collocates in most cases. (see collocations page specific help)

• **Original Concordance**: is visible if you have refined your concordance. If you select this you can get rid of the refinements and return to the original concordance.

• **ConcDesc**: provides a technical description of your query. This is useful for programmers and technical people.

5. The Word Sketch function

A Word Sketch is a corpus-based summary of a word's grammatical and collocational behaviour.

Click on **Word Sketch** in the left hand side main menu (top section of the left hand side menu), and this takes you to the Word Sketch entry form, which looks like this:

?click here (BNC) or ?click here (ACL ARC; no login required)

Choose a lemma and specify its part of speech using the drop-down list. Word Sketches are typically available for nouns, verbs, and adjectives and can be available for other word classes depending on the grammatical definitions supplied to the sketch engine (see the documentation on grammatical relation definitions for more information). Word sketches also depend on the availability of substantial amounts of data, so if you try to create a Word Sketch for a fairly rare item you will see a message saying there is no Word Sketch available. (This is perfectly reasonable: the point of the Word Sketches is to provide helpful summaries when there is too much corpus data to scan efficiently using a concordance; but when there are only a few concordance lines it is easy enough to analyse them all manually.) In general, you need several hundred instances of a word to make a useful word sketch.

This ?link shows a Word Sketch for the noun *challenge*. (?Alternative link for ACL ARC.)

Each column show the words that typically combine with *challenge* in a particular grammatical relations (or "gramrels"). Most of these gramrels are self-explanatory. For example, "object_of" lists - in order of statistical significance rather than raw frequency - the verbs that most typically occupy the verb slot in cases where *challenge* is the object of a verb. Most of the data is lexicographically relevant, though one might query the adjectival modifier *larval*: it turns out that *larval challenge* is a technical term used in parasitology, discussed in a BNC document.

You can at any time switch between Concordance mode and Word Sketch mode, and this is a useful way of getting more information about a particular word combination. Thus, if you want to look at examples of "*pose + challenge*" (where *challenge* is the direct object of *pose*), simply click on the number next to *pose* in the "object_of" list (92) and you will be taken directly to a concordance showing all instances of this combination.

If you want to create a MultiWord Sketch, look into **How to's...**

6. The Thesaurus function

The software checks to see which words occur with the same collocates as other words, and on the basis of this data it generates a "distributional thesaurus". A distributional thesaurus is an automatically produced "thesaurus" which finds words that tend to occur in similar contexts as the target word. It is not a man made thesaurus of synonyms.
The thesaurus function lists, for any given adjective, noun or verb, the other words most similar to it in terms of grammatical and collocational behaviour.

Click on the Thesaurus link on the left hand side main (top) menu and then input the word with PoS that you are interested in.

For help on the advanced options see the thesaurus help page.

**7. The Sketch Difference function**

Sketch Difference is a neat way of comparing two very similar words: it shows those patterns and combinations that the two items have in common, and also those patterns and combinations that are more typical of, or unique to, one word rather than the other. You can also use the function to compare the same lemma in two different parts of the corpus, or to compare two different word forms e.g. men and man. Click on any word in a Thesaurus entry for a word, and you will be taken straight to a screen showing the Sketch Difference between the two words. Alternatively, you can click on Sketch-Diff on the left hand side panel and this will take you to the word sketch difference entry form which gives you more options.

Suppose you want to compare clever and intelligent. In the thesaurus entry for clever, intelligent comes top of the list: it is statistically the most similar word in terms of shared contexts of occurrence. Click on intelligent and you are taken to a new screen which is in three main parts: the first part shows "Common Patterns" (those combinations where clever and intelligent behave quite similarly), the second and third parts show "clever only patterns" and "intelligent only patterns". The screen looks like this ?click here. (?Alternative link for ACL ARC.)

In the "Common Patterns" part, there are four numbers next to each collocate. The first two indicate the frequency of co-occurrence with the first and second lemma, the last two show the salience scores for the collocate with both lemmas. All collocates are sorted according to maximum of the two salience scores and coloured according to difference between the scores.

Try this out, and look at the difference in the "and/or" lists: people can be "intelligent and articulate/thoughtful/sensitive" etc, but they are often "clever and devious/cunning/brave".

For more information on the other options see the Word Sketch Difference help

**8. The Search function**

From any screen you can do a "simple" Search in any corpus by using the field and drop down list in the horizontal panel which appears just beneath the very top bar in which you can search the Help documentation. This search function provides a short cut to a simple concordance

**9. Other functions**

For an explanation of other functions in the left hand side margin you can click the help links marked with a ?

Click here for the Start Page for Sketch Engine Documentation.