

SKETCH ENGINE

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)

[HTTPS://WWW.SKETCHENGINE.CO.UK/](https://www.sketchengine.co.uk/)

- 400 ready-to-use corpora in 90+ languages
- <http://blogs.cardiff.ac.uk/linc/> - for links and these slides.

SKELL



examples, collocations and thesaurus for learners of English

What is SkELL?

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.

Try SkELL

for learners of English

ruSkELL for Russian

для изучающих русский язык

csSkELL for Czech

pro studenty češtiny

OPEN CORPORA AND SOME RECENT ADDITIONS

EUR-Lex Judgements Corpus

Extended corpus of English broadsheets

New academic English corpus

DOAJ DIRECTORY OF
OPEN ACCESS
JOURNALS

A new corpus of academic corpus was collected from the [Directory of Open Access Journals](#) words.

The DOAJ corpus contains title, country, year of publication, and is available for search to a very narrow parameter. The DOAJ corpus currently offers only the English language corpus, but is planned.

New corpus from the environment domain

You are here: Home

<u>modifiers of "climate"</u>		
	<u>6,801</u>	26.43
global +	<u>419</u>	9.55
global climate .		
humid +	<u>107</u>	8.87
warm +	<u>174</u>	8.85
wave +	<u>888</u>	8.81
wave climate		
tropical +	<u>203</u>	8.79
wet +	<u>112</u>	8.66
wet climate		
temperate	<u>91</u>	8.60
in temperate climates		

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.

[EcoLexicon corpus in detail](#)

See the most typical collocations for the noun *climate*.

[show collocations](#)

GETTING HELP

Sketch Engine

Home News & Events Get Sketch Engine **User Guide** FAQ About us Contact

User guide You are here: Home / User guide

- Quick start guide**
main features in 1 minute
- User manual**
basic and advanced
- Documentation**
for expert users
- How do I ...?**
user guide arranged by task
- FAQ**
user space, subscription and payment
- Glossary**
terminology in one place
- Video**
self-study video resources
- Attend training**
face-to-face

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.

Translators

Teachers

Terminologists

Students

Lexicographers

Historians

KEY FEATURES OF SKETCH ENGINE

- Create your own corpus
- Word sketch
- Word Sketch Differences & Bilingual Word Differences
- Automatic term extraction
- Parallel Corpora
- ...

BASIC SEARCHES

- Search for a word or phrase
- Using EcoLexicon English (Environment)
- Go to <https://the.sketchengine.co.uk/open/>

KWIC CONCORDANCE LINES

Sketch Engine   [EcoLexicon English \(Environment\)](#)

- Home
- Search**
- Word list
- Word sketch
- Thesaurus
- Sketch diff
- Corpus info
- My jobs
- User guide 

Simple query:

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

WORD LIST

- What do you think is the most frequently used word in this corpus?
- Let's check...

The screenshot shows the 'Word list options' page in the Sketch Engine interface. The page is titled 'Word list options' and features a sidebar on the left with navigation links: Home, Search, Word list (highlighted), Word sketch, Thesaurus, Sketch diff, Corpus info, My jobs, and User guide. Below these are 'All words', 'All lemmas', and 'Find x'. The main content area is light blue and contains the following options:

