Glossary of Sketch Engine Terminology (Jargon Buster)

- **attribute**: (also called positional attribute) a feature ascribed to the tokens (words or punctuation) in the corpus. Examples of attributes are word (i.e. the word form), tag (part-of-speech tag or PoS), lemma, lempos. Each attribute specified for the corpus will be associated with value for each token. For example, the token *dogs* will have attribute values as follows:

<table>
<thead>
<tr>
<th>word</th>
<th>lemma</th>
<th>tag</th>
<th>lempos</th>
</tr>
</thead>
<tbody>
<tr>
<td>dogs</td>
<td>dog</td>
<td>n</td>
<td>dog-n</td>
</tr>
</tbody>
</table>

- **ARF**: Average Reduced Frequency, a variant on a frequency list that 'discounts' multiple occurrences of a word that occur close to each other, e.g. in the same document. For more details see the ARF definition.

- **cluster**: a grouping of words in the thesaurus or word sketch which are clustered by virtue of their shared collocational behaviour (see Clustering Neighbours documentation)

- **collocation**: a sequence of words that co-occur more often than would be expected by chance.

- **concordance**: all occurrences from the corpus for a given query

- **concordancer**: a program which displays the concordance (see above)

- **corpus**: a large set of texts for studying language as it is used in real life

- **CQL**: Corpus Query Language, a formal language for allowing complex queries (for further details see SkE/CorpusQuerying)

- **default attribute**: the attribute (e.g. word or part of speech tag) that is assumed by default in a CQL expression

- **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.

- **filter**: a function which allows you to provide some further criteria to narrow the search and reduce the size of the concordance

- **freq/mill**: frequency normalized per million. If an item has normalized frequency per million 10 it means that in a corpus with million tokens it would have frequency 10.

- **global subcorpus**: a subcorpus that is shared with all users. For setting up global subcorpora for user corpora see this help. (user corpora defined below).

- **header fields**: the various types of information associated with documents e.g. the year of publication or author

- **keywords**: words which are more frequent in one corpus compared to another
• **KWIC**: the word (or phrase) in the central column of a concordance which matched your query. Also referred to as the node word.

• **lc**: word in lowercase e.g. "cat" is the lc form of "Cat"

• **lemma**: the stemmed form of the word e.g. "cat" is the lemma for the word form "cats"

• **lemma-lc**: the stemmed form of the word in lower case

• **lempos**: the lemma conjoined by a hyphen with a shortened form for the part of speech e.g. n for noun

• **metadata**: variant term for header fields. The information associated with individual documents.

• **multi level list**: these provide a list sorted at more than one level e.g. sort by lemma, for any given lemma sort by word form and then for the lemma and word form, sort by PoS.

• **n-gram**: a contiguous sequence of n items (words, letters, morphemes, etc.) in given sequence of text or speech (see search n-grams)

• **node tag**: the part of speech (PoS) tag of the word in the central column of a concordance which matched your query.

• **node word**: the word (or phrase) in the central column of a concordance which matched your query. Also referred to as the KWIC.

• **normalized frequency**: see freq/mill

• **overall score**: (in Word sketches) - score of the relation (based on logDice - see score)

• **pattern**: you can enter a pattern or regular expression to retrieve a word list (Word List). You could for example look for all words beginning with "car" by using the expression "car.*"

• **PoS**: part of speech, the grammatical class of a word e.g. noun. The exact PoS classes used are shown in the tagset summary.

• **query**: word or phrase input by user to sketch engine in order to retrieve a concordance

• **references**: these attributes of the document where the text displayed in the concordance line is found. This is displayed in the left hand side of the concordance and the attributes displayed can be selected with view options

• **regular expressions**: these expressions provide a concise and flexible means for matching strings of text, such as particular characters, words, or patterns of characters. You could for example look for all words beginning with "car" by using the expression "car.*"

• **RE**: regular expression (see above)

• **salience**: a statistical measure of how salient a word or lemma is in a given context, given the frequency of the word and the context. This is measured with logDice see section 2 of ?ske-stat.pdf
• **score**: (in Word Sketches) - the logDice metric (see above)

• **search attribute**: in the word list page you can decide what attribute you want to create a frequency list for e.g. word or part of speech tag

• **search span**: the number of words either side of the node word that will be matched in the search

• **seed words**: words to start with, the initial state of an algorithm, e.g. the key words/phrases to search for in WebBootCaT

• **simple maths parameter**: this is a number you can provide which is used for the calculations for finding key words. If you give a low number e.g. 1 you will get lower frequency keywords, whereas a higher N will get higher frequency keywords (for further details see *SimpleMaths*).

• **subcorpus**: a subpart of a corpus. You can create a new subcorpus in Sketch Engine using the Text Types defined for the corpus following a links from various places, for example from the word list form

• **tagset summary**: the list of part of speech (PoS) classes used for annotation (tagging) of the corpus (if available, it can be accessed from corpus description in wiki or from Concordance Query page with CQL query selected).

• **TBL**: *Tick Box Lexicography*

• **Text Type**: a category or subcategory of a specific partition (usually the partitions are defined as documents) of the text of a corpus. The text types of a document are sometimes referred to as *header fields*.

• **thesaurus**: Sketch Engine uses "thesaurus" to mean a distributional thesaurus (see above) which is automatically created. This is not a man made thesaurus.

• **tokens**: every word and punctuation in a corpus is referred to as a token

• **tokenisation**: the automatic process or splitting the strings in a text into tokens

• **user corpora**: corpora produced by a user either by uploading their own data or by using WebBootCat to collect data from the Web. User corpora can be shared with other users.

• **WebBootCat**: a tool for creating corpora from the web. See documentation on creating a corpus, especially the WebBootCat section. Also refer to the WebBootCat paper (paper published: Proc EAMT 2006, Oslo, Norway).

• **word**: a token beginning with an alphabetical symbol (in UNICODE sense). Usually (for many languages) there are approx. 1.25x more tokens than words.

• **word sketch**: a corpus-based summary of a word's grammatical and collocational behaviour.

Click [here](http://www.skeetch.com) for the Start Page for Sketch Engine Documentation.