An introduction

Language is never ever, ever, random (Kilgarriff)
CORPUS LINGUISTICS

“CL is not a branch of linguistics in the same sense as syntax, semantics, SLX, and so on. All of these disciplines concentrate on describing/explaining some aspect of language use. CL in contrast is a methodology rather than an aspect of language requiring explanation or description. A corpus based approach can be taken to many aspects of linguistic enquiry.”

(McEnery & Wilson, 1996: 2)
• 400 ready-to-use corpora in 90+ languages

• [http://blogs.cardiff.ac.uk/linc/](http://blogs.cardiff.ac.uk/linc/) - for links and these slides.
SkELL (Sketch Engine for Language Learning) is a simple tool for students and teachers of English to easily check whether or how a particular phrase or a word is used by real speakers of English.

No registration or payment required. Just type a word and click a button.

All examples, collocations and synonyms were identified automatically by ingenious algorithms and state-of-the-art software analysing large multi-billion samples of text. No manual work was involved.

SkELL vs. Google Search
SkELL finds good examples of the word or phrase useful for language learners.
Google Search finds web pages with information about the topic specified by the word or phrase.
OPEN CORPORA AND SOME RECENT ADDITIONS

EUR-Lex Judgements Corpus

Extended corpus of English broadsheets

New academic English corpus

A new corpus of academic English was collected from the Directory of Open Access Journals.

The DOAJ corpus contains a title, country, year of publication, and currently offers only the English part planned.

New corpus from the environment domain

The LexiCon Research Group at the University of Granada developed and provided their highly specialised English EcoLexicon corpus built up of environmental texts. The corpus is hosted as an open corpus and is freely accessible even without a Sketch Engine account.

The corpus is a great source for searching keywords and terms from the field of environment. The EcoLexicon enables the user to search in a specific language variant (British, American, etc.), sort results by a country or year of publication, even specify a domain or genre.

modifiers of "climate"

<table>
<thead>
<tr>
<th>modifier</th>
<th>count</th>
<th>frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>global</td>
<td>6,801</td>
<td>26.43</td>
</tr>
<tr>
<td>humid</td>
<td>107</td>
<td>8.87</td>
</tr>
<tr>
<td>warm</td>
<td>174</td>
<td>8.85</td>
</tr>
<tr>
<td>wave</td>
<td>888</td>
<td>8.81</td>
</tr>
<tr>
<td>tropical</td>
<td>203</td>
<td>8.79</td>
</tr>
<tr>
<td>wet</td>
<td>112</td>
<td>8.66</td>
</tr>
<tr>
<td>temperate</td>
<td>91</td>
<td>8.60</td>
</tr>
</tbody>
</table>

See the most typical collocations for the noun climate.
GETTING HELP

A guide to Sketch Engine for...

Sketch Engine is for anyone working with or interested in or learning a language. Learn which features are the best for you.
KEY FEATURES OF SKETCH ENGINE

- Create your own corpus
- Word sketch
- Word Sketch Differences & Bilingual Word Differences
- Automatic term extraction
- Parallel Corpora
- …
• Search for a word or phrase
• Using EcoLexicon English (Environment)

• Go to https://the.sketchengine.co.uk/open/
KWIC CONCORDANCE LINES
• What do you think is the most frequently used word in this corpus?
• Let’s check…
What other words co-occur with WATER? WIND? AIR?
<table>
<thead>
<tr>
<th>TASK</th>
<th>CQL CODE</th>
<th>RESULT</th>
</tr>
</thead>
<tbody>
<tr>
<td>find examples of “went”</td>
<td>[word=&quot;went&quot;]</td>
<td>concordance of the word went</td>
</tr>
<tr>
<td>find examples of all forms of go</td>
<td>[lemma=&quot;go&quot;]</td>
<td>concordance of go, goes, going, gone, went</td>
</tr>
<tr>
<td>find examples of all words tagged with the tag NP</td>
<td>[tag=&quot;NP&quot;]</td>
<td>concordance of various words tagged as NP</td>
</tr>
</tbody>
</table>
WORD SKETCHES
• Create a word sketch for AIR, WIND, WATER (or any word you like)
SKETCH DIFFERENCE

- SIMPLE vs COMPLEX
STATISTICAL MEASURES

- some understanding of the measures used is needed
MI score: a measure of how strongly two words seem to associate in a corpus, based on the independent relative frequency of two words.

1) not dependent on the size of the corpus
2) can be compared across corpora, even if the corpora are of different sizes
3) gives information about its lexical behaviour, but particularly about the more idiomatic co-occurrences
4) the highest MI scores tend to be less frequent words with restricted collocation.

• The strength of the collocation is not always a reliable indication of meaningful association.
**T-SCORE**

**t-score**: a measure of how certain we can be that the collocation is the result of more than the vagaries of a particular corpus

1) Corpus size is important.

2) cannot be compared across corpora

3) gives information about the grammatical behaviour of a word

4) the highest t-scores tend to be frequently used words (whether or not they are grammatical words) that collocate with a variety of other words.

- In some instances they may require a wider span than is commonly used with respect to ‘clause collocation’
"sour" f=4109 ;“puss” f=254
"sour" only co-occur 3 times, this gives this particular collocation a very high MI score: i.e. these two words will be very strongly associated.

However, the t-score says "maybe, but we haven't seen enough evidence to be sure that the MI is right!".

The t-score is relatively low: 1.73
FALLING PRICES

\( f("falling") = 23,209 \)

\( f("prices") = 66,352 \)

The MI figure is not particularly high (8.415) because there is plenty of evidence of "falling" occurring without "prices" and vice versa.

Statistically the strength of association between "falling" and "prices" is much less than it was for "sour" and "puss". The t-score however is quite high at 28.673 shows it has taken into account the actual number of observations.
A SAFE GUIDE:

• A high T-score says: it is safe to claim that there is some non-random association between these two words.

• MI will highlight the technical terms, oddities, totally fixed phrases, etc.

• If a collocate appears in the top of both MI and T-score lists it is clearly a solid collocate
• MI, T-Score but also Sketch Engine’s own LogDice

• LOGDICE:
  • a statistic measure based only on a frequency of words $w_1$ and $w_2$ and the bigram $w_1w_2$, it is not affected by a size of the corpus

• See https://www.sketchengine.co.uk/documentation/statistics-used-in-sketch-engine/#logdice for more detail on various other calculations.
EXPLORING SOME OTHER FEATURES

- Filters
- Text Types
- Thesaurus
- Etc.
HOW MIGHT YOU WANT TO USE SKETCH ENGINE?

• Some free time to explore ways in which you might want to use Sketch Engine
REFERENCES

- Adam Kilgariff: https://www.kilgarriff.co.uk/
  - Rich resource of papers and presentations, e.g. How Many Words are There?, and many more