- Subcorpus:** A dropdown menu set to 'None (whole corpus)' with links for 'info' and 'create new'.
- Search attribute:** A dropdown menu set to 'word'.
- Use n-grams:** A checkbox that is unchecked. Below it, a range of 'n' values is set from 2 to 2.
- Hide/nest sub-n-grams:** A checkbox that is unchecked.
- Filter options:**
 - Filter word list by:** A text input field for a regular expression.
 - Minimum frequency:** A text input field set to 5.
 - Maximum frequency:** A text input field set to 0, with a note '(0 = no maximum frequency)'.
 - Whitelist:** A 'Choose file' button, 'No file chosen', and a 'Clear' button.
 - Blacklist:** A 'Choose file' button, 'No file chosen', a 'Clear' button, and a 'format' link.
 - Include non-words
- Output options:**
 - Frequency figures:** Radio buttons for 'Hit counts' (selected), 'Document counts', and 'ARF'.
 - Output type:** Radio buttons for 'Simple' (selected) and 'Keywords'.
 - Reference (sub)corpus:** A dropdown menu set to 'English Web 2013 (enTenTen13)' and another set to '(whole corpus)'.
 - Prefer:** A slider between 'rare words' and 'common words', currently positioned near 'rare words'.
 - Change output attribute(s):** Three empty dropdown menus.
 - A note: 'You can select one or more output attributes. Please note that this option can be time-consuming.'
- Make word list:** A button at the bottom left of the main content area.

COLLOCATION

- What other words co-occur with WATER? WIND? AIR?

Sketch Engine water EcoLexicon English (Environment)

Query water 90,533 (3,163.70 per million)

Page 1 of 4,527 Go Next Last

doc#0 (rubble mound and gravity types) increase with water depth , requiring a great amount of
doc#0 . In addition , they prevent the circulation of water and so deteriorate the water quality near the
doc#0 the circulation of water and so deteriorate the water quality near the coast. In some places , they
doc#0 upper part is impermeable and extends above the water level to some distance below sea level. The
doc#0 flow exchange between the partially enclosed water body and the open sea. The present
doc#0 barrier near the free surface extends above the water level to some distance below sea level and
doc#0 facilities (wave flume dimensions and used water depths) , wave type and main investigated wave
doc#0 sketched in Fig. 1 , in which h is the constant water depth in still water ; D [m] is the draft of the
doc#0 1 , in which h is the constant water depth in still water ; D [m] is the draft of the upper part ; b [m] is the
doc#0 measured vertically upwards from the still water level. A regular wave train with wave height Hi [m]
doc#0 study. The proposed breakwater can be used for water depths ranging from 10 m to 15 m. However , the
doc#0 is carried out in the laboratory for a constant water depth of 0.50 m , slots width and gap of 0.02 m , and
doc#0 from 0.9 to 1.9 s. These ranges correspond to 15 m water depth , 0.6 m slots width and gap , and 4.9 to 10.4 s
doc#0 experiments were carried out with a constant water depth (h) of 0.5 m and with generator motions
doc#0 is 6 m each. 4.4. Wave height measurements The water level variations which resulted from the
doc#0 D/h. This can be explained by considering the water particle motions. As h/L increases the water
doc#0 water particle motions. As h/L increases the water particle velocity and acceleration increases
doc#0 the wave comes across the breakwater model , the water particle velocity and acceleration suddenly
doc#0). In addition , as D/h increases , the area which water path through decreases then the transmitted
doc#0 wave motion is minimal in the lower part of the water column for short waves (Huang 2007). 5.5. Model

Page 1 of 4,527 Go Next Last

Home
Search
Word list
Word sketch
Thesaurus
Sketch diff
Corpus info
My jobs
User guide

Save
Make subcorpus
View options
KWIC
Sentence
Sort
Left
Right
Node
Shuffle
Sample
Filter
Sub-hits
1st hit in doc
Frequency
Node tags
Node forms
Text types
Collocations
Visualize

Sketch Engine water EcoLexicon English (Environment)

Home
Search
Word list
Word sketch
Thesaurus
Sketch diff
Corpus info
My jobs
User guide

Concordance
Frequency
Node tags
Node forms
Text types

Collocation candidates

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

Minimum frequency in corpus: 5

Minimum frequency in given range: 3

Show functions: logDice Sort by: logDice

Make candidate list Save options

T-score	T-score
MI	MI
MI3	MI3
log likelihood	log likelihood
min. sensitivity	min. sensitivity

