Sketch Engine features

- **Concordancer**: show words in context
- **Word sketches**: learn how words are used
- **Word sketch difference**: compare and contrast words visually
- **Printer icon**: print queries
- **Ske.li**: get a short permanent link
- **Corpus Architect**: create and use your own corpora
- **WebBootCat**: create specialised corpora from Web instantly
- **Wordlists**
- **Keywords**
- **Terms**
- **N-grams**: search n-grams in a corpus
- **Distributional thesaurus**
- **Parallel corpora**: search many parallel corpora
- **Compare corpora**
- **Corpus Info**: show information about corpus
- **Line numbers**: show line numbers in concordance

**Concordancer**

Look up words, phrases or sentences in context matching complex criteria according to our Corpus Query Language (CQL). The resulting concordance may be subject to further standard post-processing: filtering, sorting, sampling, shuffling or retrieving multi level frequency distribution.

**Word sketches**

Learn how words are used via word sketches: a word sketch is a one-page, automatic, corpus-derived summary of a word’s grammatical and collocational behaviour. In the figure, you can see the word sketch for ‘goal’ (noun). Its collocates are grouped according to grammatical relations in which they occur.

**Word sketch difference**

Compare and contrast words visually: word sketch difference allows you to contrast words according to their salient collocational context.

In the figure, the greener a word is, the more closely it relates to clever. The redder a word is, the more closely it relates to intelligent. E.g. it is more usual to say clever trick than intelligent trick and similarly, it is more fluent to say intelligent robot than clever robot.

**Corpus Architect**

With Corpus Architect, you can build your own corpora from documents in various format: TXT, PDF, PS, DOC, HTML, VERT. When processed, you can search and query them within Sketch Engine.
**WebBootCat**

WebBootCat lets you create a specialised corpus from the web in just one or two minutes. You input around six 'seed terms' from the domain you are investigating, and the tool gathers pages that a search engine finds for those terms. You can also input a list of the URLs (web addresses) for web pages that you want to build a corpus from.

**Wordlists**

In the figure you can see several wordlists extracted from various corpora.

- A: the most frequent morphological tags used in the British National Corpus
- B: the most salient collocates for English verb 'take'
- C: the most frequent words in the British National Corpus (BNC)
- D: the most frequent character trigrams from BNC
- E: the most frequent words starting with 'b'
- F: frequency figures of wordforms of English verb 'be'

**Keywords**

The keywords of a corpus are those words that are most frequent in the corpus, in comparison with their frequency in a reference corpus. They let us compare corpora. They also let us discover the distinctive vocabulary of a domain.

In the figure, you can see keywords extracted from corpus Environment using reference corpus enTenTen08.

**Terms**

Terms are multi-word keywords. The following figure shows the top multi-word terms in the Environment corpus.

**Distributional thesaurus**

Distributional thesaurus entries can be extracted for nouns, adjectives and verbs (for English: situation may differ for different languages).

**Parallel corpora**

Parallel corpora are multi-language corpora where text, documents, paragraphs and sentences from one language and their translations in another language are connected. This allows you to study how words and phrases are used and translated accross various languages.

In the Sketch Engine you are not limited to pairs of languages but if you have proper data, you may work with triplets, quadruplets, ... of languages. In the image below there is parallel query for English key, French clé and
**Compare corpora**

Do you want to know how similar or different your own corpus is to other corpora of the same language? In the picture, there is a comparison of various English corpora. The scores in the table stand for **corpus similarity**. 1 is for identical corpora and the bigger the score (and the darker the grey), the greater the difference between two corpora. You can see that the two enTenTen corpora and the New Model Corpus are all very similar to each other, and that the 'most different' pair are 'London English' (of informal conversations in two working-class areas of London) and BAWE (of academic written English).

**Corpus Info**

If you want to know more about the corpus which you work with at the moment, you can use this function. By clicking to **Corpus Info** link in the left menu or to click to name of corpus in navigation bar. There you find out some useful and interesting information, such as statistics of various syntactic and morphological units, main information about language of corpus, encoding, date of the last changes, description of annotation, statistics of attributes etc.

For example you can see a difference between total amount of word forms and lemmas used in the corpus, which shows you how much is the language inflectional.

**Printer icon**

Use very simple way to print the results of your searching.

- Select a corpus and make a concordance.
- Click on the icon placed in the right navigation bar (showed in the picture below).

- Print preview looks like below.

**Ske.li**

**Ske.li** allows you to get a short permanent link. It is useful when you need later to return to a concrete concordance you made and you want to get sure it will not change. To get the short permanent link:

- Select a corpus and make a concordance.
- Click on the icon placed in the right navigation bar (below the printer icon).

- You can give a name (at least 8 characters) to your concordance (but this step is not necessary).
Click on Get permalink.

You can save displayed permanent link and use it whenever you need.

In this way you use the Sketch Engine’s URL shortening tool. For example when you search *anticlockwise*, you get URL containing of 360 characters:

 Nevertheless, in way described above you can get the permalink reduced to a few characters only, e. g. ?http://ske.li/anticlockwise.

**Line numbers**

You can use this function for better clarity and orientation in the text. Each concordance line will have displayed its number, for instance so as you can say ?in line seventeen you can see a rarity of ?? and nobody had to count the lines one by one.

How to visualize the line numbers:

- In the left menu click to View options.
- At the bottom of the menu select Show line numbers.
- Click to Save and change options.

**N-grams**

An n-gram is a contiguous sequence of n items (words, letter, etc.). The specially case of n-gram is a unigram (a word), a bigram (a sequence two words) and a trigram (a sequence three words), of course four-, five- until n-gram exist. There is not expected any relation among units in n-gram. For instance, the n-gram (the trigram) is "a lorry driver" but also "because the green". The result is that not each n-gram is a collocation, however, on the other hand, each collocation is a n-gram.

Firstly, you have to choose a corpus and then click on "Word List" in the left menu. Here you can choose an attribute (Search attribute), which it will search. The important thing is to tick off "use n-grams" and set the value of n (automatic is 2, maximum is 6). Clicking by the button "Make Word List" below shows you n-grams according to the selected option.

Creating n-grams can take several tens of seconds (especially 5- or 6-grams).