MORE ADVANCED SEARCHES WORD – LEMMA - TAG

TASK	CQL CODE	RESULT
find examples of “went”	[word="went"]	concordance of the word went
find examples of all forms of go	[lemma="go"]	concordance of go, goes, going, gone, went
find examples of all words tagged with the tag NP	[tag="NP"]	concordance of various words tagged as NP

MORE ADVANCED SEARCHES WORD – LEMMA - TAG

Sketch Engine   EcoLexicon English (Environment)

- Home
- Search
- Word list
- Word sketch
- Thesaurus
- Sketch diff
- Corpus info
- My jobs
- User guide 

Simple query:

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

Query type simple lemma phrase word character CQL

Lemma: PoS:

Phrase:

Word form: PoS: match case

Character:

CQL: Default attribute:

[Tagset summary](#) [CQL builder](#)

WORD SKETCHES

test/experiment

(noun) Alternative PoS: [verb](#) (freq: 1,427)
British Academic Written English Corpus (BAWE) freqs = [2,458](#) | [2,186](#)

test 6.0 4.0 2.0 0 -2.0 -4.0 -6.0 experiment

and/or	240	181	0.10	0.08
test	20	0	10.4	--
score	7	0	9.7	--
ultrasound	4	0	9.0	--
x-ray	3	0	8.5	--
examination	3	0	8.2	--
analysis	6	0	7.9	--
result	5	0	7.6	--
data	3	0	7.2	--
sample	4	3	8.0	7.8
time	0	3	--	6.0
value	0	3	--	6.4
theory	0	6	--	7.6
observation	0	3	--	8.3
concentration	0	3	--	8.3
participant	0	4	--	8.8
experiment	0	14	--	10.3

subject_of	242	307	0.10	0.14
confirm	7	0	8.9	--
give	9	4	7.3	6.1
indicate	11	5	8.3	7.1
consist	6	3	8.0	6.8
measure	6	5	8.7	8.2
produce	3	4	6.3	6.6
carry	5	7	7.9	8.3
involve	14	19	8.6	9.0
show	27	38	8.0	8.5
require	3	5	6.4	7.0
suggest	3	7	5.6	6.8
use	9	23	6.4	7.7
make	0	4	--	5.4
find	0	3	--	6.5
include	0	6	--	7.0
present	0	3	--	7.1
highlight	0	3	--	7.1
support	0	5	--	7.2
run	0	3	--	7.3
take	0	11	--	7.4
look	0	5	--	7.6
demonstrate	0	5	--	7.7
aim	0	7	--	8.3
prove	0	12	--	9.0
investigate	0	9	--	9.4

adj_subject_of	55	27	0.02	0.01
useful	4	0	8.8	--
accurate	0	3	--	9.4

object_of	492	384	0.20	0.18
pass	13	0	8.9	--
reset	6	0	8.6	--
apply	15	0	8.4	--
stretch	5	0	8.2	--
stand	5	0	8.2	--
devise	5	0	8.1	--
execute	5	0	8.0	--
fail	4	0	7.8	--

subject_of	242	307	0.10	0.14
confirm	7	0	8.9	--
give	9	4	7.3	6.1
indicate	11	5	8.3	7.1
consist	6	3	8.0	6.8
measure	6	5	8.7	8.2
produce	3	4	6.3	6.6
carry	5	7	7.9	8.3
involve	14	19	8.6	9.0
show	27	38	8.0	8.5
require	3	5	6.4	7.0
suggest	3	7	5.6	6.8
use	9	23	6.4	7.7
make	0	4	--	5.4
find	0	3	--	6.5
include	0	6	--	7.0
present	0	3	--	7.1
highlight	0	3	--	7.1
support	0	5	--	7.2
run	0	3	--	7.3
take	0	11	--	7.4
look	0	5	--	7.6
demonstrate	0	5	--	7.7
aim	0	7	--	8.3
prove	0	12	--	9.0
investigate	0	9	--	9.4

house

(noun) ukWaC freq = [391,778](#) (251.18 per million)

Haus

(noun) deTenTen [2013] freq = [7,264,685](#) (364.72 per million)

Use another candidate translation: [erinnern](#) [Hausarrest](#) [Ordnung](#) [Plenum](#) [hoch](#) [daran](#) [Parlament](#) [kehren](#) [mitteilen](#)
Click on collocates to access reciprocal bilingual search

object of	96,897	2.10	VerbY+SubstXDat (obj_dat_of)	1,476,115	3.60
terrace	1,667	9.04	wohnen	31,400	7.14
detach	1,737	8.94	fühlen	26,177	6.04
build	8,502	8.59	fahren	46,582	5.99
buy	3,998	8.10	kehren	7,991	5.64
board	853	7.94	befinden	30,253	5.35
rent	935	7.90	schicken	9,019	5.34
sell	2,452	7.58	sitzen	15,913	5.25
demolish	650	7.54	leben	23,884	5.21
situate	1,061	7.39	kommen	144,447	5.04
own	1,284	7.25	holen	10,220	4.99
occupy	798	7.12	eilen	1,891	4.84
move	2,456	7.00	rennen	2,939	4.82

subject of	58,690	1.90	SubstXNom+VerbY (subj_of)	322,711	0.80
overlook	244	6.20	beherbergen	1,910	5.78
stand	734	6.12	verfügen	16,744	5.49
belong	306	6.09	brennen	3,018	5.48
rebuild	135	5.61	abbrennen	364	4.83
date	266	5.52	schmiegen	384	4.60
front	86	5.39	erstrahlen	562	4.56
burn	113	4.98	befinden	15,624	4.47
line	84	4.93	bestechen	933	4.33
occupy	133	4.87	säumen	294	4.17
collapse	63	4.71	gruppieren	255	4.14
boast	75	4.68	liegen	32,608	3.99
survive	107	4.56	einstürzen	187	3.97

adj_subject_of	7,821	2.10	modifier	174,914	1.30	modifier	6,611,554	1.60	modifies	54,780	0.40
----------------	-------	------	----------	---------	------	----------	-----------	------	----------	--------	------

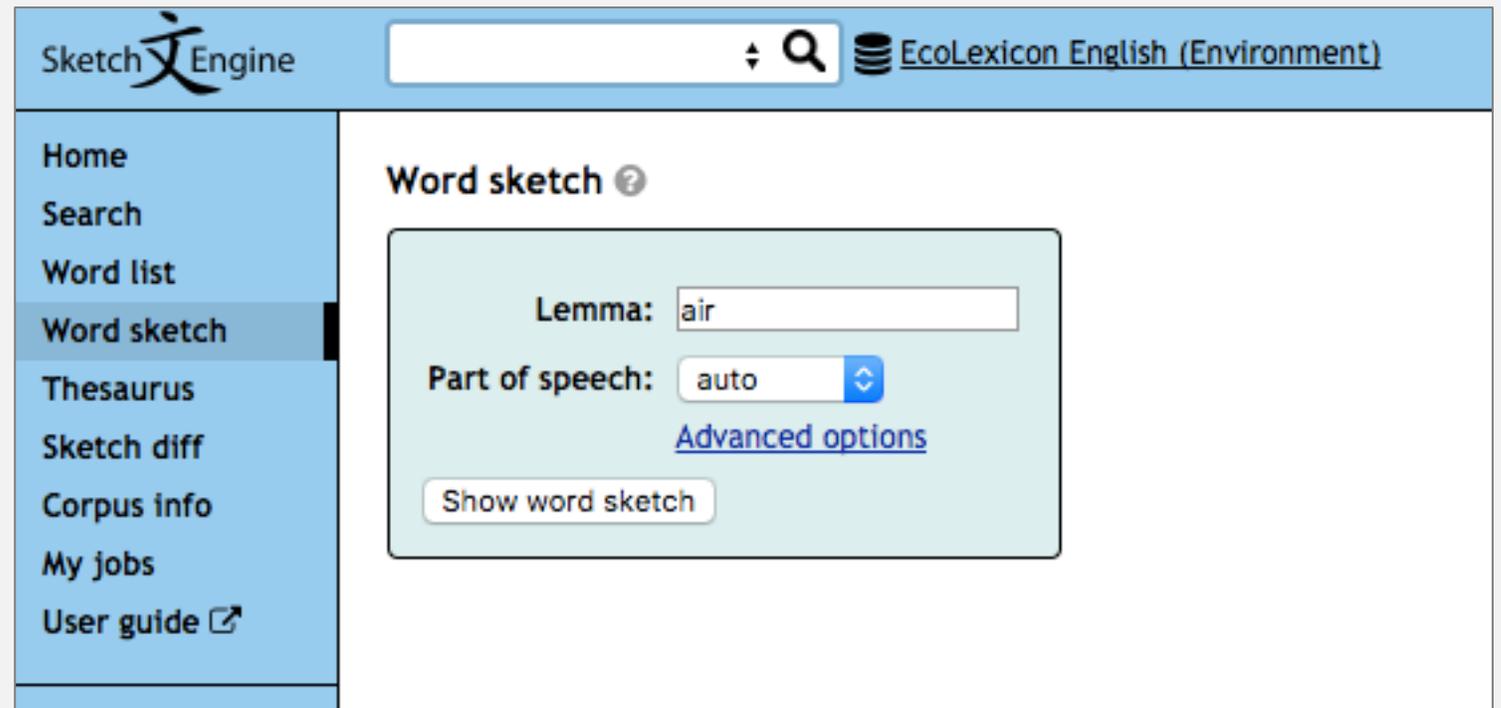
team

(noun) Alternative PoS: [verb](#) (478)
British National Corpus (BNC) freq = [22,482](#) (200.21 per million)

modifiers of "team"	13,919	0.62	nouns and verbs modified by "team"	3,166	0.14	verbs with "team" as object	4,616	0.21	verbs with "team" as subject	6,300	0.28	"team" and/or...	2,244	0.10
management +	433	9.31	spirit +	112	9.15	lead +	205	8.48	win	98	7.97	football	12	7.15
football +	207	8.63	mate	53	8.75	head	63	8.26	team won			cast	8	6.75
project +	166	8.35	the project team			team headed by			play +	105	7.86	search	9	6.71
england +	143	8.05	the england team			join +	113	8.04	work +	109	7.53	group	31	6.55
research +	164	7.83	the research team			pick	47	7.79	team working			squad	7	6.55
rescue	98	7.76	mountain rescue team			field	26	7.43	lose	40	6.78	individual	12	6.41
display	91	7.60	the national display team			assemble	25	7.17	team lost			husband	12	6.37
cup	96	7.45	cup team			beat	34	7.01	consist	31	6.78	husband and wife team		
design	87	7.38	the design team			negotiate	26	7.00	team consists of			player	10	6.35
						captain	18	6.92	perform	27	6.74	supporter	7	6.19
						send	55	6.86	compete	22	6.70	afternoon	7	6.17
						strengthen	22	6.79	teams competing in			fan	6	6.11
						investigate	27	6.77	find	57	6.55	panel	6	6.11
						the investigating team			team found			specialist	6	6.08
						championship	49	7.77	comprise	21	6.46	sale	10	6.07
						team championship			prepare	22	6.45	member	16	6.01
						selection	38	7.73	team comprising			department	10	5.93
						visit	36	6.53	105	6.36	management	12	5.91	
											manager	13	5.88	

WORD SKETCH

- Create a word sketch for AIR, WIND, WATER (or any word you like)



The screenshot shows the Sketch Engine interface. The top navigation bar includes the logo "Sketch Engine", a search bar, and the text "EcoLexicon English (Environment)". A left sidebar contains a menu with the following items: Home, Search, Word list, Word sketch (highlighted), Thesaurus, Sketch diff, Corpus info, My jobs, and User guide with an external link icon. The main content area is titled "Word sketch" with a help icon. It features a form with the following fields: "Lemma:" with a text input containing "air", "Part of speech:" with a dropdown menu set to "auto", and a link for "Advanced options". Below the form is a button labeled "Show word sketch".

SKETCH DIFFERENCE

- SIMPLE vs COMPLEX

Sketch Engine EcoLexicon English (Environment)

Home
Search
Word list
Word sketch
Thesaurus
Sketch diff
Corpus info
My jobs
User guide [↗](#)

Word sketch differences [?](#)

Lemma:

Part of speech:

Sketch diff by: lemma

subcorpus

Second lemma:

First subcorpus: [info](#) [create new](#) [?](#)

Second subcorpus: [info](#) [create new](#) [?](#)

word form

First word form:

Second word form:

[Advanced options](#)

STATISTICAL MEASURES

- some understanding of the measures used is needed

MI SCORE

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'

INTERPRETING COLLOCATES

Collocation candidates for PUSS

	<u>Freq</u>	<u>T-score</u>	<u>MI</u>
<u>p/n</u> puss	10	3.162	18.882
<u>p/n</u> glamour	4	2.000	12.646
<u>p/n</u> sour	3	1.732	12.016
<u>p/n</u> Taking	3	1.731	11.589
<u>p/n</u> Hello	4	1.999	10.861
<u>p/n</u> Little	3	1.730	9.577
<u>p/n</u> Here	5	2.232	9.253
<u>p/n</u> November	3	1.723	7.646
<u>p/n</u> black	3	1.723	7.542
<u>p/n</u> bit	3	1.721	7.238
<u>p/n</u> Britain	3	1.720	7.209

"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

INTERPRETING COLLOCATES

	Freq.	T-score	MI
p/n pheasants	4	1.995	8.672
p/n bottomless	9	2.993	8.670
p/n stretched	3	1.728	8.646
p/n 17-0	3	1.728	8.639
p/n torrents	5	2.230	8.636
p/n pro-business	3	1.728	8.593
p/n farmhouses	5	2.230	8.590
p/n 100ft	4	1.995	8.586
p/n madly	13	3.596	8.586
p/n dappled	4	1.995	8.564
p/n steadily	91	9.514	8.556
p/n bemoan	3	1.727	8.472
p/n seams	15	3.862	8.471
p/n rain	329	18.087	8.468
p/n Graff	3	1.727	8.465
p/n Recovery	18	4.231	8.462
p/n Moslems	4	1.994	8.453
p/n 37.5	3	1.727	8.452
p/n weightless	3	1.727	8.452
p/n storeys	8	2.820	8.446
p/n underperformed	3	1.727	8.445
p/n gashed	3	1.727	8.438
p/n fracturing	4	1.994	8.418
p/n prices	827	28.673	8.415

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:

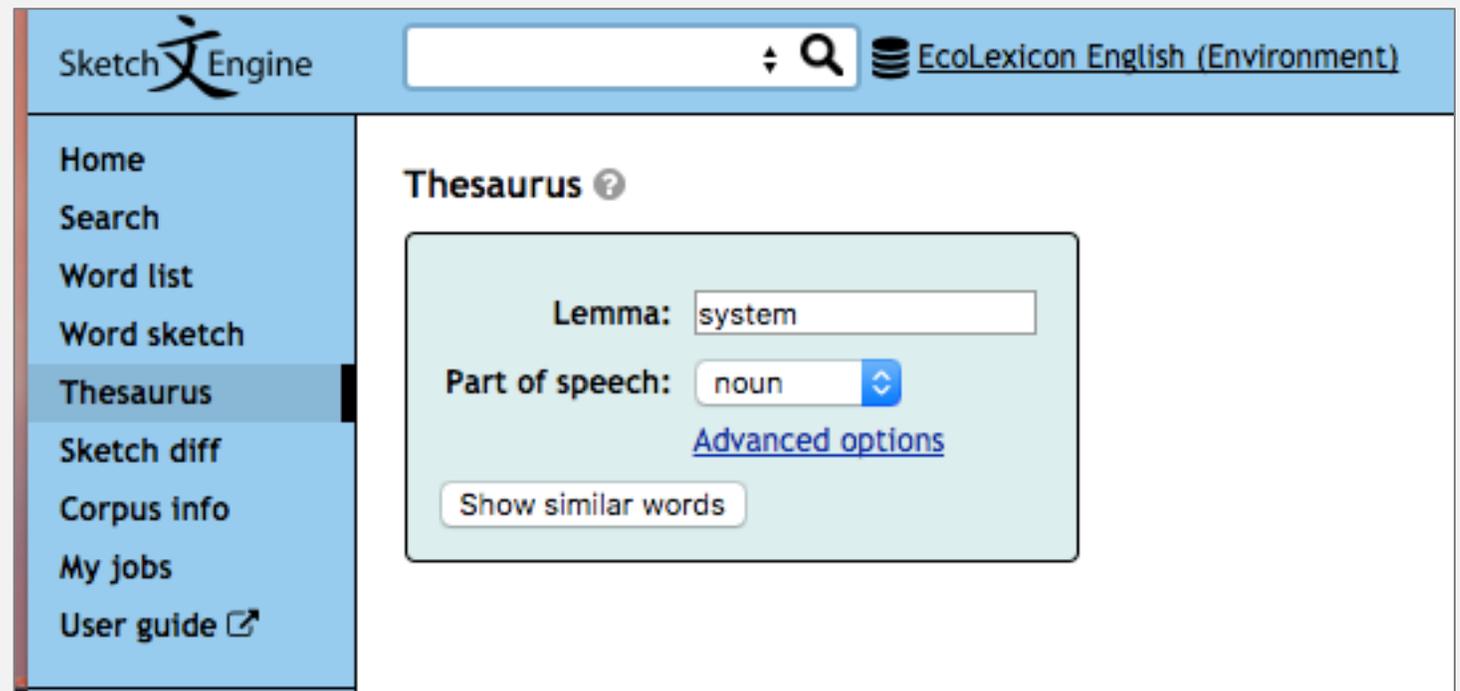
- A high T-score says: it is safe to claim that there is some non-random association between these two words.
- MI will highlight the technical terms, oddities, totally fixed phrases, etc.
- If a collocate appears in the top of both MI and T-score lists it is clearly a solid collocate

STATISTICAL MEASURES IN SKETCH ENGINE

- 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.



The screenshot shows the Sketch Engine interface. At the top, there is a navigation bar with the logo "Sketch Engine" on the left, a search bar in the center, and "EcoLexicon English (Environment)" on the right. Below the navigation bar is a vertical menu on the left with the following items: Home, Search, Word list, Word sketch, Thesaurus (highlighted), Sketch diff, Corpus info, My jobs, and User guide with an external link icon. The main content area is titled "Thesaurus" with a help icon. It contains a form with the following fields: "Lemma:" with a text input containing "system", "Part of speech:" with a dropdown menu showing "noun", and a link for "Advanced options". Below the form is a button labeled "Show similar words".

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 Kilgarriff: <https://www.kilgarriff.co.uk/>
 - Rich resource of papers and presentations, e.g. How Many Words are There?, and many more
- Hunston, S. (2002). *Corpora in applied linguistics*. Cambridge: Cambridge University Press
- Kilgarriff, Adam, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, Vít Suchomel (2014): The Sketch Engine: ten years on.. *Lexicography* 1 (1): 7–36.
- McEnery, T. & Wilson, A. 1996. *Corpus Linguistics*. Edinburgh: Edinburgh University Press.
- Rychlý, P. (2008). A lexicographer-friendly association score. In Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN, pp. 6–9.
- Thomas, J. 2016. *Discovering English with the Sketch Engine*. 2nd edition. Versatile.